The Role of Imitation in Learning to Pronounce

Piers Ruston Messum

University College London

April 2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I, Piers Ruston Messum, declare that the work presented in this thesis is my own. Where information has been derived from other sources this has been acknowledged in the thesis.

Abstract

Timing patterns and the qualities of speech sounds are two important aspects of pronunciation. It is generally believed that imitation from adult models is the mechanism by which a child replicates them. However, this account is unsatisfactory, both for theoretical reasons and because it leaves the developmental data difficult to explain.

I describe two alternative mechanisms. The first explains some timing patterns (vowel length changes, 'rhythm', etc) as emerging because a child's production apparatus is small, immature and still being trained. As a result, both the aerodynamics of his speech and his style of speech breathing differ markedly from the adult model. Under their constraints the child modifies his segmental output in various ways which have effects on speech timing; but these effects are epiphenomenal rather than the result of being modelled directly.

The second mechanism accounts for how children learn to pronounce speech sounds. The common, but actually problematic, assumption is that a child does this by judging the similarity between his own and others' output, and adjusting his production accordingly. Instead, I propose a role for the typical vocal interaction of early childhood where a mother reformulates ('imitates') her child's output, reflecting back the linguistic intentions she imputes to him. From this expert, adult judgment of either similarity or functional equivalence, the child can determine correspondences between his production and adult output. This learning process is more complex than simple imitation but generates the most natural of forms for the underlying representation of speech sounds. As a result, some longstanding problems in speech can be resolved and an integrated developmental account of production and perception emerges.

Pronunciation is generally taught on the basis that imitation is the natural mechanism for its acquisition. If this is incorrect, then alternative methods should give better results than achieved at present.

Table of Contents

Abbreviations

A, B, C	female demonstrator, male observer, male child observer
AS	awareness of sensation
BSD	breath stream dynamics
EBP	elevated background pressure
FLS	foot level shortening
IM	inverse model
L0, L1, L2	protowords, first language, second language
ME	mirrored equivalence
MP	meaningful perception
OMR	object movement re-enactment
P_{alv}	alveolar (lung) pressure
P_{sg}	subglottal pressure
P-centre	perceptual centre
PFC	pre-fortis clipping
R_{law}	lower airway resistance
R_{uaw}	upper airway resistance
RP	received pronunciation
SB	speech breathing
SBE	similarity-based equivalence
SSR	speech sound re-enactment
VC	vital capacity
VMS	vocal motor scheme
VOT	voice onset time
WGmPh	West Germanic phenomena

Acknowledgments

It has taken a long time to assemble this thesis and it would not have been possible without the help that the many people named below gave me. I am grateful to all of them, and apologise to anyone inadvertently omitted. Some people have made particularly important contributions and I am very much in their debt.

It would be impractical to also list and thank all the people who have assisted me in my practice of Vipassana meditation, but without this, too, I would not have completed the project.

There is no significance to the order that names appear in.

Peter and Pauline Messum, Ashley Messum, Chris Carnie and Jordan Jarrett, Simon and Kate Gough, Brigid Rentoul and Gerald Hughes, Sue Rentoul, Susan Attwood, Andy Willcocks, Judy Kendall, Trish Dearsley, Claire Suthren, Nina Hajnal, Mark and Shirin Thomas, Richard and Karen Starkey, Philippa Reeves, Pat Elliott, Bradley Mills, Mike Barker, Mitch Friedman, Peter and Evelyn Bacon, Anne Whaley

Roslyn Young, Fusako Allard, Michael Hollyfield, Donna and Alain l'Hôte, Glenys Hanson, Lois Rose, John Olsen, Allen Rozelle, Marie-Laure Lagrange, Mme Vigier, Catherine Rosier, Pascale Laporte, the ATM SoE group (organised for many years by Geoff Faux)

Steve Lancashire (Charterhouse in Southwark), Susan Morland, Richard Gaskell, Marco Federighi, Daniel Roder (INSEF Conseil), Pedro Tromoso, Kurdish Resource Centre (Kennington)

Michael Vaughan-Rees, Adrian Underhill, Adam Brown, Barbara Bradford, Jonathan Marks

Alessia Bolis, Dolores Ditner, Marc-Georges Nowicki, Mireille Michel, Rudolf Lüer, Sylvie Battle, Tina Dickson

Michael Ashby, Adrian Fourcin, Evelyn Abberton, John Maidment, Molly Bennett, Stefanie Anyadi, Martyn Holland, Lisa Migo, David Cushing, Mahen Goonewardane, John Wells, Jill House, Neil Smith, Warwick Smith, Paul Iverson, Gordon Hunter, Bronwen Evans, Volker Delwo, Mary Wykes, Yves Le Clézio, Val Hazan, Mark Huckvale, Steve Nevard

Brenda Cross, Bill Potter, Charles Maxwell, Wilf Francis, Harold (Ces) Williams, Duncan Farquharson, Alan Hogben, Graham Crowther, Jamie Ingram, Mike Goring, Sally Page, Michael Hanley, Tony Gardner-Medwyn, Bruce Lynn, Lynn Bindman

Ian Howard, David Shanks, Sophie Scott, Peter Howell, Celia Heyes, John Goldstone, Andrew Smith, Sylvia Warner, Geoff Cusick, Rod Lane, Nuccia Quinn, Elizabeth Royston, Mary Plackett, Peter Swartz, UCL Graduate School (for their support of cross-disciplinary research)

Janice Chapman, Ingrid Rugheimer, Jill McCullough, Julian McGlashan, Meredydd Harries, David Howard, Kirsten and Borge Frøkjær-Jensen, Valerie Morton, John Rubin, David Clark, British Voice Association, Laryngograph Ltd, Acoustical Society of America

Mike MacMahon, Marilyn Vihman, Alan Bell, Carol Fowler, Mary Morrell, Robert Schroter, Elaine Stathopoulos, Kevin Murphy, Harm Schutte, Mamoru Kinjo, Gabrielle Konopczynski, Pierre Badin, Louis-Jean Böe, Lucie Ménard, Bernard Gautheron, Gerald Moon, Alan Cruttenden, Michael O'Neill, Celia Scully, Pamela Davis, Sidney Wood, Anna Barney, Christine Shadle, Hal Edwards, Shushan Teager, Richard McGowan, Steve Kelly, Paul Sharp, James Hannah, Peter Jones, Graham Hassall, Robert Jaruwczelski, Tom Hixon, Jeannette Hoit, Joseph Milic-Emili, Georges Boulakia, staff in U.O. Pneumologia at IRCCS Pozzolatico (Fondazione Don Carlo Gnocchi), Chin-Wu Kim, Mark Tatham, Peta White, John Hewson, Carol Boliek, Eileen Finnegan, Peter Watson, Robert Mayr, Lise Menn, Sarah Hawkins, Rachel Smith, Geoff Morrison, Gautam Vallabha, Peter McLeod

Peter Roach, David Faber, Gary Weismer, John Catford, Esther Thelen, Gerd Gigerenzer, John Ohala, Johan Sundberg

1 Introduction

> "Imitation is obviously so important in learning to speak that at first sight it might seem to be the only thing that matters, the one essential condition of a child's progress in language. It is obvious that English can become the mother tongue of children only if they can imitate speakers of English...
>
> A simple answer to the question, how does a child learn to speak? might therefore seem to be, By imitation.
>
> This answer is certainly too simple."
>
> M.M. Lewis, *How Children Learn to Speak* (1957:48)

It may be too simple to say that children learn to speak by imitation, but there is a great deal of converging evidence that has seemed to insist that they learn to pronounce, at least, this way. Firstly, infants can judge sound similarity across a wide range of others' voices (e.g. Kuhl 1991), so why should they not use this skill to guide the production of the qualities of their own speech in some way? Secondly, they also replicate some apparently arbitrary patterns of speech timing, like the vowel length differences in *sit* and *seat*. Since many of these patterns are language specific, this seems to be particularly strong evidence for imitation to be playing a role, for how else could they be learned? Thirdly, mimicry is something that infants, older children and adults are able to do, and imitation clearly does play a role in a child's more general learning of language and of other skills. Why not in the learning of pronunciation, too?

Note that the final question asked specifically about the learning of pronunciation, by which I meant the mature skill of pronunciation. The first words of an infant may be produced on a different basis, by being re-created holistically[1]. However, the way an older speaker pronounces a word depends on how he reproduces the speech sounds that make it up. So the genesis of word production and the genesis of the mature skill of pronunciation may not be the same. The latter relies on the learning of speech sounds, which may not depend upon earlier skills of mimicry.

A similar distinction can be made between drawing and writing a word. We learn to handwrite so that we can reproduce a word by writing the letters that make it up. But we

[1] 'Holistic' is explained by Studdert-Kennedy (2002:213): "Early words are said to be holistic because, although they are formed by combining gestures, gestures have not yet been differentiated as context-free, commutable units that can be independently combined to produce new words."

may draw the very first marks on a page that can be read as a word. Asked to copy some Bangla script, for example, the best that I could do would be to draw it: to re-create it 'holistically', for which it would make little difference if I copied it right way up or upside down. To understand any skill of handwriting I might develop in future, however, we would need to look at the stage where I start to reproduce Bangla letters by writing them, instead[2].

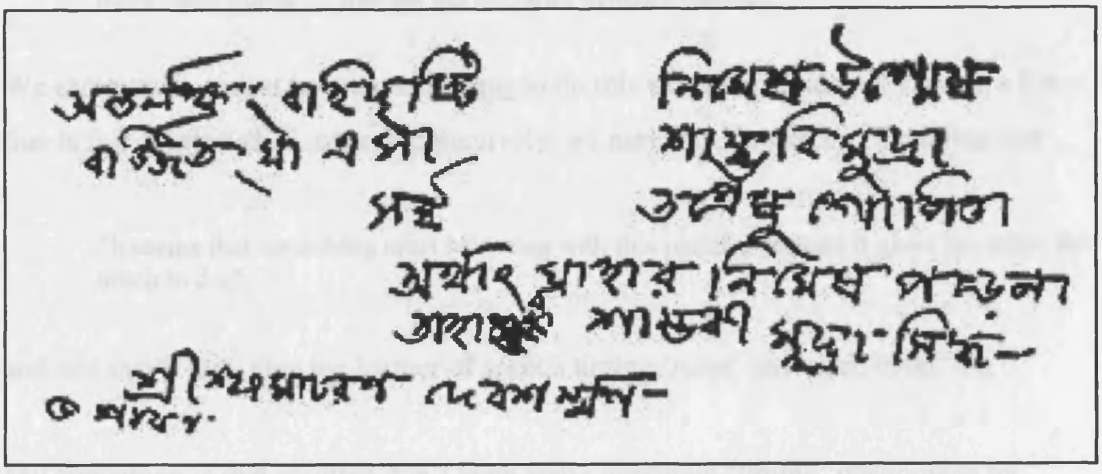

Figure 1-1. Example of Bangla script. For a non-writer of Bangla, this can be re-created (drawn) equally well whatever the orientation of the original.

This thesis is about the path to the mature skill of pronunciation, so one issue it examines is how speech sounds develop. The theoretical problems with this occurring by a process of imitation are not widely appreciated, although Locke and Snow (1997:282-283) draw attention to one of their possible effects, the "mystery" of why adults fail to achieve perfect foreign accents.

A second issue examined is the development of some timing patterns. If these are imitated then the child must certainly be a talented modeller, who discerns and then reproduces the sometimes fine temporal detail he finds in the speech of others. This modelling will include the interaction of timing 'rules', of course, when several different phenomena affect the duration of a single phonetic interval. This is something that speech researchers have been unable to replicate (e.g. Gopal 1996).

[2] Both producing words from a model holistically and producing them via a particulate principle could be said to be production 'by imitation'. As I have started to do already, I will use 're-creation' for the former and 'reproduction' for the latter, to try to keep these processes distinct.

In fact, the more we understand about the fine-grained temporal patterning of speech, the more remarkable a child's modelling achievements become. Twenty years ago, Fowler (1985:231) discussed just a limited selection of the processes believed to be operating in speech, and invited us to,

> "... sit back and marvel at how wonderful the mind is to be able to keep track of all these things in the course of language production (all of these things and then all of those other things ... that get the utterance actually uttered)."

We can marvel, too, at how a child learns to do this while simultaneously living a life that is full of other challenges. Alternatively, we may join Fowler in concluding that,

> "it seems that something must be wrong with this picture, because it gives the talker too much to do,"

and add that it may give the learner of speech timing 'rules' too much to do, too.

My contention is that children don't learn some important 'timing' phenomena by imitation, and nor do they learn the pronunciation of speech sounds this way. To take imitation's place, I will describe different mechanisms for replicating timing relationships and for producing acceptable sound qualities. For that reason, this thesis is divided into two parts.

In Part 1, I will investigate speech breathing and aerodynamics in a child, to show how his speech may be canalised towards producing phenomena which include durational compression effects, the properties of tense and lax vowels, and long lag voice onset times (if he is learning English or German).

In Part 2, I will argue that if a child can be as attentive to others' responses to his vocal output as we now believe him to be attentive to their speech, then it is possible for him to learn the pronunciation of speech sounds in this way.

The idea that young children learn to pronounce by imitation is deeply embedded in our assumptions about speech; but it has only ever been an assumption, made in the absence of any more plausible alternative. In fact, there are good reasons why it is unsatisfactory. There are also many advantages, theoretical and practical, to the alternative mechanisms I will be proposing. Those of Part 1, concerned with 'timing'

phenomena, would make learning to talk a more straightforward achievement for a child than this seems to be at present. They would also explain anomalies in the developmental data and make the phonetics of English (and other West Germanic languages[3]) considerably more coherent.

The account of speech sound acquisition presented in Part 2 has similar advantages. It avoids problematic aspects of any account that involves acoustic matching by a child, it simplifies the child's task by giving a more significant role to his caregivers, and it generates a new form of underlying representation for speech sounds. This may explain some longstanding issues, including whether speech is best characterised as gestures made audible (as Stetson and others have claimed), or as an acoustic code.

From a practical point of view, the proposals in Parts 1 and 2 offer the prospect of improved results in the teaching of pronunciation and in therapeutic intervention.

- § -

Please note that in writing this thesis I have been constrained by both time and the maximum number of words I am (rightly) allowed. So some section headings are followed by a star (*). This indicates that the treatment of the topic is not as full as I would have liked.

The terms 'infant' and 'child' are sometimes applied on the basis of linguistic ability, sometimes by age. I have tried to reserve the former for the more or less pre-linguistic state. When it is impossible or unimportant to make a distinction I have used 'child' or 'young child'.

Slightly confusingly, Part 1 generally deals with children at an older age than those in Part 2. I have arranged matters this way partly because the arguments in Part 1 are more straightforward, and partly because some of the early conclusions are relied on in later discussions.

---

[3] I have considered English, German and Dutch in the research reported here, but not others such as Afrikaans, Frisian, etc.

Where I have inserted direct quotations, any emphasis made by the original author appears in italic or as an underlined word. Any text in bold has been highlighted by me.

Finally, to avoid continual use of "he or she", "his or her", etc, I describe all childhood interactions as being between a male child and his mother. When describing more general issues in imitation, the demonstrator of a behaviour is female and designated 'A'; a mature, male observer is designated 'B'; and a young, male observer is designated 'C'.

Part 1 – The Replication of Patterns of Timing

2 Introduction to Part 1

2.1 *Four 'temporal' phonetic phenomena*

Part 1 of this thesis is largely concerned with the replication by children of what are believed to be timing effects in adult speech. I will concentrate on four of these, which I will introduce here and then describe in more detail later. The first is found in many languages; the next three are particularly characteristic of English and other West Germanic languages.

1. **Pre-fortis clipping (PFC)**

   When a phonologically voiceless consonant closes a syllable, the speakers of many languages reduce the length of the preceding vowel. So, in English the vowel in *seat* is shorter than that in *seed.*

   Naeser (1970) found this effect in children as young as 22 months.

2. **Rhythmic adjustments**

   Speakers make various adjustments in the timing of segments that create the impression of 'rhythmic' differences between languages.

   There is disagreement about whether the basis for these adjustments lies in rhythmicity *per se,* or in the operation of other processes with non-rhythmic motivations which then combine to give an impression, for example, of a 'stress-timed'[4] rhythm for English (Laver 1994:532). In either case, the phenomenon which contributes most to the pattern of the final output is 'foot level shortening' (FLS)[5]. This describes how syllables in a foot shorten as their number grows. At present, this is an outstanding piece of evidence in favour of rhythmicity as a factor in normal speech production, since no satisfactory non-rhythmic motivation for FLS has been demonstrated.

---

[4] It would be tiresome to always enclose *rhythm* and *stress-timing* in single quotation marks to indicate my doubts about how real they are. I will use the words only as labels for the timing effects attributed to stress-timed rhythm rather than as explanatory terms.

[5] Also called 'rhythmic clipping' (Wells 1990:136), the 'stress-timing effect' (Zhang 1996), and so on.

3. Vowel classes

The 'pure' vowels of English[6] can be categorised into two classes, for which I will use the traditional labels 'tense' and 'lax'. The classes can be generated in at least three, apparently independent, ways:

- By the relative lengths of vowels in comparable, high prominence contexts: tense vowels are then longer than lax ones.
- By their phonotactics: tense vowels can appear in open syllables at the end of words, while lax vowels must be followed by a 'checking' consonant.
- By considering the distortion of the vocal tract in production: tense vowels are articulated with more 'extreme' deviation from its neutral position[7].

If these criteria are unconnected, it seems quite remarkable that each divides the inventory of vowels into classes with the same memberships. I have not seen attention drawn to this in the literature except by Lass (1976:39), and then from a different perspective. I will propose that there is a linkage that has not so far been understood[8].

4. A complex set of production correlates for obstruents

The production correlates for sounds such as /p/ or /b/ include:

- Contextual phonetic voicing that does not always correspond to a sound's phonological status as 'voiced' or 'voiceless'.
- Contextual gradation of aspiration for syllable initial plosives.
- Long lag and short lag voice onset times that also vary contextually. (Although Dutch shows a different pattern to English and German.)

---

[6] Unless otherwise specified, my examples and analysis will be based on RP. I have considered General American pronunciation, of course, but not tested my proposals in detail against Scottish English, for example.

[7] Munhall (2001:102) shows striking mid-sagittal images of a talker producing [i], [u] and [a], with a pharyngeal constriction for [a] very apparent.

[8] The only partial attempts to link them that I am aware of are (i) Abramson and Ren (1990:90), who put forward a tentative proposal to link length distinctions with articulation (that differs from earlier proposals about the timing of gestures which now do not appear to be valid), (ii) a historical explanation for the co-occurrence of length and quality distinctions in English in Hill (1936:16), and (iii) 'syllable cut' phonology, discussed in section 6.4.2.

It will be convenient to be able to refer to the second, third and fourth phenomena together. I shall call them the 'West Germanic phenomena', abbreviated to WGmPh. As a group, they are noteworthy in a number of ways:

- They are not present in the pronunciation of small children, but emerge at around the same time as a child starts to use stress-accent for syllable prominence.

- Some aspects of the WGmPh seem quite arbitrary: for example, the 'checking' constraint on English lax vowels (whose sound qualities alone adequately distinguish them from their tense counterparts). Other languages allow vowels with similar sound qualities to end words. Why, in the face of pressures from the adoption of foreign words and from the need to generate new words (for product names, for example), has there been no innovation in English away from this limiting phonological 'rule'?

- In other ways, the WGmPh complicate English in ways that do not seem necessary. For example, French speaking children produce the, "typical syllable structure and ...trailer timed rhythm of the French language," by age 1;5 (Konopczynski 1995:25). However, English speaking children have still not acquired its rhythm by age 4 (Grabe et al. 1999:1201).

  If languages are under evolutionary pressure to make them as 'learnable' as possible, why do the old native-speaker varieties of English persist with the complications of stress-timed rhythm, when other dialects of the language – those used in the West Indies and Singapore, for example – demonstrate that it can be dispensed with?

It seems odd that some characteristics that are an additional burden on a child learning English and are not, it seems, systemically necessary, should be so resilient. How is it that they reliably reappear in each new generation of speakers, when we might imagine pressures for the language to evolve away from them? I hope that the answer to this question will be clear by the end of this part.

2.2 Replication by imitation (modelling)

It is widely believed that PFC and the WGmPh are replicated by imitation, meaning that a child comes to model them (even if he might start by mimicking the pronunciation of words where the phenomena appear). This would involve:

1. Observation of the durational patterning in others' speech (and/or his own mimicked versions of words and phrases). He will need to note how units of different sizes change length (sizes that are segmental, sub-segmental and supra-segmental), by how much, in what context, etc.
2. Analysis (probably) of variable data of this kind from a range of speakers.
3. Synthesis of a model (or rules) of the processes involved.
4. Adaptation of the model to his own speech production capabilities.

2.3 Outline of Part 1

I will argue that PFC, the WGmPh and some other phenomena are not imitated, but emerge as a result of a set of processes arising from the embodiment of speech.

The features of child production that will be of significance in my account include the following:

- Children do not start to speak knowing how to create the aerodynamic driving forces needed for extended periods of speech production. They have to learn the complex sensorimotor skills involved, at the same time as learning to say words. As with other motor skills, the initial forms of speech breathing are very different from the smooth, final form we see in older speakers.
- Children have greatly reduced lung volumes and smaller airways than adults. But they use higher subglottal pressures and only moderately reduced airflow. There is, therefore, an asymmetry in the scaling of the variables of speech aerodynamics in the child compared to the adult model.
- Children's physiology means that their speech breathing is only assisted by pressure generated from elastic recoil to a minor extent if at all (Netsell et al. 1994). The adult model of speech breathing is probably not possible until after 7 years of age.

I will use the phrase 'breath stream dynamics' (BSD) to cover these and related aspects of speech production. The term was used by Peterson (1957), and by Rothenberg (1968)

who explained that it describes, "the space-time distribution of the energy associated with air flow in the vocal tract."

In the next chapter, I will develop a model of pulsatile speech breathing in young children, proposing that this creates a frame into which activity of the upper articulators must fit.

In chapter 4, I will apply this to syllable production, to demonstrate that PFC may appear as the result of the distribution of limited aerodynamic resource rather than the imitation of timing relationships. The timing changes we observe would then be epiphenomenal.

In chapter 5, I will consider how a child implements stress-accent, an important feature of West Germanic languages. Stress-accent as defined by Beckman (1986:1), "differs from non-stress accent in that it uses to a greater extent material other than pitch." In particular, it makes stressed syllables in English longer and louder. West Germanic languages are considered to use stress-accent to achieve rhythmic prominence, while Japanese, for example, uses pitch accent for this.

I will argue that stress-accent is probably implemented by English speaking children with active respiratory system gestures, so that their speech breathing continues to be pulsatile for some time.

In chapter 6, I will reapply the ideas of chapter 4 to explain foot level shortening. Then I will reason that the pulses of extra effort a child must produce for stress-accent prominence will heighten the pressures and/or flows of air in the child-size vocal tract. The consequences would be potentially harmful to vowel production, so a child makes various accommodations which come to characterise vowels as being tense or lax.

A third application of BSD ideas explains VOT, aspiration and, perhaps, the fortis/lenis distinction. Finally, I will show how some other phenomena can be explained this way.

In chapter 7, I step back from these issues to discuss some aspects of imitation that are common to both parts of this thesis. I then criticise the conventional idea that the speech phenomena I have discussed are learnt by imitation.

In chapter 8, I draw some conclusions, and consider some further points relating to the issues I have been discussing.

Part 1 is summarised at the end of the thesis, in chapter 16.

3 Speech breathing in adults and children

The purpose of this chapter is to arrive at a model for a young child's style of speech breathing (SB). The respiratory system delivers a smooth and stable supply of power for adult speech. I will consider whether or not this is also the case for children, asking how a system of power supply that must be created while speech itself is being learnt could be organised and controlled.

Definitions of SB are given by Hixon and Hoit (2005), Netsell (reported by Solomon and Charron 1998:61), and Von Euler (1982). I consider it to be the coordinated actions of the muscles that change lung and airway volume with the muscles that have a valving effect on the vocal tract, in order to produce the aerodynamic conditions required for speech.

Some of the terminology used to discuss SB is problematic, but the niceties of this are not important for what follows. So, for example, I will use 'chest wall system' and 'respiratory system' interchangeably.

3.1 Speech breathing in adults

Hixon and Hoit (2005) give a good introduction to SB. Hixon (1987) is a comprehensive collection of articles, including reprints of Hixon (1973; 1976).

As background, I would note the following:

- The main interest that phoneticians have had in SB is its possible role in the production of stress-accent. I discuss this in chapter 5.
- As Scully (1990:112) points out, it is not possible to isolate SB as the 'responsibility' of just the respiratory system: speech breathing is an activity of the whole vocal tract. It is only possible to understand the moment by moment actions of the muscles of the chest wall in the context provided by the resistances to flow created by structures downstream of the trachea.
- Measurements of either pressure or flow in isolation are not very useful in understanding how SB functions. Since they are emergent properties of the system they must be measured simultaneously (and at appropriate points) if they are to throw light on the underlying activities that ultimately produce them, including changes in respiratory drive.

- Hixon and Weismer (1995) have demonstrated that the classic experiments conducted by Ladefoged and his colleagues in Edinburgh (summarised in Ladefoged 1967) can no longer be taken as a reliable guide to SB.

The most important point for what follows, though, is that SB is a very complex motor skill. The alveolar (lung) air pressure (P_{alv}) is partly the result of pressures generated from distended body tissue (as in a balloon with an elastic rubber skin). The volume of the lungs changes continuously while we speak, and with it the non-volitional contribution to P_{alv} made by this 'relaxation pressure'. So there is a need for constant adjustment of the volitional contribution made by the chest wall musculature if subglottal pressure (P_{sg}) is required to remain constant. At the same time, the load seen by the respiratory system is constantly changing as a speaker alters the resistance of the vocal tract through changing glottal and oral articulations, also requiring adjustments.

As a result,

> "... there is great complexity in the events of respiration during speech, complexity that equals and in many respects surpasses that of events in other parts of the speaking machinery." Hixon (1973)

> "When the complexities of airway resistance generated by the articulators, coupled with the complexities of the laryngeal resistance during conversational speech are considered, the interactions between muscular effort, air flow, airway resistance, and subglottal pressure virtually defy description." Zemlin (1988:92)

3.2 Speech breathing in children

Despite its complexity, adult SB is performed smoothly and usually without a need for conscious attention.

However, it is a motor skill that has to be learnt, and children's SB reflects the difficulty of the task they face. Boliek et al. (1997) make the following observations about 2- and 3-year-olds (who will, of course, have been vocalising for many months already):

> "Data from this investigation showed substantial variability. Breathing adjustments for vocalisation were not stereotypical either within or between subjects ... [who] demonstrated a variety of ways to achieve the aeromechanical drive required for vocalisation. This could reflect experimentation by these subjects to control a continuously changing breathing apparatus, while simultaneously learning and exploring more complex phonetic productions.

> Lung volume excursions were accomplished through an extremely wide range of rib cage and abdomen contributions. This wide range of behaviours might reflect the "trying out" of various chest wall displacement patterns and their impact on vocalisation.
>
> This investigation clearly demonstrated that the motor-skill refinement process has a relatively long developmental course.
>
> [These and other findings] suggest that the emergence and refinement of oral communication involves the integration and coordination of many sub-systems. The assembly of speech gestures has not resolved into stereotypic patterns by age 3 years. Furthermore, consistent behaviour observed in 7-year old subjects is clearly not achieved by 3 years of age."

SB is clearly a significant challenge for the young speaker. Solomon and Charron (1998) summarise other aspects of what is known about its developmental path, warning, however, about some inconsistencies between studies, particularly with respect to normative values for airflow. It seems, though, that children initiate speech at higher lung volume levels than adults, and tend to speak,

- on greater P_{sg} (which may be related to their tendency to speak louder);
- using a greater proportion of their vital capacity (VC);
- with lower average and peak air flows.

There is less data about comparative lung volumes and capacities during the period of speech acquisition than might be expected. The respiratory system of a child is, though, very much smaller than an adult one. See graph 6 from Stathopoulos (2000), redrawn on the left in my figure 3-1, for some illustrative data.

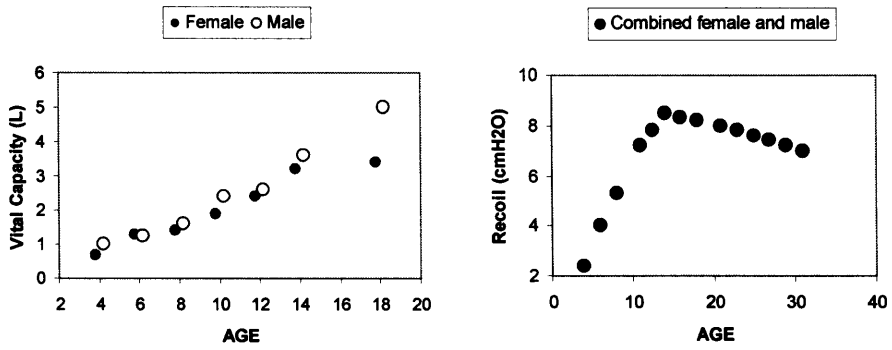

Figure 3-1. Redrawn from Stathopoulos (2000) figures 6 and 7. Original captions: "Vital capacity" and "Static recoil pressure".

The % of VC expended by a child on each syllable is an important measure, since this will affect the length of utterance that he can attempt. Hoit et al. (1990) found 7-year-olds expending approximately twice the %VC/syllable of 16-year-olds. For young children the relative value is hard to measure accurately, but will certainly be higher still (Boliek et al. 1997:387). See table 6-1 for numerical data.

The right hand graph of figure 3-1 shows the relaxation (recoil) pressure generated by speakers of various ages[9]. Children's relaxation pressures are lower because their lung and chest wall compliances are greater than those of adults (Russell and Stathopoulos 1988:146).

Netsell et al. (1994) considered two further aspects of young children's SB. The first was the pattern of net respiratory muscle pressure (P_{mus}) applied over typical lung volume speaking ranges. The three groups of subjects in this investigation were just under 4 years of age, just over 7 years and young adults. Netsell and his colleagues combined data collected on the subjects' subglottal pressure and airflow during a speech task, with information from the literature about the lung mechanics of children and adults. In this way, they produced hypothetical functions for the volitional contribution speakers must make to supplement inherent relaxation pressures.

These functions showed that as lung volume decreased over an utterance in a normal range from 54 to 30 %VC, the adults would begin speaking with relaxation pressure greater than that required for normal speech. They therefore needed to apply net inspiratory muscle pressure, so-called 'inspiratory checking', to reduce this to the pressure required. Their volitional contribution changed to net expiratory muscle pressure as lung volume fell.

7-year-old children were shown to typically begin speaking with relaxation pressure less than (but close to) the pressure they require for speech. They would therefore be expected to use predominantly net expiratory muscle pressure, except on the occasions where they inhaled to a greater than usual lung volume[10].

---

[9] Although not stated, the data points presumably relate to lung volumes at the end of a normal inspiration.

[10] Curiously, Russell and Stathopoulos (1988:152) found that 9-year-old children use a range of VC that starts and ends lower than adults, i.e. *contra* the trend for younger children reported by Solomon and Charron above. This finding was rather unexpected. The lower relaxation pressures supplied by a child's

The relaxation pressures of 4-year-olds **always** fall short of what children at this age require for speech. They were shown to use only net expiratory muscle pressure.

So, during the first 4 years of life, relaxation pressures at the beginning and over the course of an utterance form a relatively small part of the pressure that children must generate for speech. This will have implications for their style of SB and, perhaps, for the phenomenon of declination (section 6.4.5).

Netsell et al. also analysed the expiratory work of SB. Reworking their calculations with more recent data (for air volume expired per syllable), suggests that 4-year-olds are working at least four times as hard as adults on this aspect of their speech production. For children younger than this, the multiple would rise[11].

The comparative work of breathing can be approached from another hypothetical angle. Catford (1977:83) shows a graph of the relationship between volume-velocity (airflow), channel areas and 'initiator-power'. It shows that to move air through a channel at a given volume-velocity while reducing the cross-sectional area by a half requires a fourfold increase in power. On the other hand, reducing the volume velocity of an airflow by a half when channel area remains constant reduces the power required by a factor of one eighth. (In other words, square and cube laws are opposed.)

In very rough and ready terms, a 4-year-old might have a glottal area one quarter that of an adult, based on data for adult female vocal fold length in Goldstein (1980:65). This proportion would be smaller if the comparison were made with adult males, and significantly smaller if based on membranous vocal fold length (see Stathopoulos 2000). The child's typical airflow may be half male adult values (Netsell et al. 1994;

---

physiology in comparison to an adult's at an equivalent %VC had led them to expect that children would start speaking at a higher %VC, in order to generate appropriate passive P_{sg} for the initial part, at least, of their utterances.
Their figure 4 shows this effect across a variety of conditions. It raises the possibility that the 9-year-olds are avoiding moving into the region of lung volume where they might need to apply inspiratory checking. In other words, at this age children may still prefer the speech breathing strategy they have been used to: that of adding expiratory muscle pressure throughout an utterance. Younger children would initiate at a higher %VC because this gives them the benefit of increased relaxation pressure, but not to the extent that they need to apply inspiratory checking.

[11] Both this and the calculation that follows can only be approximate. There are many differences in the breath stream dynamics, physiology, vocal fold action, etc of children and adults, making comparisons of this type difficult.

Stathopoulos and Sapienza 1993) although possibly similar to that of adult females. If the child is half the height of an adult at this age, he may have a volume of muscle tissue that is one eighth of the adult's[12]. Taking the most conservative of these figures suggests that the power developed by the child to achieve a given airflow in speech must be double that of an adult, and that this is created with very significantly less musculature.

Both of these calculations indicate that the relative work being done by a child to create the aerodynamic conditions for speech is much greater than that done by an adult, a conclusion also reached by Titze (1994:180). The disparity will be greater for children younger than 4 years.

Learning to control his respiratory system may be the single most difficult motor challenge for speech faced by a child, with a correspondingly long apprenticeship. In his review of Faber and Best (1994), Hewson (1998) writes:

> "They present a great variety of evidence to show that the motor skills of speech develop with maturation and practice, in much the same way as walking skills, where the uncertain gait of an 18-month-old may be compared with the dancing and skipping feet of a four year old."

During the period of early speech, a child's SB is probably closer to "uncertain gait" than to "dancing and skipping". Because this developmental change is not as apparent as the changes in oral articulatory skills neither it nor its implications are usually considered.

In summary, there are four aspects of SB in children I would draw attention to:

- SB is a motor skill which has to be learnt at the same time as a child is learning to speak. During early childhood it is clearly not stable and routine. It probably does not approach adult levels of performance until several years after speech has begun.
- Young children use expiratory muscles to generate pressures for speech throughout an utterance. The adult style of SB – inflating the lungs and then speaking, to some extent, on relaxation pressure – is not available to them.

[12] Netsell et al.'s calculation normalised for different body size in a different way, by using %VC per syllable as their measure of volume change.

- The aerodynamics of child speech (pressures and flows) is neither the same as that of adults, nor a uniform scaling of it.
- SB is significantly harder (physical) work for young speakers than for adults.

3.3 Style of adult speech breathing

I will consider stress-accent and the controversy over its nature and how it is created in chapter 5. However, I need to briefly preview one aspect of this to describe the possible ways in which the respiratory system can be controlled for speaking English.

Ohala (1990) describes a relatively loose coupling between the pulmonary system and the upper articulators. He acknowledges that extra expiratory effort is made for emphatically stressed syllables. However, he describes the rises in P_{sg} that are observed for routine stress as being the result of increases in downstream resistance rather than increased respiratory drive (i.e., that they are 'back pressures'). So even for English, many aerodynamic models only assign simple overall goals to the respiratory system. These can include the maintenance of constant P_{sg} (e.g. Scully 1990), of constant lung volume decrement or of constant net force applied to the lungs (Ohala 1976).

An alternative view (e.g. Finnegan et al. 2000) has the respiratory system involved in the production of routine stress. In this case, the control of SB can be described as, "a volume solution overlaid with a pulsatile solution" (Hixon 1973).

Existing data tell us what adult speakers actually do during normal speech. However, if we consider what the range of legitimate control strategies for SB might be, then we must enlarge the solution space. In Svend Smith's Accent Method remedial therapy (Thyme-Frøkjær and Frøkjær-Jensen 2001), an active preferment of pumping muscle over laryngeal activity is taught. This style of SB also produces results that sound entirely natural[13].

[13] I was fortunate to be able to attend a five day course on the Accent Method organised by the Speech and Language Therapy Department of Belfast City Hospital in 1997. The exercises proposed to us encouraged me to adjust the balance I had previously struck between use of my abdominal muscles and use of my larynx when producing stress, so that I spoke with greater chest wall system contribution (or, perhaps, 'support'). I discovered that this new style of speech breathing did not in any way mark my speech output unnaturally. There was, however, a noticeable improvement in the pleasantness of my voice. I imagine that this effect is analogous to the one that the singers investigated by Schutte (1984) were aiming for. They preferred to sing 'inefficiently' (from a biomechanical standpoint) when this improved the timbre of their output.

Kneil's (1972) study bears directly on this. He set out with a focus on rate variation to discover, *inter alia*, if there are two modes of respiratory system activity, at low and high rates. (As this work is not easily available I have included some extended notes about it in Appendix A.)

He did, indeed, discover that respiratory system activity is reorganised as the rate of production increases. (As is seen with other motor skills: the move from walking to running, for example, and aspects of speech articulation (Boucher and Lamontagne 2001).) Figure 3-2 illustrates this. At a slow rate of repetition of a syllable train, the system returns to a relaxed position between the discrete 'pulses' that support each token. As the rate rises, a speaker starts to develop a higher background pressure level, from which only smaller excursions of pressure increases for each token are needed. Finally, the higher level is maintained continuously, and token-by-token activity is minimal.

I will describe the behaviour of P_{sg} in these three situations as excursions from ambient pressure, excursions from (partly elevated) background pressure, and elevated background pressure (EBP). In the first case we can imagine pulsatile muscle activity, and in the final case, smooth activity.

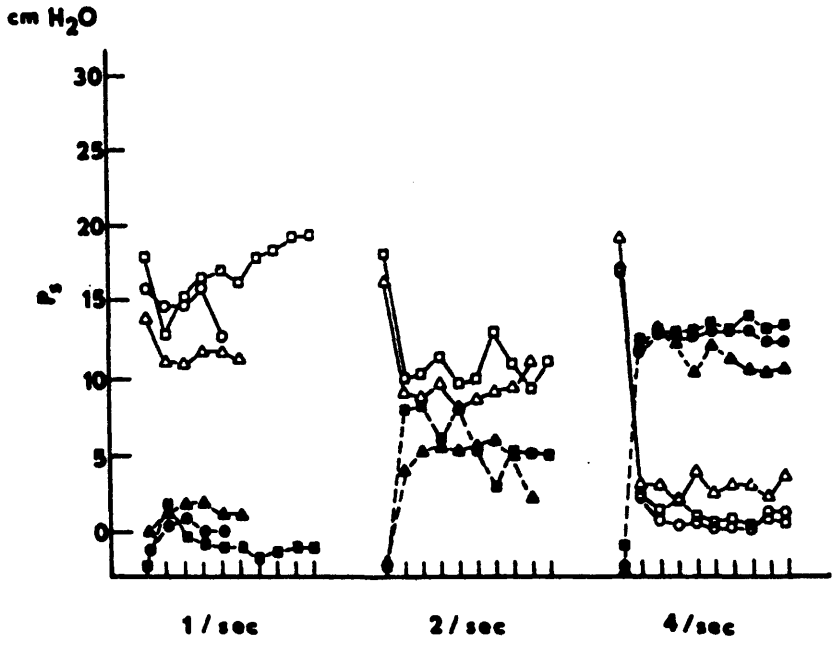

Figure 3-2. The evolution of elevated background pressure (EBP) with increasing rate. Figure 37 from Kneil (1972). Original caption: "Ps data: subject TK repeating /s/ at fixed rates." The unfilled symbols are the change in subglottal pressure; the filled symbols are the background level of subglottal pressure.

Kneil drew attention to the articulatory skill needed to achieve and maintain EBP (the coordination of valving activity by the larynx, the tongue, the lips, etc). He also demonstrated that the reorganisation of SB varies by syllable type. The high flow segment /h/ resists a move to EBP, presumably because of the extra demands it places on the respiratory system. However, it was clear from when subjects altered rates during syllable trains that EBP is the preferred mode of SB for adult speakers. It would seem to be highly compatible with the use of relaxation pressures.

It now seems reasonable to view the legitimate control strategies for SB in English described earlier along a continuum, as in Figure 3-3. Given Kneil's results, the example of the Accent Method, and a characterisation of the learning process for speech breathing as falling within a dynamical systems framework (Boliek et al. 1997), we should allow that performance will not be fixed at any one point on such a scale. It will vary both across individuals and, as circumstances change, within an individual.

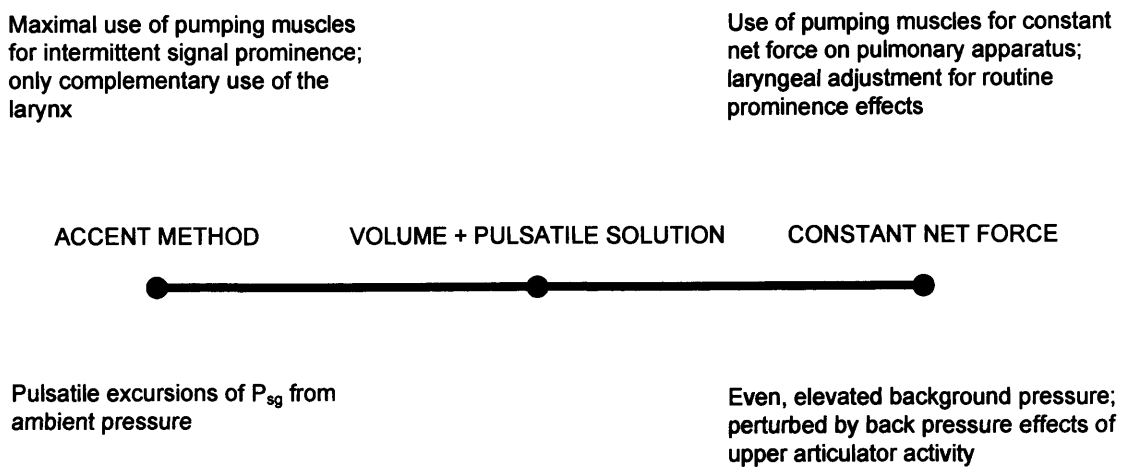

Figure 3-3. A simplified continuum of speech breathing styles, with strategies and sub-glottal pressure changes for a stress-accent language collapsed onto a single scale.

As described in the lower section of Figure 3-3, the subglottal pressure variation most naturally associated with the extreme left of the continuum would be excursion from ambient pressure; that associated with a central position, a partly elevated background pressure plus ripples caused by stress–related respiratory system activity, i.e. excursion from background pressure; and that associated with the right end, the same profile but with any (smaller) ripples explained through back pressures resulting from upper articulatory activity that increases resistance to airflow (EBP)[14].

For speakers of non-stress-accent languages the need for the transient changes in loudness associated with stress-accent prominence (Kochanski et al. 2005) is absent. The normal control strategy of mature speakers of French, Japanese, etc, can be assumed to be at the right hand end of the continuum.

For all languages, factors that will move the position of speakers' SB control along the continuum will include their rate of speech and their skill in maintaining EBP. Factors

[14] The associations made here may be broadly correct, but the diagram does simplify the real possibilities considerably. Even with pumping muscle activity causing variation in P_{alv}, P_{sg} might stay constant if R_{uaw} were varied appropriately in synchrony. Conversely (and as described) P_{sg} may vary due to back pressures if R_{uaw} varies even without any change in P_{alv}. (Although not by very much, if the Moon et al. (1993)/Finnegan et al. (1999) model is valid; if the respiratory system functions to an extent as a constant flow source then presumably this is more likely.)
Of course, this whole discussion presupposes that the aerodynamics of the vocal tract can be approximated with simple circuit theory. The larynx may, though, behave as a much more complex circuit element, as found by Nasri et al. (1996) in dogs (albeit at higher pressures than those seen in human speech). If so, then some of the inferences about P_{alv}/pumping muscle activity that have been drawn in the past from P_{sg} data may be unreliable.

more specific to speakers of stress-accent languages will include the degree of prominence required and the speaker's skill in use of a laryngeal mechanism for loudness variation. These are discussed further in chapter 5.

3.4 Style of child speech breathing

Infants may start speech with some experience of achieving EBP during activities like cooing, where the resistance of the vocal tract does not continuously or dramatically change in the way it does during speech.

However, the general picture of motor skill acquisition is for a progression from an activity being performed in discrete steps ('jerkily') to a smoothly integrated sequence. For this reason, and since the first word-like utterances of young learners would typically require just one respiratory system gesture, I would suggest that the starting point for their style of SB is located at the left end of the scale in figure 3-3.

I have three further observations in relation to this:

1. Moving rightwards on the scale is likely to take some time. To begin with, children's articulation rate is slower than adults'. Also, although their absolute flow rates may be lower, the % of lung volume decrement associated with any segment is higher. So all of their syllables may be relatively 'high flow', in the same way as a syllable beginning with an /h/ is for an adult.

2. A child will have to learn to coordinate and balance the pumping and valving activities for EBP. This will take time. If it happens during early production, then it will be further complicated by the instability of the other subsystems which form the complete skill of speech. (A pulsatile style of SB, in contrast, is a simple way to achieve a given vocal output. At a time when there are many demands on a child's attentional resource, it seems unlikely to be a priority for him to improve a part of his system that is already adequate.)

3. Finally, at low lung volumes adults have to add expiratory muscle pressure. They then use chest wall gestures of greater magnitude and are less likely to be in EBP mode. The mechanics of child speech mean that children are always

adding expiratory muscle pressure, so they, too, may find it hard to achieve EBP in this situation.

So I will assume that to control SB, children use pumping muscles in a pulsatile style both initially and while learning to concatenate words into simple phrases.

Kneil (1972) has shown that a smooth, EBP mode of control is ultimately preferred. The first indication that children are on the way to attainment of this in normal speech may be the impression of adult-like prosody given by French children at around 17 months (Konopczynski 1995)[15]. On the other hand, the deployment of stress-accent by child speakers of West Germanic languages after 18~24 months is characterised by the lengthening of the units over which respiratory system gestures apply and differential syllable prominence relations based on loudness and length. So the speech breathing of these young speakers will be pulsatile for significantly longer. This is discussed in chapter 5.

3.5 The nature of a pulsatile style of child speech breathing

I now take a motor skills perspective to explore the basis on which this pulsatility is controlled, and what determines the magnitude of the pulses.

3.5.1 Motor skill acquisition

We can expect both speech and SB to be acquired in a similar fashion to other motor skills, while remaining aware that the communicative characteristics of speech production and perception might also cause their development to be atypical in some ways.

As we have seen, Boliek et al. (1997) used a dynamical systems framework to explain the kinematic data they collected from their 1- to 3-year-old subjects. In this theoretical perspective, the SB of a young child is not assumed to be either (i) a scaled down version of the adult skill, or (ii) moving in a linear development from an immature starting point towards that model. Instead the skill is seen as a 'soft' assembly of sub-systems which are developing at different rates. The child finds successions of

[15] The resyllabification of codas to onsets in connected French speech may make underlying SB gestures more regular, and EBP therefore easier to achieve.

provisional solutions to the demands being made upon him, created with the resources he has at his disposal at any given time. As the characteristics of these sub-systems or the task change, so a new configuration may be adopted if this is better adapted to the new circumstances. Exploration and selection are important parts of this adaptation, so it may take some time for the child to discover new solutions[16].

Learning to breathe for speech involves the control of over two dozen muscles which contribute to the chest wall system alone (Hixon 1987). Bernstein's classic degrees of freedom problem – how to control a system with so many possibilities for action – will therefore present itself from the very start of vocalisation. It will then repeatedly present itself during speech acquisition proper whenever the child addresses a new aspect of the challenge. SB is, of course, only one of the motor skills which underpin speech, so it in turn will be a developing sub-system within the wider context of the child's speech development.

In general, learners of motor skills solve Bernstein's problem by 'locking' many of the possible interactions of joints and muscles and then tackling the simpler problem this now creates. Once a workable solution has been found, the possibilities for action and their consequences can be progressively explored, and constraints on coordination gradually relaxed on the way to smoother and more economical gestures (Schmuckler 1993:164).

Other features of motor skill acquisition include the following (adapted from Magill 1993):

- Changes take place in the movements of the learner's attention during the performance of the skill. It is only to the extent that foundational sub-skills can be made automatic that a learner becomes freer to attend to more subtle aspects of the challenge.
- There are identifiable stages of skill development. Authors vary in their analyses, but Gentile (1972) describes a first stage of "getting the idea of the movement", prior to a second stage in which the learner both works on (i) the capability of

[16] Thelen (1995) gives an introduction to motor development seen from this perspective.

achieving the goal regardless of the situation, and (ii) consistency in achieving the goal. Bril and Brenière (1993) describe stages of 'assembly' and 'tuning'.

- Changes can be observed in how the goal of the skill is achieved (the criteria used for control). So early in practice a learner may focus on, say, the spatial components of a task while later a learner may refine his skill by attending to the accuracy and consistency of velocity and acceleration components (Marteniuk and Romanow 1983). Another way of describing this is as a 'structural displacement' of the resources used for task performance (Ivry 1996:286).

3.5.2 The control of pulses

I have proposed that SB in early childhood is pulsatile. In this subsection I will discuss what aspects of SB might be controlled. In the next I will ask what goals the child may have for SB in terms of what he controls.

Before L1, a child has already been babbling for some time. He will have developed forward and inverse models[17] of the relationship between SB, laryngeal activity, some movements of his upper articulators, and what he hears of his sound output. To some extent he will be able to judge his SB 'by ear': he will be able to tell this way if his SB has not contributed appropriately to what he says, and to adjust it accordingly.

However, to achieve this, there are several aspects of his SB that he might attend to. One possibility is the effort he makes.

The sense of effort is believed to derive from a centrally generated signal related to the descending, efferent motor command. Thus an increased effort is assumed to reflect an increasing drive to motoneurons involved in performing a task (Solomon et al. 2002). The exact mechanism is still poorly understood and it is presently impossible to measure directly (Killian and Gandevia 1996), but the concept is familiar and effort is used via self-report against psychophysical scales in respiratory and exercise physiology (e.g. Borg 1998).

In a healthy adult, SB is so well practised and taxes us so lightly that we are not usually aware of the respiratory system effort we make during conversational speech. However, when the system is stressed this effort becomes apparent. It seems likely that a child will

[17] See section 10.4.2.

be more sensitive to the effort he makes for his normal SB. In section 3.2, I suggested that the relative work of SB for a 4-year-old versus an adult would differ by a factor of at least 4, with this disparity increasing further for a younger child.

Apart from effort, a young child might also attend to muscle tension, muscle fatigue, metabolic energy expenditure and, perhaps, 'power'[18] (calculated in some way from the output of pressure and flow receptors[19]). Effort is likely to be so much more apparent and accessible than any of these that it seems reasonable to assume that it is the primary informant for control[20].

3.5.3 The magnitude of stress pulses

The typical syllables produced by a young speaker are in CV, VC and CVC formats. The development of more complex initial and final consonant sequences may start as young as 2 years, but McLeod et al. (2001:104) report 'ages of acquisition' to start at 3;6 and extend to 8;0 for the most complex clusters. (A variety of devices are used by children prior to this to avoid tackling sequences that are beyond them.) So early syllables[21] will not vary greatly in the time they take to produce.

Given this, is it plausible for us to imagine that a child takes account of the composition of a syllable when judging the SB effort required to produce it?

It seems implausible to me. Firstly, to do so he would have to have a clear model for how his own production of the syllable 'should' sound. Although some researchers have asserted this, I explain in Part 2 of this thesis why I think it is unlikely.

---

[18] Catford (1977:80) describes a model of SB based on 'initiator power', which he defines as the product of pressure and flow. He uses this to explain certain features of English. Although I will be making very similar proposals to his, it doesn't seem likely that these aerodynamic variables would be controlled by a young child in the way he describes them being used by adult speakers.
However, while 'initiator-power' may not reflect exactly how speech aerodynamics is controlled by children, the aerodynamic output may sometimes be a reasonable proxy for experimental investigation of effort.

[19] Moosavi et al. (2000) are unable to resolve whether or not we can perceptually distinguish aerodynamic work and effort with respect to the respiratory system.

[20] I note that Rothenberg (1968:8) recognised the desirability of representing a neural innervation function in modelling speech breathing, presumably because he considered that it is a realistic parameter of control. Effort would be the perceptual correlate of such a modelling function.

[21] In chapter 6, I am going to dissociate syllables as conventionally defined using acoustic/articulatory characteristics, from units of production that the respiratory system might 'see'. At this stage, therefore, I am disregarding weak syllables in the proposals I am making.

Secondly, the result would be an extremely complex system of control. At a time when speech presents innumerable challenges, it is surely more likely that he starts with a default strategy of applying the same effort to all types of syllable, containing whatever 'segments'.

This strategy would be modified, of course, if he discovered that the result was unsatisfactory. It may be that the breathy onset to a vowel, /h/, would create this problem (it appears relatively early in production inventories (Yavas 1998:131)). If so, then I could imagine extra respiratory system effort being applied for syllables with this onset.

However, in general and as a summary of the main point of this chapter, I suggest that a child will create a 'frame and content' relationship between the pulsatile activity of his respiratory system and the upper articulator activity that creates the segmental contents of a syllable. The size of a pulse will be determined by how loudly the child wants to speak, his affective state, and so on; but not by the contents of the frame. A pulse will be controlled by the effort made on it, a readily available percept.

This relationship is, above all, simple to learn. For this reason, as we have seen, it is normal in the early stages of motor skill acquisition, and also in speech structures[22].

Later on, as the 'joints' of speech are unlocked, SB may become integrated with, and responsive to, upper articulator activity (Bucella et al. 2000). But at the start, speech will be learnt as other motor skills are learnt: by rough and ready assembly of a working system, using, in the case of SB, the most straightforward form of control compatible with successful output.

[22] See the proposals of MacNeilage and Davis described in section 10.4.3. Analogously, MacNeilage (1998) quotes Levelt (1992:10): "Probably the most fundamental insight from modern speech error research is that a word's skeleton or frame and its segmental content are independently generated." See also Fowler et al. (1980:386) and Fowler (1996:527) for discussion and examples of 'coarse-grained' processes constraining 'fine-grained' ones.

4 Pre-fortis clipping (PFC)

4.1 Introduction

In English, vowels shorten before phonologically voiceless consonants. I shall describe this phenomenon as 'pre-fortis clipping' (PFC), but it has other names including 'voicing conditioned vowel duration', 'extrinsic vowel duration' and 'the vowel length effect'.

Gimson (1989:155) describes PFC in more detail:

> "It is a feature of RP that syllables closed by fortis consonants are considerably shorter than those which are open, or closed by a lenis consonant. We have seen in the chapter on vowels that this variation of length is particularly noticeable when the syllable contains a 'long' vowel or diphthong, cf. the fully long vowels or diphthongs in *robe, heard, league* (closed by lenis /b, d, g/) with the reduced values in *rope, hurt, leak* (closed by fortis /p, t, k/). Preceding consonants, notably /l, n, m/, are also shortened by a following /p, t, or k/, especially when the consonants are themselves preceded by a short vowel, e.g. compare the relatively long /l/ in *killed, Elbe*, /n/ in *wand*, and /m/ in *symbol* with the reduced varieties in *kilt, help, want, simple*."

As with all the phonetic phenomena I discuss in this part, PFC needs a more extensive introduction than I can give it here. But for the adult skill, I note the following:

- PFC is significant for primary stressed syllables, less so for secondary stressed ones and negligible or absent in unstressed syllables (De Jong 1991).
- It is clear in citation forms, but often absent in normal speech (Crystal and House 1988).
- PFC affects the /e/ and /n/ of "tent", and the /ɪ/ and /l/ of "milk" (Wells 1990:136).
- It occurs in esophageal speech (Gandour et al. 1980). It also occurs in whisper (Sharf 1964), which has sometimes been taken as evidence that the phenomenon is systemic/linguistic rather than 'physiological'. However, Hamlet (1972) and others have shown that the glottis narrows in whispered lenis consonants compared to fortis ones, so the aerodynamics of the two situations are not neutralised. (Sharf himself concluded that linguistic structure is at least a perpetuating factor in vowel duration variation, but that this does not rule out the possibility of physiology as a precipitating factor.)
- PFC is almost a language universal (in languages allowing syllable codas that have not been neutralised with respect to voicing). Polish, Czech, Swedish and Saudi

Arabic are the known exceptions (Keating 1984; Buder and Stoel-Gammon 2002; Mitleb 1984).

- Lisker (1974) summarised the explanations for PFC that were then current, and dismissed one based on a rule of constant energy expenditure for the syllable. However, it had been framed in terms of upper articulator activity; SB was not considered.
- Laeufer (1992) analysed French and English to demonstrate that the degree of PFC in each is comparable. Previously it had been thought to be more marked in English (e.g. Chen 1970), implying it has a phonological basis.

As I described in section 2.2, Naeser (1970) found that the speech of children learning English showed PFC from at least 22 months of age. It has seemed that young children imitate (i.e. model) a temporal feature of adult speech even at this early stage. In this chapter I will suggest an alternative to this: that the breath stream dynamics (BSD) of children's speech canalise their output to create this effect, but that it is epiphenomenal. In reality the child is just balancing SB and upper articulator activity, and PFC emerges as a by-product of this.

4.2 A breath stream dynamic (BSD) explanation for PFC

In the previous chapter I proposed a developmental model of SB. In this, the pulsatile activity of a young child's respiratory system creates a syllabic frame which is independent from the upper articulator activity which provides the contents. The magnitude of the pulse (controlled via the percept of effort), is invariant with respect to the particular 'segments' being produced in the syllable. (I qualified this proposed invariance in one way, with respect to high flow onsets.)

From this, a simple explanation for PFC emerges. If, in otherwise comparable syllables, the final segment of one syllable requires more aerodynamic resource for its production than its counterpart in the other, then that resource will be taken from the other segments, particularly the preceding vocalic nucleus. The result may be a shortening of the duration of those segments.

PFC is usually studied with respect to final stops, since these create convenient landmarks for measurement in an acoustic record. Here I will first illustrate my proposal with final fricatives, since these demonstrate its workings more clearly.

4.3 PFC before final fricatives

Compare the aerodynamic resource required to produce the final sibilants of *peace* [p^{h}is] and *peas* [p^{h}i:z]. Hogan and Rozsypal (1980:1768) demonstrate that PFC will occur in this context. Both instrumentally (Stevens et al. 1992) and introspectively (O'Connor 1973:38-40), it is apparent that the [s] requires greater respiratory system activity than the [z] in order to produce a perceptible output from its single, turbulent sound source.

C.Smith (1997) shows that even if a final /z/ is devoiced, its aerodynamics are similar to a voiced one rather than an /s/, and PFC occurs with both types of /z/.

My proposal is that the additional aerodynamic resource required to distinguish an [s] from a [z] or an open syllable (*pea*), leads a child to redistribute the (invariant) resource he has available. More is devoted to the [s], less to the other elements of the syllable, particularly the vocalic nucleus[23]. Without any conscious intention, as would be needed if the vowel duration were being imitated, the vowel is shortened (or 'clipped').

This epiphenomenal effect is illustrated in (a) and (b) of figure 4-1.

[23] In adults, a little less is typically devoted to the onset, explaining the otherwise puzzling durational results reported by Weismer (1979), who found a reliable shortening of VOT's in the initial consonants of syllables closed by fortis rather than lenis tokens. The degree of shortening, was below a normal perceptible limen, making it hard to reconcile with an imitative account of PFC. (This explanation, however, relies on the account I shall give of VOT in section 6.4.)

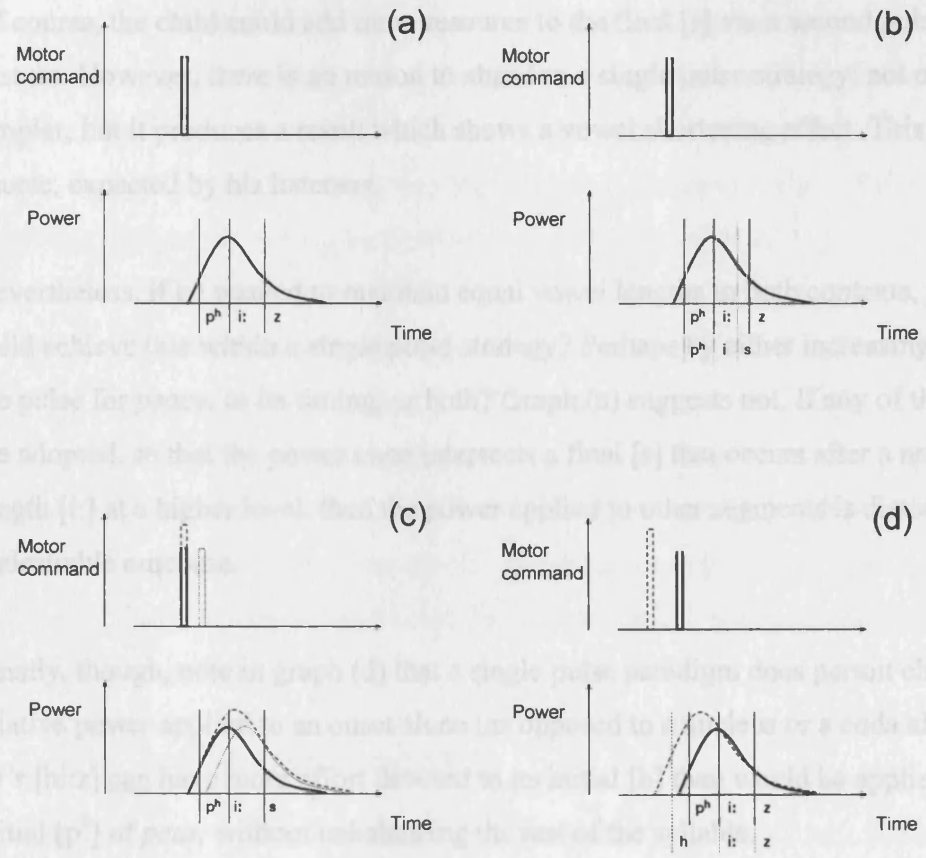

Figure 4-1. A simple control strategy for respiratory system output, and its perceptual consequences. See text for explanation.

In (a), imagine a single impulsive signal from the central nervous system sent to the respiratory system (considered to be operating as a coordinative structure[24]). Since this process will have considerable inherent damping, the result is a pulse of power production extended over time (cf. Fujisaki 1993). P-centre research suggests that the segments will then be arranged so that the CV boundary aligns with the power peak.

In (b), the child speaker reserves more resource for final [s] by reducing the allocation to the nucleus of *peace* as compared to *peas*. Given his strategy of keeping respiratory system and upper articulator activities decoupled, he achieves this by reducing the duration of the vowel (starting final consonant production sooner).

[24] "A coordinative structure can be described as a set of constraints between muscles that are set up to make the set of muscles behave as a unit. Thus, control is simplified in the sense that the individual muscles need not be controlled independently of each other but rather as a functional unit." (Löfqvist 1997:415).

Of course, the child could add more resource to the final [s] via a second pulsatile gesture. However, there is no reason to abandon a single pulse strategy: not only is it simpler, but it produces a result which shows a vowel shortening effect. This is, of course, expected by his listeners.

Nevertheless, if he wanted to maintain equal vowel lengths in both contexts, could a child achieve this within a single pulse strategy? Perhaps by either increasing the size of the pulse for *peace*, or its timing, or both? Graph (c) suggests not. If any of these tactics are adopted, so that the power trace intersects a final [s] that occurs after a normal length [i:] at a higher level, then the power applied to other segments is distorted, an undesirable outcome.

Finally, though, note in graph (d) that a single pulse paradigm does permit change to the relative power applied to an onset alone (as opposed to a nucleus or a coda alone). So *he's* [hi:z] can have more effort devoted to its initial [h] than would be applied to the initial [p^h] of *peas,* without unbalancing the rest of the syllable.

Figure 4-1 is only meant to be an illustrative sketch, and it is clearly unrealistic in some ways. There is an immediate problem in the relationship between the effort pulse I have shown as occurring prior to power output and the actual perception of a P-centre at the CV boundary. (In section 6.5, I will be suggesting that the latter is obtained from the former.)

However, it may not be so unrealistic in its restriction of simple pulse variation to amplitude and timing changes. Rosenbaum and Krist (1996:31), discussing Gottlieb et al. (1989), speculate that pulse height modulation is an open-loop or predictive mode of control. It would be simpler to implement than a closed-loop pulse width control.

More recently, it has been proposed that, "the transition from pulse height to pulse width control is not the result of the implementation of different control schemes *per se*, but rather the relation between the force demands of the task and the biomechanical properties of the joint" (Khan et al. 1999:157). In either case, the preferred mode of control (where possible) is pulse height modulation. Pulse width modulation might be needed, therefore, where a steady state rather than a transient force must be applied.

An alternative to the account I have described might still have a child imitating the durational effect of PFC, with this giving him an appropriate resource for final consonant production. I.e., the aerodynamic requirement would not be shaping the output, even if (perhaps) in some way it precipitated the phenomenon of PFC in the origins of speech. Present day perpetuation would be by imitation.

I have already mentioned Weismer's (1979) results that might tell against this. In addition, Buder and Stoel-Gammon (2002) report 24-month-old Swedish children exhibiting PFC before, apparently, suppressing it 6 months later in order to properly express the phonemic vowel length of Swedish. This pattern of development is incompatible with an imitative account. (The suppression, though, might require Swedish children to have moved to a different mode of SB control, possibly including pulse width modulation.)

4.4 PFC before final stops

The presence or absence of voicing is the prototypical way of making a distinction to create two types of sibilant speech sounds. I have just described how if a young child discovers this within the developmental framework for SB I have described, then PFC is the natural result. By maintaining some difference in his glottal configuration between /z/ and /s/ even when he devoices the former, PFC is also seen in this circumstance.

(It is interesting that speakers do maintain such a difference. It suggests that they are not operating within a purely perceptually-driven model of their own speech; if they were, we might expect them to distinguish /s/ and /z/ by PFC alone, for example, producing [s] for /z/ in final position.)

The situation with final plosives is more complex. A child must find some way to maintain a three-way contrast, between $\_VC_F$, $\_VC_L$ and _V#.

In French this is straightforward. Taking *coûte*, *coude* and *cou* as examples, the environment will expect a young child to produce *coûte* with a final burst and to maintain voicing through a reasonable portion of the final segment of *coude* (Flege and Hillenbrand 1987; Laeufer 1996). Both these activities serve to distinguish *cou*, and the

aerodynamic requirements of the final burst in *coûte* are clearly greater than those of *coude*. So PFC is seen.

In English, a child must eventually contrast triplets with tense vowels such as *seat, seed* and *see,* and *cart, card* and *car*, and pairs with lax vowels such as *hit* and *hid.* This may not be so straightforward. Among adults speakers, voicing sometimes continues for a short period into final fortis stops (B.Smith 1979), and final lenis stops can be voiced, partially voiced or devoiced. Both fortis and lenis prepausal stops may be silently or audibly released, "more or less randomly" (Lisker 1999:44). Yet for my hypothesis on the cause of PFC to be valid, children's articulations must make clearly different aerodynamic demands in syllables closed by fortis and lenis stops.

As a first step in the discussion of this, I think we can draw aerodynamic distinctions between the audible releases after fortis and lenis stops. Locke (1983:102) describes, "a puff of air in the case of voiceless stops or a faint schwalike vowel in the case of their voiced cognates." For the former, heightened intraoral pressure will have been maintained during the closure interval. For the latter, it may be instead that the oral structures which were displaced to maintain transglottal airflow during closure now recoil, and in the process shunt a small amount of air through the opening constriction (Stevens 1997:331,346-347). Alternatively, if voicing has not been maintained, a percussive sound may be created by the opening, for example, of the lips (Pike 1943:103).

In either case, the noise following a lenis stop does not require any additional aerodynamic resource.

However, is more resource used by an English speaking child to produce final fortis stops? I do not know of any study that provides direct evidence on this issue. Of course it may be that young children are unlike adults: when PFC occurs they may have always ended the syllable with an audible release burst. If so, then the situation is analogous to French, and therefore straightforward. I note that Mack and Lieberman (1985:542) reported that for their single subject analysed at ~1½ and ~3 years old,

> "... there seemed to be a great deal of aspiration associated with [the] word-final stops [in *book* and *cat*] at both stages, as evidenced by their long durations."

Similarly, Paul-Brown and Yemi-Komshian (1988:633) found just 7% of final stops were unreleased by their 5½ year old subjects. However, these were isolated tokens.

However, it seems more likely that a child will not always release a fortis stop with an audible burst. If not, why might he nevertheless devote more aerodynamic resource to its production than to an unreleased lenis stop?

His needs his listeners to be able to distinguish words he produces: in this case with syllables that either end with a lenis stop or are unclosed. He may discover that they can do this based only on the formant transitions from the vowel into the consonant[25]. However, to create appropriate vowel transition cues, he has to re-create the aerodynamic conditions that would also produce a final burst: a strongly produced vowel abruptly terminated by an occlusion. He may find he can do this without subsequent audible release, or by glottalisation, but in both these cases – as with a released token – he will require greater aerodynamic resource than the lenis cognate. So PFC can be expected, and is seen, under all three circumstances.

If I am wrong, he may realise that it is possible to create a difference between his final stops by shortening the previous vowel, as conventional accounts of PFC would have it. I have given two reasons why I think this is unlikely (based on children learning Swedish, and the covert contrasts in initial consonant VOT found by Weismer (1979)). However, given that an unreleased fortis consonant has little to indicate its existence apart from the preceding vowel length and postvocalic transitions, it may be that a more convoluted developmental path than the one I am suggesting actually occurs[26].

4.5 Extensions and exceptions

This process I have described should apply wherever codas require different levels of resource. So it would explain the PFC of both vowel and nasal consonant in English *lent* and *lend*, for example.

---

[25] See Walsh and Parker (1984) for arguments in favour of vowel transition cues over vowel length, and Nittrouer (2005:352) for children's own perceptual preferences in this direction.

[26] Naeser (1970) suggests, for example, that children might initially view the vowel length changes they hear as phonemic, and imitate this durational distinction.
Blevins (2004:274) describes some such process as having led to Appalachian English now having phonemic word-length contrasts before final consonants that are now always devoiced: [be:t] and [bet] for *bed* and *bet*.
There is also some confusion about what cues children use perceptually (e.g. Greenlee 1980; Wardrup-Fruin and Peach 1984; Nittrouer et al. 2005:352), and a need to understand the input better (e.g. Bernstein Ratner 1984).

The principle is not limited to fortis/lenis contrasts. It would also explain an apparently distinct phenomenon: the 'compression' of vowels and consonants in the rhyme as segments are added to the coda of a syllable (Fowler 1983; Munhall et al. 1992). For example, as in *ram, ramp, ramped.*

The nature of this process suggests that it should be a language universal where contextual conditions permit it. This is not the case. However, of the four reported languages where PFC might be expected but is not seen, three (Swedish, Czech, Saudi Arabic) use vowel length phonemically. In Swedish, Buder and Stoel-Gammon (2002) found that PFC appears in child speech in the /i/-/ɪ/ contrast before being suppressed, presumably so as not to disrupt the vowel length contrast. They concluded that this supports the idea that, "duration is naturally correlated with final stop consonant voicing" (p.1862). It would be interesting to see if their results extend to other tense/lax vowel contrasts in Swedish, and to Czech and Arabic.

I have no explanation for the non-appearance of PFC in Polish (Keating 1984). There may be some significance in the fact that Polish historically made use of vowel length phonemically. However, it no longer does so.

4.6 Summary

I have described how pre-fortis clipping (PFC) would be a natural consequence of the model of developmental speech breathing (SB) that I proposed in chapter 3. Rather than modelling their output on adult speech, I have suggested that young children do what is necessary to make a final fortis consonant distinctive and that this requires more aerodynamic resource than the lenis cognate. Constrained to articulate a syllable with the same aerodynamic resource for both tokens, the child redistributes energy to the fortis segment away from earlier segments. One way that this can be manifest is in a shortening of the vowel.

While this explains PFC in a child using a pulsatile style of SB, how can it extend to an adult speaking on an elevated background pressure (EBP)? One possibility is that a timing relationship which appears in childhood with a non-temporal basis, is subsequently incorporated into the adult production model of timing control.

However, I prefer an alternative explanation. PFC is often not found in conversational speech yet reliably appears in a more careful style. Also, it is more significant in conditions of primary rather than secondary stress. So it appears under conditions where the adult style of SB might be expected to move leftwards on the continuum described in section 3.3, perhaps generating ripples or pulses overlaid on the 'volume solution' that again allow for a resource-based explanation for the phenomenon. Some evidence for this argument will be presented in the next chapter.

I do not have the space here to document my review of this proposal against the literature, but I have not found any evidence that would seriously threaten it. It seems to explain well the contextual variation that is observed.

There are a number of arguments that can be made against the plausibility of the conventional time-based modelling alternative for the emergence of PFC in a child's speech. Since these also apply to some of the accounts of phenomena to be discussed in chapter 6, I will defer presenting them until after that, in chapter 7. It may be worth noting here, though, that my proposal in no way denies the often demonstrated importance of PFC as a cue to the identity of the following consonant. It just denies that this is the motivation for at least the early emergence of the phenomenon.

5 Respiratory system involvement in the realisation of stress-accent

In chapter 3, I combined the views of Hixon and Ohala with the practice of the Accent Method to develop a continuum approach to styles of speech breathing (SB), which I then applied to pre-stress-accent speech in children.

However, Hixon and Ohala's ideas were developed for adult speech, and relate to a controversy about how the prominence created by stress[27] is produced by speakers of English and other West Germanic languages. For routine sentence stress, do speakers make a laryngeal adjustment, or is the respiratory system recruited?

According to Beckman (1986), quoted earlier in section 2.3, 'stress-accent' languages use duration, loudness and spectral variation to give syllables greater prominence. Increased loudness can arise from a 'purely' laryngeal adjustment because the slope of the trailing edge of the glottal air pulse - the maximum flow declination rate - seems to be a principal determinant of intensity. In theory this can be made steeper by increasing the stiffness of the vocal folds, by increasing respiratory drive or by a coordinated combination of both manoeuvres (Gauffin and Sundberg 1989; Sundberg et al. 1993; Alku et al. 1998)[28].

In the next chapter I will propose that various phonetic phenomena that characterise West Germanic languages can be explained if children implement routine stress with increased respiratory system activity. I have suggested that their initial speech is produced this way (with pulses of respiratory system activity), but if the adult model of SB is different then by the time children adopt stress-accent they might have already developed this style, in which case my proposals would not be valid.

---

[27] Where I use 'stress' I will be referring to actual prominence (a rhythmic beat), rather than the lexical property of potential for prominence. A stressed syllable is thus one, "marked by greater loudness than unstressed syllables, and often by pitch prominence, or greater duration, or more clearly defined vowel qualities" (Wells 1990:683). An 'accented' syllable will be one carrying a pitch movement.
The words are used in other ways, e.g. by van Heuven and Sluijter (1996:238), but one advantage of this (UCL) approach is that the term 'stress' conforms to its usage by non-phoneticians.

[28] For increased respiratory drive, I will be assuming that there is also an appropriate laryngeal adjustment (Kostyk and Putnam Rochet 1998). So RS activity does not increase in isolation but will be accompanied by a 'tuning' of the vocal folds. The nature of the 'laryngeal adjustment' mechanism, however, is that the vocal folds are stiffened to an extent that would not naturally be associated with a particular P_{sg} (cf. Nasri et al. 1994), and that this is done without change to the respiratory drive.

There is no direct evidence to decide this issue. I will start by looking at adult performance, and then discuss how children's SB may be similar or different to this.

5.1 Adults

There are three main lines of evidence bearing on respiratory system involvement in routine stress in adults: one from 'observers' ('introspective' phoneticians and teachers of phonetics and pronunciation), a second from instrumental investigation of the production system, and a third from acoustic studies of the speech output. I discussed this topic in Messum (2003), since when Jensen (2004) and Gordeeva (2005) have produced fuller reviews. What follows, then, only goes into minimal detail.

5.1.1 Evidence from observation

Since we are concerned with the speech production of a child learning English, it may be reasonable to draw parallels with second language learning. Pronunciation teachers get good results by teaching stress through active recruitment of students' chest wall musculature. It seems that this is an effective pedagogical route to mastery, to start with a 'rough and ready' version of the skill, whatever the final, 'tuned' form will take[29].

Jensen (2004:3-17), Yamamoto (1996:231) and Fox (2000:120) describe the views of numerous phoneticians who have described stress as the result of extra respiratory system activity. In addition, Sievers (1901) drew a distinction between *Drucksilben* and *Schallsilben* ('pressure' syllables and 'sonority' syllables) which seems similar to a conclusion I shall come to for child speech.

Even if we assumed (as I do not) that so many phoneticians' impressions of their own speech and that of others were unreliable, then we would still have to account for the source of these impressions. Presumably they would derive either from how the speaker would hyperarticulate or from echoes of the developmental path he has taken, or from both sources. Then this evidence would support the idea that a child realises stress with increased respiratory system activity when he first deploys stress.

---

[29] A student I taught in Metz in 1995, Georges Dussy, likened speaking English to blowing a trumpet, and speaking French to playing a clarinet. Using this notion seemed to strikingly improve his own pronunciation of English.
On the somatic differences between speaking the two languages, see also Gattegno's observations in Appendix B. My experience with the Accent Method, described in chapter 3, also seems relevant.

Among phoneticians, Catford has taken the aerodynamics of speech very seriously as a field of inquiry. He is quite clear that, "initiator power is the organic-aerodynamic phonetic correlate of what is often called 'stress'." (1977:84)

5.1.2 Evidence from instrumental investigation of the production system

Stetson (1951) claimed that every syllable is produced by a 'chest' pulse. (See my note of Kneil (1972) in Appendix A). Ladefoged and his colleagues (Ladefoged 1967) could not find electromyographic or pressure trace evidence to support this assertion, and pointed out that the rise in P_{sg} they observed on stressed syllables could have been due to increased upper airway resistance (but see the discussion of Finnegan's work below). Similarly, Adams (1979) could establish no reliable link between stress and chest wall system activity, as measured with EMG.

Ohala's experiments (Ohala 1976; Ohala et al. 1980) led to him model 'pulmonic system' activity as being no more than the application of constant net force over the course of an utterance:

> "[T]he primary function of the pulmonic system during speech is simply to produce P_{sg} that is reasonably constant and above some minimum level. Short term passive variations in lung volume decrement occur in reaction to variations in lung pressure which in turn are reactions to changing downstream resistance to airflow created by oral articulations. Active short-term variations in lung volume decrement are probably limited to the production of variations in the loudness of speech: lesser decrement for less intense speech, greater decrement for very loud speech.
>
> Claims that independent action of the respiratory system underlies the production of syllables, stress ... and certain segment types are called into question." (1990:39)

Sundberg (1995:101) comes to a similar conclusion to Ohala,

> "In neutral speech, void of emphatic stress, it seems sufficient to signal stress by F0 gestures and syllable duration while in emphatic and emotional speech also subglottal pressure is recruited."

Laryngeal adjustment is believed to be quicker and less costly to the speaker for the production of sentence stress than increasing the effort made by the respiratory system. Of course, all speakers are expected to produce emphatic stress by increased respiratory drive.

However, Finnegan et al. (1999; 2000) have cast doubt on the basis for these conclusions. Earlier, Moon et al. (1993) had developed an 'ideal pressure source' model to explain the stability of P_{sg} in the face of a change in upper airway resistance (R_{uaw}) that might have been expected to reduce it (caused by the presence of a bleed tube, or by velopharyngeal port inadequacy). They claimed this was a passive reflection of the pressure drop across resistances in series. The converse of this, however, is that when P_{sg} is observed to change, as often seen in stressed syllables, the model does not allow that a change in R_{uaw} has caused this.

So Finnegan et al. modelled the effect that a change in upper airway resistance (R_{uaw}) would have on the division of the pressure drop from alveolar pressure in the lung to atmospheric pressure at the mouth. Taking lower airway resistance (R_{law}) to be more or less invariant, they pointed out that during phonation the resistance at the larynx (which is effectively the totality of R_{uaw} during vowels) was already so much greater than R_{law} that to attempt to increase P_{sg} by increasing laryngeal resistance would require a greater increase than any observed drop in airflow has ever suggested occurs. They therefore concluded that P_{sg} is controlled by changes in alveolar pressure (P_{alv}), i.e. by the respiratory drive of the speaker. Thus any change in P_{sg} that we observe during stress will not, as Ladefoged and Ohala suggested, be the passive result of changing downstream conditions, but of an active respiratory system gesture.

A further important conclusion from this is that a speaker has some freedom to vary the resistance of the larynx without affecting P_{sg}, providing a useful means of independently regulating airflow. I return to this idea in section 6.2.4.

Finnegan et al. found that extra respiratory drive was responsible for both sustained and transient increases in intensity, i.e. for changing the overall loudness of a phrase but also for stressing syllables within a phrase.

Her conclusions are consistent with the views of experimentalists who have looked at the kinematics of SB. Thus Hixon (1973, 1987:47) described the action of the chest wall system during conversational speech as the combination of a 'volume solution' which sustains a constant P_{alv} as lung volume changes, with a 'pulsatile solution' that meets, "the frequent demands for rapid changes in muscular pressure."

5.1.3 Evidence from acoustic studies

Jensen (2004:3-17) documents the shift in research/theoretical perspective during the middle of the twentieth century from stress as subjectively felt speaker action ('breath force') to hearer perception. This led to the experimental work of Fry and others to determine which factors are responsible for creating a sensation of prominence. These seemed to demonstrate that fundamental frequency and durational changes were the principal correlates of stress, and that intensity was placed a rather poor third.

Recently these conclusions have been challenged on at least two grounds. Firstly, it had not been appreciated that intonational pitch movements should be seen as a separate phenomenon from stress (Ladd 1993:11). Secondly, earlier studies made a mistake in assuming that simple intensity variation could adequately represent loudness variation.

So, for example, Sluijter and van Heuven (1996; 1997) carefully controlled for the effects of accent lending pitch movements, and found that stressed syllables in Dutch have acoustic properties which are consistent with greater effort being used in their production. (But see Heldner (2003) on natural conversation.)

5.1.4 Summary

The three lines of evidence considered do not disallow that in conversational speech routine stress may be produced by laryngeal adjustments without increased respiratory drive. However, there is increasing evidence of respiratory system involvement and the continuum approach described in section 3.1 seems to reflect adult performance of routine stress in stress-accent languages.

5.2 Children

5.2.1 When is stress-accent deployed?

Kehoe et al. (1995) examined the acoustic correlates of stress in children from 1;6 to 2;6, focussing on fundamental frequency, duration and amplitude (and making no attempt to control for differences between accent and stress). They found that even 18-month-olds produce all the correlates. They noted that:

> "All subjects marked the stressed syllable with higher overall intensity than the unstressed syllable. There was a tendency for the intensity difference to increase with age ... Intensity has generally been underplayed as a cue in stress perception, but the

finding of age-related differences in intensity suggests that this parameter may deserve closer attention in stress measurement studies."

An acoustic study by Pollock et al. (1993) found that 2-year-olds only used duration to mark stress, while 3- and 4-year-olds used a combination of the same three qualities as Kehoe et al. found to be used at an even younger age. However Schwartz et al. (1996) supported Kehoe et al. in finding that 2-year-olds mark stress with intensity differences.

Goodell and Studdert-Kennedy (1993:712) found that on durational criteria, "the children's ability to make a stress contrast [had] become virtually adult-like" over the period from 22 to 32 months[30].

Stress in some form, then, is a feature of quite early language use in English speaking children. For many children, 18 months is around the end of the 'first words'/'50 word' mark, and at the beginning of the so-called 'vocabulary spurt'.

I will now continue with a discussion of how children might implement greater loudness for stress, but, of course, they may not perceive the correlates of stress in the individual way that we measure them. It is not essential to my argument, but it seems very possible that they directly retrieve increased effort on the part of other speakers as the signal of prominence, and interpret this as the result of heightened articulatory and respiratory system activity which they then deploy.

5.2.2 What style of speech breathing is used for stress?

Various features of child speech make it likely that a child's implementation of stress will have a significant component of respiratory system involvement, rather than a predominance of laryngeal adjustment in the absence of greater respiratory drive.

Previously, the child will presumably have developed a 'tuned' relationship between his respiratory system and his larynx, such that they move in step to produce greater loudness. Disregarding this adequate and available automatism in favour of a laryngeal mechanism for stress would require the development of a new relationship at a time when he has other demands on his attention. Further, both his larynx and his respiratory

[30] Davis et al. (2000) argue that the motor capacity for adult-like stress production is developed pre-linguistically.

system are developing physiologically (Beck 1997:263; Koenig 2000), and his SB is not yet stable (Boliek et al. 1997).

In fact, a laryngeal mechanism may not be possible at all. In discussing the experimental data reported in Stathopoulos and Sapienza (1993), Stathopoulos (1995:78) reports how children vary loudness:

> "One noticeable difference between children and adults is that [children] do not use their laryngeal mechanism in the same way to increase vocal intensity: they do not change their open quotient (increase the closed time of the vocal folds within each cycle) as do adults to increase their vocal intensity levels. [However], children do use the 'MFDR mechanism' which is closely tied to aerodynamic events. **Maximum flow declination rate, for children, may be controlled by the respiratory system alone – through increases in subglottal pressure.**[31]"

All researchers agree that adults implement emphatic stress with increased respiratory system activity. In chapter 3, I argued that SB is a considerably more effortful task for young speakers than for adults. 'Routine' stress for a child speaker may require a form of implementation that would be experienced as emphatic stress in an adult.

In chapter 4, I suggested that Swedish children might modify a pulsatile style of SB to overcome the pressure for pre-fortis clipping. Might an English-learning child do something similar?

This would be a process of self-supervised learning, and to do so he would therefore have to have a model for how his production should sound. Given that what he hears of himself will be so different from what he hears others produce (in rate, overall loudness, segmental balance, etc) it is hard to imagine him having such a model. But if he did, then Catford (1985:346) has shown how it would anyway be compatible with a single pulse form of implementation.

Finally, as I suggested in section 5.1.1, the intuition of native speakers, as embedded in pedagogical and phonetic traditions, suggests that stress is initially implemented by respiratory system activity even if an alternative form of production is developed later[32].

---

[31] See also Moore (2004:193) quoted in section 6.2.5.

[32] There is also something unrealistic about the baton of even routine stress (in a stress-accent language) being handed from generation to generation via its acoustic attributes. These attributes are not arbitrary,

It seems reasonable to conclude that a child's initial strategy for creating stress-accent prominence is to raise respiratory drive on stressed syllables. So at the time when the WGmPh are being acquired, the forces developed by the activity of a child's chest wall system will be pulsatile, rather than being essentially constant as they are sometimes described as being in adults.

5.2.3 Control and magnitude of stress pulses

For the same reasons that I gave in section 3.5.2 it seems to me that the child would control stress pulsatility through the percept of effort.

Similarly, as in section 3.5.3, I see no reason for him to set the magnitude of the pulse with regard to the 'segmental' content of the frame (while this continues to give results which are acceptable to his listeners). However, the nature of the frame may change. I will shortly argue that from the respiratory system's perspective, a foot is the appropriate domain of action rather than a syllable.

5.3 Summary

In this chapter, I put forward various arguments in favour of a child enhancing his original style of SB for the new demands of stress accent, modifying it by increased effort to create appropriate prominences. After this 'assembly' stage he will be able to 'tune' this arrangement on his way to adulthood, but without any pressing requirement to do so.

I concluded that stress production in young children probably has two important characteristics:

1. Stress-accent prominence in young learners is largely created by activity of the respiratory system rather than by laryngeal adjustment.
2. The magnitude of this activity is independent of the segmental composition of the 'syllable' to which it applies.

---

after all; they arise from the connection between stress and extra effort. Surely at some developmental stage, stress must be embodied, with its correlates arising not from imitation of the acoustic cues from others' speech but from the actual use of extra effort in its production. If not, it is hard to see how the complexities of cue trading could survive. Stress would surely become phonologised, and resolve into one of its components: increased loudness, longer duration or spectral change.
In other words, while I can believe that in adults, "speech movements are programmed to achieve auditory/acoustic goals," as Perkell et al. (2000) describe, this surely cannot be true at all stages of development.

The independence of the respiratory system frame from its articulatory contents will persist to the extent that it does not give results that are unacceptable to listeners. When could this occur? It is unlikely to be the result of differences in the aerodynamic requirements either of vowels or of the consonant(s) in the coda (which is probably of little importance in the perception of stress[33]). As suggested earlier, though, /h/ (and perhaps some onset plosives) might require more effort from the young speaker than other consonants in order to be perceptible, and initial /s/ in onset consonant clusters may have a similar requirement (cf. the EMG results reported in Ladefoged (1967)).

In fact, the arguments I will be advancing would not be invalidated if there were some differences in stress pulse effort based on the nature of syllable onsets. So although from now on I will speak in terms of invariant stress pulses, the phrase should be understood to encompass this non-significant variation.

Finally, although it was implicit in my explanation of PFC, I should be explicit about a further assumption I am making. For a pulsatile style of SB, not only will no more than the resource created be available within the domain of a pulse, but all of that resource will be expended before a new cycle begins.

In the next chapter, I will describe a number of mechanisms which account for the appearance of the WGmPh through production constraints and the accommodations made by a child in response to these. Common to these proposals is the idea that the changes in the child's speech are precipitated by his use of a pulsatile style of SB for stress-accent.

[33] Cambier-Langeveld and Turk (1999:277) found coda consonants in an accented syllable to be lengthened less than onset consonants (more so in English than Dutch).

6 A breath stream dynamic (BSD) account of the emergence of the WGmPh

In chapter 2, I described some apparently independent phonetic phenomena that characterise English, German and Dutch, which I called the West Germanic phenomena (WGmPh). These have appeared to be the result of topographical learning[34] because they are not present in a child's early speech. However, by five years of age this has changed; almost all children learning these languages will have a rhythm that is stress-timed, tense and lax vowels which differ in duration as well as quality, and voice onset times that reflect the pattern of the adult norm.

My account of how these emerge will explain them in different ways, but not through the imitation/modelling of surface features that topographical learning would involve. Rather, they will be the result of a child reconciling the demands of stress-accent and the opportunities it affords with (i) the acoustic requirements of speech, (ii) the aerodynamics and physiology of his production system, and (iii) the need to take a path that achieves successful communication at every step.

In sections 6.1 to 6.3, I present new accounts of how each of the WGmPh might develop. In section 6.4, I mention some further phenomena that may be explained by the breath stream dynamics (BSD) of child speech.

6.1 Foot level shortening (FLS)

In English, the duration of a foot increases as syllables are added to it, but not proportionately. Instead, the syllables are compressed as their number grows (e.g. Rakerd et al. 1987). Compare '*one* '*two* '*three* '*four* with '*one and a* '*two and a* '*three and a* '*four* (spoken at a normal pace with normal reductions).

---

[34] Locke (1993:168) defines 'topographical learning' as, "... the reproduction of surface physical features that were not previously a part of the child's output repertoire, and which therefore shift the vocal contour or constituent parts of utterances in the direction of ambient stimulation. Topographical learning requires that the infant perceive differences between ambient sounds and the sounds of its current repertoire and possess whatever articulatory control is needed to achieve adultlike patterns."
As examples of this, Locke discusses how children refine their sound patterns to closely match those of the language surrounding them, and develop voice onset times that are similarly characteristic of their linguistic environment.

This 'foot level shortening' (FLS) is considered strong evidence in favour of English being stress-timed. Proponents argue that the effect is an attempt on the part of speakers to produce stresses with a tendency to isochrony.

6.1.1 Appearance of rhythm and vowel reduction

I am not aware of any research on the development of FLS *per se* in young children, but Allen and Hawkins (1980) surveyed their own and others' research into the related issue of the development of rhythm, making the following observations:

> "Two-year-olds tend to use far fewer reduced syllables than do adults, so that their speech rhythm has fewer syllables per foot, or more beats per utterance; in short, it sounds more syllable timed." (p.231)
>
> "Children's early polysyllabic utterances typically show a high frequency of unreduced syllables ... and one of the first rhythmically important skills the child apparently must learn in order to produce fluent English phrases is that of reducing weak syllables in acceptable ways." (p.235)
>
> "By the age of 4 or 5, the rhythm becomes more adult-like, with increased rate and greater numbers of reduced nuclei." (p.233)
>
> "Finally, we must make some mention of an important aspect of this research mentioned by virtually every investigator: variability within and between children. Some children may begin to modify their speech rhythm in their second year, others not until their fourth." (p.240)

Hawkins (1994:4180) added that, "much the most rapid development [of stress patterns] happens between the ages of about 3 and 5 years."

6.1.2 Aerodynamic effects of syllable reduction

A two-foot phrase like *fricatives and resonants* can be transcribed as |ˈfrɪkətɪvz ən |ˈrezənənts |. However, with respect to the reduced vowels the conventional analysis reflects the auditory outcome, but not necessarily its means of production.

Catford (1977:217-226, 1985) points out that all vowels are not created equal. In the final position of a phonological word he accepts the conventional analysis of schwa as a vowel. However, he characterises a CVC context containing a weak vowel as an 'open transition' between the consonants (C·C), where the articulation of the first consonant is completed before the second begins. This contrasts with a 'close transition' (CC), where articulations overlap, which is conventionally called a consonant cluster. He applies a

similar analysis to weak vowels in word initial position, and calls syllables containing open transitions 'pseudosyllables'.

Reversing the order in which I have introduced them, his examples of contrasting close transitions, open transitions and consonants separated by 'true' vowels include:

> "a brief lunch – a Brie for lunch – a briefer lunch"
> "make names – make an aim – make Ann aim"
> "take part – take apart – take up art"[35]

Catford compares the articulation of open transitions and vowels. He reports durations in comparative contexts of between 10 and 60 ms for the former, and 60 to 200 ms for the latter, with means of 30 ms and 110 ms respectively. In terms of channel area, he reports the open transitions to involve minimal release of the articulatory stricture: a maximum of 0.2 cm^2 for a bilabial open transition as compared to more than 2 cm^2 for a corresponding vowel opening. Stevens (1998:574-578) also discusses weak vowels in these terms, and reports similar data.

Thus the 'vocalic' portion of an open transition between consonants is brief and involves minimal change in the aerodynamic state of the upper vocal tract. From the perspective of the respiratory system, open transitions are not dissimilar to close transitions (consonant clusters) and syllabic consonants. To a first approximation, all create periods of high resistance to airflow.

So where a foot contains a stressed syllable followed by zero or more reduced syllables, the respiratory system sees the coda of the stressed syllable plus all of the following reduced syllables as a single unit; in a sense, as a complex consonant cluster. *Fricatives and resonants* would appear as || CCVC·C·CC·C | CVC·C·CCC ||.

---

[35] "Brian - O'Brian - Oh Brian!" is an example with word initial open transitions. Catford describes speakers creating these differences by control of the timing of their RS gestures:
"The difference between #C_ and #·C_ is that in the close transition the articulators are already fully in position for the articulatory stricture before the initiatory effort begins. In open transition, #·C_, the pulmonic initiation begins at about the same moment as the articulatory organs begin to move together to form the articulatory stricture. In the third case, #VC_, the pulmonic initiation for the vowel starts long (i.e. 10cs or so) before the articulators begin to move into position for the C-." (1985:340) (cf. Slifka 2003)

Children will not produce feet of such complexity to begin with. But in chapter 3, I proposed that the distribution of limited aerodynamic resource would explain both pre-fortis clipping and the shortening of the nucleus as the coda of a syllable grows (e.g. in *ram, ramp, ramped*). Continuing to apply this mechanism now explains what is effectively the same phenomenon, but now described as the level of the foot rather than the syllable. FLS is the result of a speaker distributing an invariant amount of resource over the domain of a stressed syllable plus following unstressed ones.

6.1.3 Varieties of foot

So far, I have characterised a foot[36] as consisting of an initial stressed syllable followed by syllables containing weak vowels which from an aerodynamic point of view create an undifferentiated, high resistance load. However, feet in English can also contain strong but unstressed vowels. Catford (1985:345) shows how the concept of an open transition allows this to be accommodated within his or my accounts.

For example, *photographs* would seem to have a final syllable that is stronger than its middle one, which would imply a pattern of initiator power applied that differs from the single pulse per foot model. (I.e., it would require an additional, minor pulse within the foot.) However, if the second syllable is reanalysed as a pseudosyllable (consonants separated by an open transition), then the profile of energy expenditure over the foot can again be seen to be that of a peak followed by a decline to the end of the word/foot.

Similarly, the concept of a pseudosyllable allows a rationalisation of the foot structure for an utterance beginning with an anacrusis. Rather than having an unfooted initial element, the utterance can be seen to begin with a complex consonantal onset. Catford gives the following illustration:

/ ʌ Pa / rade Street's / where we're / going /
/ P·rade Street's / where we're / going /

---

[36] In the UCL scheme of terminology, below the level of stressed syllables it is possible for a syllable to be either full (containing a strong vowel) or reduced (containing a weak vowel). As we have seen, in Catford's terms a closed reduced syllable is reanalysed as a 'pseudosyllable', with a close transition for its 'nucleus'.
With respect to 'foot', White (2002:57ff) describes the 'cross-word foot' (based on strong and weak syllables), the 'within-word foot' (similar, but not crossing a word boundary) and the 'Abercrombian foot' (based on pitch-accented syllables or, in another interpretation, on syllables with primary lexical stress). My use of 'foot' differs from all these, being similar to the cross-word foot but based on stressed and unstressed syllables rather than strong and weak ones. It therefore reflects performance rather than an abstract analysis.

These arguments enable Catford to ground his view of stress firmly in physiological reality. He reverses the usual direction of causation, which has feet being created out of the patterning of syllables already marked by different levels of stress, arguing that,

> "It is the stress-contour of the foot that imposes different degrees of stress upon the successive syllables it dominates, according to their location within the foot." (p.345)

In this way he says,

> "... we can account for four degrees of perceived stress in English without having any independent system of stresses at all. The strongest stress is the power-peak of a tonic foot, the next is the power-peak of a non-tonic foot, the third corresponds to the later, declining, part of the power-curve, and the fourth, and weakest stress, is that of the open-transition." (p.346) (cf. Hewson 1980)

Catford's reversal of the conventional analysis of stress is attractive. It would apply very naturally to child speech, making 'learning to stress' more straightforward than must be assumed to be the case at present.

6.1.4 The foot as a unit of production

My account of FLS only differs from Catford's (1977:87) in two ways. Firstly, he proposes a principle of 'isodynamism': equal 'initiator power output' for each foot, as defined by the aerodynamic variables of pressure and flow. I have suggested that invariant effort (other things being equal) is a more likely basis of control.

Secondly, he describes isodynamism applying in adult speech, while I have developed it out of a model of children's SB. I will discuss its status in adult speech in chapter 8.

Fowler (1996:547) proposed a similar idea, positing a pool of resources for producing an utterance on a single breath group, with the more there is to say the less being available for each unit. She speculated that this might explain the latency effects found by Sternberg and his co-workers (Sternberg et al. 1978, 1980, 1988). In these production experiments, the delay between a starting signal and the onset of speech was measured. Subjects knew in advance the sequence of words they were to utter, so the delay was taken to reflect production processes but not cognitive processes. It was found that it increased with the length of utterance, but that a coherent pattern only

emerged when the length of the delay was measured in feet. The relationship was then an extra 12 ms delay for every additional foot in the string.

This robust effect was taken by Sternberg and his colleagues as strong evidence for the foot as a unit of production[37]. My proposals suggest that the 12 ms latency increments reflect the planning of aerodynamic resource allocation for each foot[38].

I have described how conventionally defined syllables (the units of activity of the upper articulators) dissociate from the unit of activity of the respiratory system. Mead and Reid (1988) reported a similar dissociation. They found chest wall muscle activity whenever the airstream of an utterance was broken into syllables by repeated glottal gestures, but not when the interruptions were produced by lingual or mechanical ones. Aerodynamically, the effect of all three interventions were equivalent, so it seems that the tongue has an independence from the respiratory system in this regard not shared by the larynx.

Note that the mechanism I have described is consistent with proposals made by Fudge (1999) and Sluijter and van Heuven (1995) for a replacement of a single prosodic hierarchy linking tone groups and syllables with two parallel hierarchies, one of which is concerned with accent and the other with stress. Also, Kent et al. (1996:34) discuss 'frame and content' theory with respect to rhythm, explaining that in this view, "speech is organised into two channels, one representing syllable markers and the other representing the phonetic content associated with syllables."

6.1.5 Timing

I do not have the space to review my proposals against the extensive literature on metrical timing effects. However, they seem to be able to explain previous results, including those of the extensive recent investigations made by Turk, White and their colleagues.

I will, however, pick up on one global issue these researchers raise. Firstly White (2002) points out that,

---

[37] Marshall and Chiat (2003) provide alternative evidence for this in child speech.
[38] A.Smith and Goffman (2004:235 and 243) report evidence from lip/jaw movements that some motor commands for speech are planned at a phrase level.

> "For accentual lengthening ... no single unit fully characterises the locus, because lesser degrees of lengthening extend beyond its apparent boundaries." (p.77)
>
> "... there is not an obvious candidate for the domain of accentual lengthening within a constituent hierarchy." (p. 78) (See also Zhang 1996:92.)

As a result,

> "It remains an open question whether attempts to unify prosodic constituents and prominences within a single type of representation will ultimately succeed." (p. 78) (See also Cambier-Langeveld and Turk (1999:276) on Dutch.)

Similarly Turk and Shattuck-Hufnagel (2000) observe that:

> "In addition to word-initial lengthening, the pattern of results in our data appears to require polysyllabic shortening that operates relatively uniformly across the components of a syllable, accentual lengthening and/or pitch-accented word-final lengthening, and syllable ratio equalization that operates on the nucleus of word-initial primary stressed syllables. **We look forward to the day when the patterns we account for with a complicated set of mechanisms can be explained by a more parsimonious model.**" (p. 428)

The more detailed our understanding of the changes that actually occur in speech becomes, the more improbable it seems that children manipulate timing *per se* within a conceptual framework defined by prosodic units. Instead, the need for a speaker to distribute limited aerodynamic resource, may form the basis of a parsimonious and more plausible model.

There is, of course, no *a priori* reason to expect that resource distribution will always result in a timing pattern that can be modelled in terms of conventional prosodic units. The timing adjustments, whether FLS, 'anticipatory lengthening' (Van Lancker et al. 1988; Bolinger 1981), or anything else, will be epiphenomenal. They will be predictable to the extent that the child's resource allocation is implemented in a way that maps onto changing the time a segment takes, but this may vary by speaker, by occasion and so on.

Speakers will only be constrained by timing models in the cases where non-phoneticians already intuitively recognise this to be the case: in speaking verse, for example.

6.1.6 Learning

I have two further observations to make about what we observe of the learning of temporal prosody by children.

Firstly, the omission of weak initial syllables may be more understandable in the light of Catford's characterisation of them as pseudosyllables. From a production perspective, an open transition between initial consonants seems more like an onset consonant cluster than a normal syllable. A word initial open transition requires a complex coordination of the respiratory system and upper articulators, as Catford describes it. In both cases it may be the motor challenge that delays the child's production of the feature.

This suggestion is in accord with Faber and Best (1994), who suggest that the motor aspect of speech is the primary determiner of the developmental schedule. Also, Carter (1998) presents data showing that children do not entirely omit initial syllables, and hence argues for a phonetic explanation of the phenomenon to augment previous phonological ones.

Secondly, while English speaking culture has its nursery rhymes, to my knowledge at least two non-stress-accent languages, French and Japanese, have nothing similar (and certainly nothing to the same extent). Both have short sung verses that children learn (*contines* in French), and both have number chants, but neither have similar spoken verses for such young speakers.

The rhythmicality of nursery rhymes must be an important part of their appeal, but if speaking rhythmically were the attraction in itself then surely we would see more French material, for example, to support this.

So the appeal of nursery rhymes presumably goes deeper. For a child, they may highlight the gap between his speech and that of adults (they 'force' this awareness, as Gattegno would say), and challenge him, acting therefore as a pedagogical device to promote the reorganisation of the use of his respiratory and articulatory systems.

The simplest nursery rhymes (e.g. "Round and round the garden"), emphasize the loudness contrast between their stressed and unstressed syllables, particularly when performed in an exaggerated style. The SB challenge is clear. More sophisticated nursery rhymes ("Three blind mice"), add the requirement for complex strings of

syllables after the stressed syllable to be reduced, in order for time to be kept. This will encourage their articulation as open transitions.

6.1.7 Summary

I have suggested that in normal speech the child does not have an extrinsic goal for the timing of a foot, whether defined rhythmically or otherwise. He has the goal of producing a string of segments with a limited aerodynamic resource. Thus the timing adjustments that are called FLS are epiphenomenal.

A number of compression effects that have up to now been analysed separately, including PFC and FLS, may all be manifestations of a single underlying mechanism: a preference for simplicity in the early development of the motor control of speech breathing. Temporal effects are not the primary features we have taken them to be, and are not acquired by imitation.

6.2 Vowel changes under the influence of stress-accent

Three well-known characteristics distinguish tense and lax vowels in English:

1. Length differences in prominent environments.
2. Close and open articulations[39].
3. Differing phonotactic possibilities, for appearance in open and checked syllables.

These characteristics are believed to be independent of each other. So, for example, while it has been proposed that the first and second might be linked because the tense gestures take longer to execute, this does not seem to explain the data adequately (Wood 1975:110; Ohala 1992:307).

If these characteristics are indeed independent, then it is remarkable that using them to divide up the vowel inventory generates classes with exactly the same membership in each case. This may be a coincidence[40], but I shall propose that the implementation of stress-accent is a single underlying cause that connects them.

---

[39] Based on the point of maximum constriction in the whole vocal tract including the pharynx, as distinct from 'close' and 'open' auditory realisations, as indicated by the labels on the axes of the IPA quadrilateral.

[40] The (un)likelihood of it being a coincidence may be more apparent from the following example, which I think would be analogous.

There are several steps in my argument, beginning with the finding that that young children produce acceptable tokens of at least some of these vowels prior to the deployment of stress-accent. However, children's speech has a number of aerodynamic vulnerabilities. I will ask how a child can respond to the consequential threat that stress-accent poses. His challenge is to produce an acceptable output (appropriate vowel qualities, stress prominence on appropriate syllables) within the constraints imposed by a child-size speech production system. In this, however, he will be assisted by ways in which English pronunciation has evolved to be compatible with stress accent: he will be allowed to change the duration of his original vowel allophones and he will not be asked to produce certain vowels in aerodynamically unstable environments.

6.2.1 Appearance of tense and lax vowel classes

Stoel-Gammon et al. (1995) examined the development of /ɪ/ and /iː/ in nine American children aged 30 months. In this and a later study (Kehoe and Stoel-Gammon 2001) they reported reliable qualitative differences that distinguished the two phonemes in the speech of these children, and of some much younger children aged down to 15 months.

Previous data from adults showed a length ratio of 0.71 for these lax and tense phonemes. For the child subjects reported in 1995, the mature durational relationship was actually reversed. The overall ratio of 1.06 reflected lax tokens that were longer than tense ones.

This and other studies paint a picture of high variability. The same authors' reanalysis of Naeser's (1970) raw data for 20-month-old children gave a ratio of 0.74, while a later study (Stoel-Gammon et al. 1999) with twenty 24-month-olds showed seven of the

---

In most cultures, three properties of a book aimed at the general reading public, (1) the gender of its author, (2) its binding (hardback or paperback), and (3) its content (poetry or prose fiction), are all completely independent.
If we went to Ruritania and chose 11 books completely at random from a comprehensive general catalogue, we would, surely, be surprised if the five books that were hardback were all written by women and were all poetry, while the six books that were paperback were all written by men and were all prose fiction. (In fact, the exact correlation of any two of these properties would be surprising, let alone all three.) Of course, it could be that random selection would generate this neat division but we might instead suspect some underlying linkage between the three properties concerned. In this case, it might be that Ruritanian culture only allows poetry to be written by women and then values it so highly that it is always published in hardback for longevity, while prose fiction can only be written by men and has then to be as affordable as possible so it is always published in paperback. Some such underlying connection would be probable; the likelihood of our highly structured sample being the result of a random choice from a catalogue in a typical country would be very, very small.

subjects with lax vowels longer than tense ones in at least one of the contexts examined. Intrinsic vowel length is clearly not mastered by all children before 2 years of age yet they reliably produce distinct qualities for each vowel, as judged perceptually and by analysis of formant patterns.

With this background, I will now consider the likely aerodynamic consequences of greater respiratory system activity following the adoption of stress-accent by a child. I will start by considering two potential problems, and then look at the different solutions he might find, each adapted to the aerodynamic characteristics of particular classes of segments.

6.2.2 Aerodynamic vulnerabilities to stress-accent

One aerodynamic constraint on (literally) small speakers will arise from their lung capacity being very considerably reduced compared to that of adults. Figure 3-1 in section 3.2 indicates typical vital capacities for 4-year-olds to be less than 1 litre, compared to adult values of over 3 and over 5 litres for women and men respectively. At the same time the airflows observed in child speech in various studies have ranged from being about half to being almost the same as adult values (e.g., Netsell et al. 1994; Tang and Stathopoulos 1995). Whichever figures are most representative, it is clear that children's airflow as a proportion of the air they have available is much greater than that of adults.

Table 6-1 attempts to quantify this, to give a sense of the aerodynamic challenge a young learner faces. (Figures are very approximate, and data has been combined from two studies with different methodologies.)

	18 – 36 months	7 year-olds	16 year-olds
Syllables/breath group	1 – 3	8	16
(Predicted) vital capacity [(P)VC]	0.9 – 1.5 l	1.6 l	4.4 l
Volume/syllable	100 ml	40 ml	55 ml
Volume/syllable, as a % of (P)VC	**7-11%**	**2%**	**1%**
Lung volume excursion, as a % of (P)VC	13%	19%	17%

Table 6-1. Approximate values for some aerodynamic measures of child speech. (Sources: Boliek et al. 1997; Hoit et al. 1990)

The row of figures in bold – a normalised measure using vital capacity for cross-age comparison – shows the youngest age group using volume resource at, perhaps, 6 times the rate of the 16-year-olds. (For a healthy adult it may be hard to imagine speech made on such an extreme basis, although panting while speaking may give some sense of it.)

The final row of figures show the proportion of vital capacity used in practice. The youngest speakers' figure may be the result of saying few syllables at a time or the cause of it, but even if this group used a proportion comparable to older speakers, the number of syllables they could produce would still be very constrained.

Thus young children are faced by the real possibility of 'running out of breath' if they do not use their volume resource carefully. Even adult speech seems to be conditioned to some extent by this consideration. Hoit et al. (1993:519) use "air conservation" to explain the tendency for VOT's to be shorter at low lung volumes. It would therefore be surprising if child speech were not similarly conditioned.

I have proposed that a child increases chest wall system activity for stress-accent. In the absence of any glottal or upper articulator adjustment of tract resistance, one consequence would be increased airflow. This is in direct conflict with an important objective for the child at 2- or 3-years of age. He would like to be able to say longer utterances, not to find himself further constrained by the new feature of stress-accent, which has the potential effect of depleting his air resource more quickly than previously.

A second possible aerodynamic constraint might have more immediately disruptive consequences. An interval of high flow experienced as a result of a stress-accent pulse might threaten a child's subglottal pressure head. Protection of the pressure head is an important consideration for speech, highlighted by the difficulties faced by those with inadequate closure of the velopharyngeal port[41]. Young English speaking children might face an analogous challenge to that faced by speakers with cleft palates in this respect.

For these two reasons, I will assume that it is important for a child to mitigate the higher loss of air volume threatened by his adoption of stress-accent. I will shortly propose that

[41] Warren and others have worked on the mechanisms that protect the pressure head. See, for example, Warren et al. (1990) and Smith Hammond et al. (1999).

he does this (i) by control of length of exposure (for lax vowels), and (ii) by control of airflow (for tense ones).

6.2.3 Vowel articulation: approximants and resonants

As a preliminary to my explanation of tense and lax vowel characteristics, I will follow Catford (1977:120), who divides vowels into categories of 'approximants' and 'resonants' using aerodynamic criteria. All vowels are produced with smooth airflow during normal voiced production, of course. Catford points out, however, that the restricted channel area of some vowels leads to the creation of turbulence when airflow increases as a result of devoicing. In this way, these vowels are similar to the four sounds in English conventionally described as approximants: the frictionless continuant /r/[42], liquid /l/ and semi-vowels /j/ and /w/ (Gimson 1989:35). In adult RP, the tense vowels /iː uː ɔː ɑː/ are produced with sufficient deformation of the vocal tract to be called approximant. /ɜː/ is a special case that I will discuss further below.

Other vowels have a larger channel area[43], which means that airflow will be smooth in both voiced and voiceless conditions. This is true of the lax vowels of RP /ɪ e ɒ ʊ ʌ æ/, which Catford labels 'resonants'. The consonants called 'fricatives' are a familiar third category, with the smallest channel areas. For these, turbulence occurs with the airflow rates of both voiced and voiceless conditions. See figure 6-1 for Catford's diagrammatic summary of this.

[42] Phonetically [ɹ] in the contexts relevant to this discussion, but I shall use phonemic symbols.

[43] Note that the arch of the hard palate (in coronal section) means that lowering the tongue increases the cross-sectional area of a constriction in a non-linear fashion. (Illustrated in Guenther et al. (1998:617) from original sketches by K.N. Stevens and J.S. Perkell.)

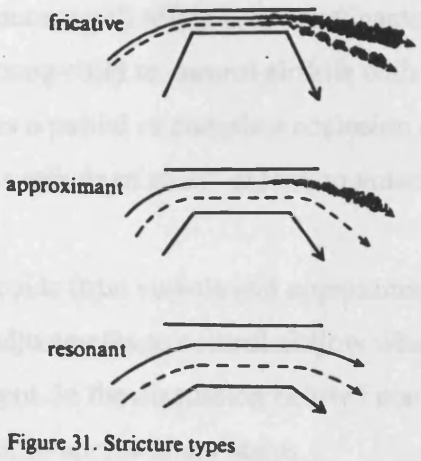

Figure 31. Stricture types

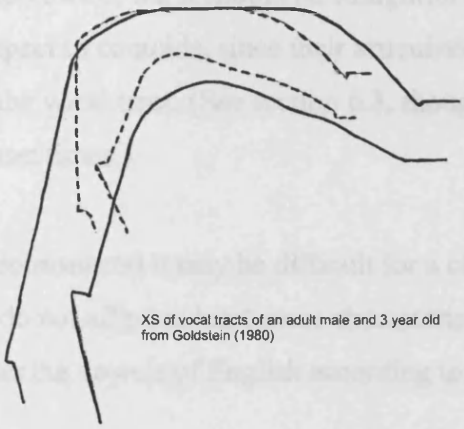

Figure 6-1 (left). Stricture types from Catford (1977). Each diagram shows (from the top) the palate, unvoiced airflow, voiced airflow and the top surface of the tongue. Turbulence occurs in some conditions.
Figure 6-2 (right). Mid-sagittal cross sections of the vocal tracts of an adult male and 3-year-old (dashed) superimposed. Derived from Goldstein (1980).

Young children have vocal tracts that are very significantly smaller than adult ones. Crelin (1987:75) shows rubber casts of the vocal tracts of a newborn, a two-year old, a six-year-old and an adult male, which strikingly illustrate the differences in both size and shape. Figure 6-2 is a similar comparison based upon Goldstein (1980).

It seems natural from this reduction in tract size that the cross sectional area of any child's vowel constriction should be smaller than the corresponding adult one (or at least no larger). Consistent with this, Goldstein (1980:207) gives reasons for believing that infants' vowel constrictions are very much smaller than adult ones. Further, acoustic modelling of the vocal tract suggests that F1 is proportional to the inverse of the square root of the area of maximum constriction. Thus a child's higher F1 across all vowels implies smaller such areas.

We have seen that from an early age children produce formant patterns for tense vowels that reflect the adult pattern, so let us assume that children's tense vowel articulations are at least approximant. They may, of course be even more sensitive to increases in airflow than this label implies, i.e. the increase in airflow needed to provoke the onset of turbulence may not even be as great as that resulting from devoicing.

Stress-accent will affect both consonants and vowels, but it should be straightforward for a young child to control airflow with respect to contoids, since their articulation involves a partial or complete occlusion of the vocal tract. (See section 6.3, though, for possible effects of stress-accent on voice onset times.)

For vocoids (true vowels and approximant consonants) it may be difficult for a child to make adjustments to control airflow which do not affect at least some characteristics of the output. In the discussion below I consider the vowels of English according to their resonant or approximant status.

6.2.4 Vowel adaptation to stress

Resonants (short and 'checked' lax vowels)

The simplest case we have to consider is that of a resonant immediately followed by a consonant as, for example, in *give*. The state of aerodynamic instability that may result from the vowel being stressed is promptly closed and thereby limited, so there is no need for the child to make any adjustments as a result of stress-accent. This class of vowels will come to be labelled 'lax', with the following consonant being one of its defining features phonologically.

If resonants were not checked in this way and no other precautionary articulatory adjustments were made, then there would be the negative consequence of air loss and the danger of the loss of the pressure head[44]. Another solution to this problem is diphthongisation, which I consider after I discuss approximant vowels. The phonotactics of English suggest, however, that producing a resonant in an open syllable in the same way as it is articulated in a closed one would be unacceptable, and is therefore not seen.[45]

---

[44] Willis (1919:34) also reasoned this way, although he did not relate the idea to children specifically: "Vowel sounds are of two species, long and short, i.e. those which, like the a in 'father', can be prolonged, and those which, like the e in 'let', cannot. The reason of this seems to be that in forming the short vowels the throat is in such a position as to emit a large quantity of air, so that the lungs are immediately emptied of wind; hence it is necessary to close or partly close the mouth in order to lessen the expense of wind, if speech is to continue; in other words, such vowels must always be followed instantly by a consonant ..."

[45] In further support of this being problematic, Weiss (1976:12) reports that many Germans find it difficult to lengthen a lax vowel. (Stetson (1951:67-68, 222) described the same problem in singing.) It may be that adults as well as children find the aerodynamic conditions associated with unchecked lax productions uncomfortable. My own attempts to prolong lax vowels (while being careful to maintain my larynx, in particular, in the attitude it adopts for a normal rendition) generate very different aerodynamic conditions from the stable situation of prolonged tense vowels.

Approximants (long tense vowels)

As I argued earlier, while there is no reliable data for the channel areas of young children's vowel productions, there are reasons to believe that they are smaller than the adult equivalents.

The exact place and manner of young children's articulations are likely to differ from adult ones for another reason: the vocal tract is a different shape in young children, particularly in the regions where [ɔ] and [a] type vowels, and /ə ɜː r/ type sounds are articulated[46].

I will assume in this discussion that the aerodynamic characteristics of the constrictions of /ɔː/ and /ɑː/ in a young child's speech are similar to those of /iː/ and /uː/, even though the cross sectional areas of the former in adults are usually reported to be greater than those of the latter. As we shall see, if this is incorrect then there are other mechanisms which will account for the behaviour of /ɔː/ and /ɑː/ under stress-accent.

For approximant vowels like these, increased airflow poses a threat in addition to the two already mentioned. It may lead to the onset of turbulent sound production. This would add an unacceptable sound quality to the vowel.

It might be possible to model the aerodynamics of child speech to examine whether or not this threat is real. A central issue would be how close the child's previous configuration of constriction/flow is to the point where turbulence will begin. Two lines of enquiry seem to cast some light on this, and add to the plausibility of the idea:

1. Fant et al. (1997:28) found a significant pressure drop across a tense vowel constriction in an adult speaker, "in the maximally constricted interval of a long, highly stressed [u:]," where they noted a supraglottal pressure of 2.1 cmH_2O. This level of pressure drop across an oral constriction is at the point which Stevens (1998:515) takes as a threshold for the onset of turbulence. If this is found with an adult's tense vowel articulation, then one can imagine greater potential effects in a child when creating stress. His baseline P_{sg} will already start at a higher level than

[46] For general discussions on the way that children's (and women's) articulations must differ from those of men beyond simple dimensional scaling see Nordström (1977), Böe (1999).

an adult's and his tense vowel constrictions are likely to be smaller.

2. Stevens (1998:32ff) discusses airflow and pressures in the vocal tract for typical static configurations. His figure 1.26 (see figure 6-3) extends the usual analysis of mean values to encompass the alternating flows of voiced production. It shows (i) an overlap between the region of peak flows for vowels and that where glottal aspiration will occur, and (ii) the region of peak flows for vowels abutting the region containing values for fricatives.

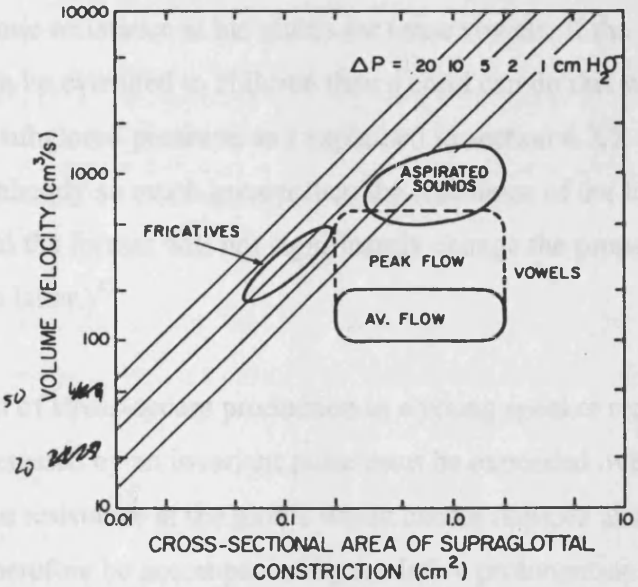

Figure 1.26 Sketch of approximate range of constriction sizes, airflows, and pressure drops at the supraglottal constriction for vocal tract configurations corresponding to various classes of speech sounds as indicated. (Adapted from K. N. Stevens, 1971.)

Figure 6-3. Airflow during adult vowel production. Peak flows approach flows that produce turbulent sound sources at glottal and oral constrictions. From Stevens (1998).

Stevens suggests that the unwanted sound produced as a result of the overlap will be minimal under normal conditions, but the fine tolerances which speech production works to are clear. The ways that a child's vocal tract aerodynamics will differ from an adult's – smaller constrictions, higher subglottal pressures – would already lead to overlap of the region of peak flows for vowels with the regions for fricatives and for glottal aspiration if his other aerodynamic variables were similar. If pressures now increase as a result of stress-accent, then an adjustment to avoid turbulence by reducing airflow would seem likely to be necessary.

If turbulent sound is indeed the potential consequence of the deployment of stress-accent by a child, how can he restore the balance between acceptable acoustic output and the aerodynamic conditions in his vocal tract?

The turbulence could be eliminated by easing the lingual constriction, but this would change the acoustic character of the vowel. Alternatively, the effort associated with the stress pulse could be reduced, but then the syllable would lose its prominence.

A third possibility is for the child to reduce airflow directly, by increasing the aerodynamic resistance at his glottis for tense vowels. If the model of Finnegan et al. (2000) can be extended to children then a child can do this without significantly affecting subglottal pressure, as I explained in section 4.2.2. (The resistance at the glottis is already so much greater than the resistance of the lower airways that an increase in the former will not significantly change the proportion of pressure dropped across the latter.)[47]

My model of stress-accent production in a young speaker requires that all of the resource created by an invariant pulse must be expended over the course of a foot. An increase in resistance at the glottis which causes reduced airflow but minimal change in P_{sg} will therefore be accompanied by a relative prolongation of the vowel. Hence the child will come to produce context-dependent durations for tense vowels: relatively lengthened when in stressed contexts, not lengthened in others, as we observe in

---

[47] I am not aware of anyone having measured laryngeal resistance appropriately during a young, English speaking child's vowel production, but in principle this should be possible. (In practice it may be too invasive.) In support of the plausibility of there being different levels of resistance for different vowel classes I would offer the following:
1) Laukkanen et al. (1995) show that a manoeuvre which increases resistance need not produce any perceptual effects (at least, they report none).
2) Catford (1977:203) gives an example of, it seems, differing levels of aerodynamic resistance at the vocal folds for vowels in Javanese, indicating a precedent for this aspect of the proposal.
3) The variability in the intrinsic pitch of vowels already indicates that there are some differences in their laryngeal settings.
4) Warren (1996:57) portrays the glottal resistance of vowels to be at least three times that of voiced fricatives. If such wide discrepancies in glottal resistance during voicing exist, it would not be surprising if there were also some variation within the vowel category.
5) Stevens (1998:297) gives guarded support for the notion of laryngeal adjustments to accompany tense-lax distinctions in adult speech.
Phoneticians who have thought that the qualities of tenseness and laxness extended to the vocal folds include Heffner (1964:96) and - as reported by Fischer-Jørgensen (1990:130) - Sievers, Meyer and Russell. See, however, my discussion in section 6.2.5 for how this might be manifest.

practice[48]. Since this is compatible with the perceptual expectations of his interlocutors, there will be no adverse reaction to this and hence no reason for him to intervene to modify the process. (If, indeed, he is even aware of the change in timing.)

Examples of aerodynamically motivated accommodations of this type in adults have been reported by Scully (1992:4) for a French speaker, and Rothenberg et al. (1987) for soprano singing. A different proposal, but one in a similar spirit, is made by Stevens (1998:270):

> "Limitations on the minimum pharyngeal cross sectional area that can be achieved for low vowels are determined by acoustic and aerodynamic factors. ... [As] the cross sectional area at the constriction is reduced, a point is reached where airflow through the constriction ... causes a pressure drop ... and can cause turbulence noise to be generated at the constriction. This build-up of pressure and possible generation of noise is not consistent with production of a vowel, and thus the reduction in constriction size for a low vowel should be limited to cross sectional areas that are greater than this critical value. This value is probably in the vicinity of 0.2 to 0.4 cm^2 for adult vocal tracts..."

Stevens is suggesting that the cross sectional area of a constriction must be greater than a certain minimum value to satisfy an aerodynamic constraint. I am suggesting that given the pre-existence of distinct acoustic targets, there is little scope for the child to vary the cross sectional area of vowel constrictions, but that variation of airflow is an alternative strategy with the same motivation. It seems that using the larynx as a throttle valve would be a natural way to achieve this.

Laryngeal adjustment is needed to avoid turbulence, but by reducing airflow it also addresses the issues of depletion of volume resources and loss of the pressure head. So approximant vowels with such a glottal adjustment need not be constrained in the way described for lax vowels, above, and can appear in both checked and free syllables.

---

[48] Research on another stress-accent language, Dutch, demonstrates that for a tense vowel to lengthen, it is not enough for it to appear in a phonologically strong position (as opposed to a phonetically strong one). Rietveld et al. (1999) examined the durational behaviour of two vowels ('half long' /i/ and 'long' /a:/) in all (nine) possible prosodic structures of Dutch. They found that,
"... lengthening of vowels in the heads is not restricted to primary stress, but is observed in syllables with secondary stress as well, i.e., is the result of 'headedness' in the foot. Vowels in the heads of weak or strong feet are longer than their counterparts in weak syllables."
"... both vowels have the same [underlying] duration, and ... the difference between the two only surfaces in foot heads. This result corroborates the constraints (and ranking) on syllable and foot structure in Dutch proposed by Gussenhoven, **in which the long duration of 'long vowels' is derived from stress** ('Stress-to-Weight Principle')." (p.466)

They form the class of tense vowels in RP, characterised by prolongation under stress-accent and substantial lingual excursion in their articulation.

I would like to add three further observations here:

1. The perceptual-centre (P-centre) of the stressed syllable in a foot probably reflects the peak of the pulse activity, at a position close to the junction between the onset and nucleus (see section 6.4.1 on P-centres). This is where there is the greatest danger of turbulence. Tense vowels in the syllables subsequent to the stressed one (i.e. within what will usually be a sequence of consonants linked by open and close transitions) will therefore have no need for the adjustments described above, and will be of similar length to lax ones.

2. The extra length of /æ/ in certain contexts (Gimson 1989:93, 107) may be a result of its production by some speakers "with considerable constriction in the pharynx". Gimson accounts for the length as a feature to better distinguish /æ/ from /e/, but it seems possible that the pharyngeal constriction creates a risk of turbulence, which is countered by a raising of glottal resistance. This leads to increased length for the reasons described above.

3. I suggested that given the pre-existence of distinct acoustic targets, there is little scope for the child to vary the cross sectional area of approximant vowel constrictions to control airflow. However, with the respect to these targets, the lowering of RP /i:/ from [i] (or the French /i/) and the relaxation and fronting of /u:/ from a true back position (Gimson 1989:121), may both be reflections, on the one hand, of the systemic need for vowels within these parts of the acoustic/articulatory space, and on the other, of the advantages of minimising turbulence under stress-accent by increasing the cross-sectional channel areas.

Diphthongs

The standard diphthongs of RP can be divided into 'closing' and 'opening' classes, based on the movement of the tongue relative to the hard palate during their glides. The closing diphthongs, /eɪ aɪ ɔɪ aʊ əʊ/, have resonant starting points and glides that move in the direction of [i] and [u]. While the use of the symbols for /ɪ/ and /ʊ/ is now standard in their transcription, different approaches have been proposed and are still

used for other varieties of English. These alternatives include the use of /j/ and /w/, and of /^{i}/ and /^{u}/. The latter approach is used by Catford (1985) and in some treatments in the tradition of (Firthian) prosodic phonology.

Given a resonant starting point, part of the function of the second element must be to stabilise the aerodynamic situation in the same manner as a consonant following a lax vowel. The glide can help to do this by occluding the airway, not (usually) to the degree that consonant-like frication occurs but to the extent that the resistance at the oral constriction can be increased to a point just short of this.

At the same time there is presumably also a need for some glottal adjustment. The aerodynamics of the child's production are therefore controlled by these two resistances in series, with the lips possibly providing a third resistance in glides towards [u].

For the opening diphthongs, /ɪə eə ʊə/, the second element must achieve a similar overall result. Schwa is usually considered to be the result of a neutral configuration of the vocal tract, which would suggest that the glide has no resistive effect. However, this may not be the case, even for adults (Gick 2002). Given the very different morphology of the pharyngeal section of the vocal tract in young children, there is every possibility that schwa production creates an aerodynamically significant constriction in young speakers. In this case the schwa 'target' would function in exactly the way described for the final elements of the opening diphthongs. If this is not the case, then increased laryngeal resistance alone may stabilise the aerodynamic state during the glides.

While the discussion above describes an abstract form of RP, in practice it allows some variation in the balance between use of glottal and oral resistance. So, for example, speakers are free to realise /i:/ as /ij/ in many situations (Gimson 1989:101). This variation respects the aerodynamic constraint. Further, speakers may monophongise the RP diphthong /eɪ/ by prolonging the initial element (ibid, p.95), resulting in an extended resonant production with airflow presumably regulated through applying increased glottal resistance throughout. This may also be the case for opening glides towards /ə/ which, "some RP speakers [realise] merely as an extension of the preceding syllabic vowel element." (ibid) Such generally increased resistance would also account for the patterning of RP /ɜ:/ as a tense vowel.

As part of the same pattern, dark /l/ and its vocalic allophone [ö] are used as diphthong-like final elements in many words (Gimson 1989:205).

Extending our gaze beyond RP, we see that English exploits the variation allowed between aerodynamic resistance at oral and glottal constrictions in a more widespread way. Rhotic varieties of the language presumably change the balance in favour of oral resistance. Many dialects generalise the optional monophongisation of the RP diphthong /eə/ to all its realisations.

The possibilities that English can exploit for what are the tense and diphthong phonemes of RP are graphed in figure 6-4 (from Stevens 1998). Airflow is shown as a function of the area of the supraglottal constriction for a family of curves representing different sizes of the glottal opening. Between the extremes of no glottal constriction and of the area of the supraglottal constriction being large there is, "a wide range of cross-sectional areas ... in which the airflow is determined by both constrictions" (p.36)[49]. Dialects of English have adopted different positions on the curves, with varying acoustic consequences. English allows free variation among these possibilities, not only between dialects but to some extent within the speech of any given speaker[50].

[49] Note that the graph is based on a subglottal pressure of 8 cmH_2O, so the values given are probably not representative of a child's speech when creating stress.

[50] As an aside (while on the subject of diphthongs), Ladefoged (1983:177) points out that the vowels in American English words such as *beard, bard* and *bird*, "are most peculiar, and occur in few, if any, of the other languages of the world." If this is correct, then it may be a consequence of the rarity of stress-accent, which shapes vowels for aerodynamic reasons in the way I have been describing.

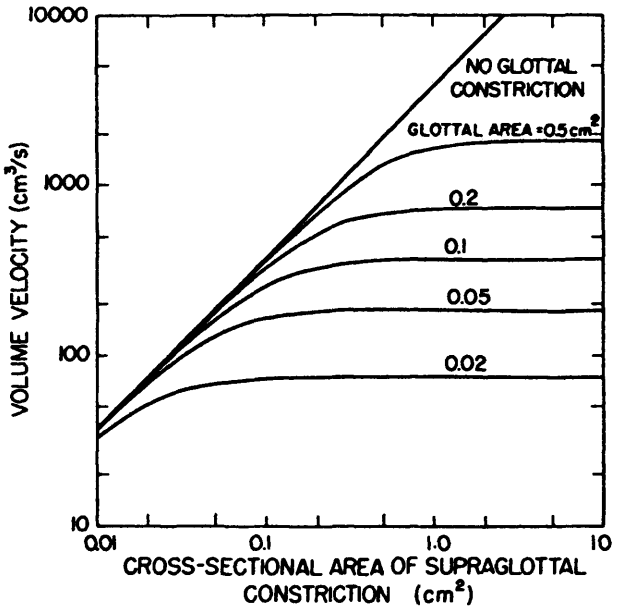

Figure 1.29 Airflow vs. supraglottal constriction size for various glottal openings, based on equation (1.12). A fixed subglottal pressure of 8 cm H_2O is assumed.

Figure 6-4. The possibilities for trading off oral and glottal resistances in the production of tense vowels and diphthongs. Figure 1.29 from Stevens (1998:36).

Schwa, schwi and schwu

The analysis of the final element of diphthongs as high resistance may generalise to word final schwa and to the 'happY' and 'inflUence' vowels (Wells 1982:165; 1990), which I have suggested we might christen 'schwi' and 'schwu' (Messum 2002:23). High resistance will limit airflow. This and the stable subglottal pressure predicted by Finnegan's model means comparatively little aerodynamic resource will be expended in their production, in line with the intuition of native speakers that these are 'low energy' segments.

Systemically, the key attribute of word final schwa, schwi and schwu would be increased tract resistance. Speakers therefore have wide discretion as to how to realise these segments acoustically, as long as the final result in the case of schwi and schwu is somewhere in the vicinity of [i] or [u].

This analysis provides an explanation for the perception/production difference in word pairs like *windier* and *reindeer*. Listeners hear the final parts as the same. Speakers feel

that there is a difference[51]. Both can be correct if in the case of *windier* the two final segments are both produced with a high glottal resistance /i̤ə/, whereas in *reindeer* only the glide of the diphthong /ɪə/ moves towards this state.

In standard analyses, the reduced vowel system of RP includes the weak /ɪ/ that distinguishes the rhymes of *rabbit* and *abbot* and a similarly weak /ʊ/. Aerodynamically these are open transitions coloured by [i] and [u] like sounds (Catford 1985).

6.2.5 Inconsistencies

All this having been said, recent investigations into voice quality during vowel production have produced results that are problematic for my account. Kingston et al. (1997) report that, "tense vowels in American English have laxer voice qualities than their lax counterparts." Epstein (2003:2406) describes the vocal/laryngeal differences concerned as follows:

> "Tense voice is associated with high values of adductive tension, medial compression and longitudinal compression of the vocal folds; lax voice is associated with the opposite"
>
> "Low values of OQ (open quotient) are correlated [with] ... a tenser voice quality"

Tense voice, then, tends towards creak, and lax voice towards breathiness. Di Paolo and Faber (1990:162) noted further that, for at least some languages, such creaky and breathy voice qualities have been correlated with reduced and increased airflow, respectively.

Gordeeva (2005:216) reports her child subjects between 3;4 and 4;9 producing, "a more breathy laryngeal configuration for the tense /i/, and a less breathy configuration for the lax counterpart."

All this appears to tell against my account, where tense vowels are supposed to have greater aerodynamic resistance at the glottis.

In response, I can at present offer some thoughts which might help to reconcile what I have described with this evidence.

---

[51] John Wells, personal communication.

Firstly, there would be no conflict if a breathy laryngeal setting could have higher aerodynamic resistance than a tense setting. At first glance this seems contradictory, but many qualities of the larynx are, or can be, more or less orthogonal: phonation type, pitch and loudness, for example. All these relate to the larynx operating as a resonator. It may be that its valving characteristics can be, to some extent, independent of phonation type. Features such as [± stiff], [± slack] and [± ATR] (Stevens 1998:251ff), all affect the tension of the vocal folds, but in ways with potentially different aerodynamic consequences, for example.

Secondly, there is a potential confusion between (i) vocal effort and its acoustic correlates, and (ii) respiratory system effort. The respiratory system resource available for a lax vowel is expended over a shorter period of time than for a tense one in my account, so this might be reflected in a measure of vocal effort in some way without this necessarily being derived from a setting of higher laryngeal resistance.

Thirdly, the functioning of the larynx in small children may be sufficiently different from that in adults to require a separate analysis. Stathopoulos (1995:78) points out:

> "One noticeable difference between children and adults is that they do not use their laryngeal mechanism in the same way to increase vocal intensity."

She goes on to relate this to fine motor co-ordination and/or the physical development of the laryngeal muscles. Moore (2004:193) develops this theme:

> "Another especially intriguing aspect of the physiologic development of speech is the rapidly changing biologic context in which it occurs. Many of the characteristics, properties, and capacities incorporated into the adult model are simply not present in, nor attainable by, young children."
>
> "[T]he vocal folds appear to lack the essential biomechanical properties that would permit modulation of fundamental frequency (F0) and intensity using adult-like strategies. Whereas the elasticity of the vocal ligament in the mature system imbues it with the capacity to generate a wide range of tensions, the immature vocal folds have no such capacity. Modulation of F0 and intensity then requires the use of a poorly understood, immature control mechanism."
>
> "More significantly, these differences reinforce the notion that speech is developed using a mechanism that is fundamentally different from that of the mature, target system."

So, for example, the close integration and fine balance between the respiratory system, larynx and upper articulators that Bucella et al. (2000) discovered (which indicated vowel-specific adjustment of all three systems during adult speech), is unlikely to be a feature of a young child's speech, particularly when SB is still highly pulsatile and not stereotypic.

Finally (but connected to my first point), in two footnotes I have mentioned observations by Willis, Weiss, Stetson and myself which I have taken to be evidence that lax vowel configurations are aerodynamically unstable, i.e. that they 'need' a following consonant[52]. It is only personal introspection, but while I find it easy to sense greater 'vocal effort' in my lax vowels than my tense ones, I find it hard to believe that I have a greater glottal resistance in the former than the latter. Tense vowels, after all, can be prolonged indefinitely. Lax vowels, seem to require a laryngeal adjustment to allow this.

6.2.6 Summary

The early vowels of English-speaking children do not exhibit tense and lax characteristics, but are adequately distinctive on the basis of sound quality alone (Buder and Stoel-Gammon 2002). Prior to tense/lax differentiation, then, the child has a mode of production that is aerodynamically balanced.

I have proposed that the impact of stress-accent is to potentially unbalance a child's production of vowels, threatening faster depletion of air resource and loss of the pressure head. There are different ways that the child can respond to this:

- Resonant articulations are safe if the vowel is immediately followed by a consonant. The aerodynamic state of 'lax' vowels is controlled in this way.
- Resonant articulations can also be stabilised with a glide towards a high resistance, vocalic element. The resistance will result from the combined effects of the oral constriction(s) and any laryngeal adjustment; with the relative proportions of these articulations recognisable in the acoustic output. This class of sounds forms the diphthongs. A further variant of their production is for the speaker to apply a high resistance laryngeal adjustment throughout a resonant production.

---

[52] I can add Durand (1947:147, 160) to this. She reports increasing airflow in lax vowels, decreasing airflow in tense ones, for an American English speaker.

- Approximant articulations create an additional threat: unwanted turbulent sound production. The airflow during their production must therefore be controlled in all stressed contexts, which is achieved by increasing laryngeal resistance. As a by-product, the child prolongs the vowel in relative terms, to use up all the resource of the respiratory system pulse. The increased resistance also allows it to be produced in all contexts, 'checked' as well as 'free'. Such vowels are labelled 'tense'.

The different ways that tense and lax vowels produced under stress-accent are reconciled with other production demands and constraints, have now explained why it is possible to generate the two classes in the three ways I described earlier: by the increased length of tense vowels in high prominence contexts, by the extent of the lingual excursion for tense vowels, and by the phonotactic constraints on lax vowels. What has previously seemed to be a remarkable coincidence can now be seen to be part of a more coherent system of production.

However, evidence of voice quality in tense and lax vowels may contradict this. If the characteristics of these vowel classes do emerge naturally rather than through imitation, then clearly the way this happens is going to turn out to be more nuanced than the account I have been able to give.

*6.3 Voice onset time, aspiration etc**

What follows is a sketch of how the development of VOT and associated phenomena could receive an alternative treatment in a BSD account.

There is a widespread assumption that children attempt to reproduce target values of VOT that they have inferred from their linguistic environment (e.g. Cho and Ladefoged 1999). This would be a process of temporal imitation or modelling.

However, there are many experimental findings that seem inconsistent with this[53]. Also, the developmental path of VOT, as described by Whiteside and Marshall (2001) (drawing on Macken and Barton (1980) and others) is far from what one might expect:

[53] I am particularly struck by Van Dam's (2003) discovery that the adult state has not two, but three or four distinct ranges of VOT for voiceless plosives. It is hard to imagine any child of any age ever noticing such fine distinctions, and then being motivated to model and replicate them.

> "Stage 1. From the latter half of the first year to around the age of 18 months, some children display no distinction in the production of VOT values between the 'voiced' and 'voiceless' adult forms. In these cases, most stop productions fall within the range of 0-30 ms.
>
> Stage 2. A distinction develops with 'voiceless' stops produced with longer VOT's, although they are still perceived as 'voiced'. For some children, this stage occurs around 18 months of age, but can extend up to the age of 28 months. This is an example of **covert contrast production where the differences, although produced, are too subtle to be perceived by listeners.**
>
> Stage 3. With further development, **overshoot of adult VOT values** is noted in the production of 'voiceless' stops which are later retracted back to more adult-like values around the age of 4 years. The observation that 4-year-old children display an overshoot of adult VOT values therefore contrasts with the interpretation of VOT data from 2- and 6-year-olds [by earlier researchers] which [suggested] that the development of VOT is a continual movement towards adult values.
>
> Stage 4. By the age of 6 years, a bimodal distribution of VOT values is produced for the contrastive 'voiced' versus 'voiceless' stops although some overlap continues to be observed.
>
> Stage 5. From the age of 6 years, stops are produced with adult-like VOT's with non-overlapping bimodal distribution. What is worthy of note at this point, however, is that **variability in VOT productions continues well beyond the normal period of phonological acquisition, and they generally reach an adult-like minimum level around 8 years of age.**"

In general, it seems that [p t k] is the initial stop series for all children, interpreted as /p t k/ or /b d g/ depending on the linguistic environment. For many of the languages where a child must differentiate a second series, the adult distinction is based on actual vocal fold vibration. This appears to set an articulatory challenge to a young speaker that takes some time to master (Allen 1985). In English and German, however, the two series are distinguished by short and long lag VOT's (at least in prominent positions).

Given the model of SB I have developed, there is a natural alternative to the imitative account. The deployment of stress-accent in English will precipitate the **discovery** of [p^h t^h k^h], which will, of course, by accepted by listeners as /p t k/. The degree of aspiration, and hence the duration of long lag VOT's, will largely be determined by the characteristic stress pulse applied by the speaker, not by imitation of perceived timings. For /b d g/, on the other hand, narrowing of the glottis will control the stress pulse, and aerodynamic factors will lead to the short lag VOT's observed (cf. Berry 2004).

There are three articulators which can be used to control aspiration and delay voicing: the respiratory system, the larynx and the oral articulator (tongue or lips). I suggest that in adults these are united into a synergy which re-creates or re-describes the results of a

child's speech production system. Mature SB is smooth, stable and stereotyped. On the foundation of a well-controlled subglottal pressure a speaker can play on a continuum of glottal width and interarticulator timing to achieve distinct characteristics for plosives. But I would contend that the underlying model for aspiration and VOT will be driven by the nominal strength of the stress pulse in any context, not by linguistic/phonological rules generating timing targets[54]. (Cf. my discussion of planning in section 8.2.)

It seems that this approach has further explanatory potential. The fortis/lenis distinction, which manifests, for example, in almost universal closure interval differences[55], may be the result not of greater power input, as sometimes suggested, but of greater resource utilisation. These ideas, however, will have to be pursued elsewhere.

6.4 Further effects

The mechanisms I have described provide natural explanations for the emergence of the four phenomena I identified in section 2.1: PFC and the WGmPh. They also generate explanations for some further speech phenomena, which I will briefly mention.

6.4.1 P-centres

The perceptual-centre (P-centre) of a word is regarded as the moment in time that both speakers and listeners use to make rhythmic judgments (Fowler 1979). Most investigations have focussed on the perceptual aspect of this phenomenon (e.g. Scott 1998), but research by Fowler and her colleagues (for example, Whalen et al. 1989) and more recently by De Jong (1994) and Patel et al. (1999), looked for the peaks in upper articulatory activity that might coincide with a P-centre. These investigations proved inconclusive.

---

[54] My proposal gives a good account of the data for English and other European languages. However, a different explanation is needed for how a child learns Hindi, for example, where voiced and voiceless plosives all occur in unaspirated and aspirated forms (so /p/ /p^h/ /b/ /b^h/ /t/ /t^h/ and so on).
A natural account is possible here, too. It seems reasonable to imagine that infants and young children will discover two ways to release the pressure built up behind an obstruction. One way is through a crisp, positive releasing gesture. The second is by relaxing the obstructing articulator so that the pressure behind it naturally opens the obstruction. In the first case, oral pressure will quickly drop, creating the pressure differential across the larynx needed for voicing. In the second, the drop of pressure in the oral cavity will be slowed, leading to the resumption of voicing after a time delay and with some intervening airflow. Thus /p/ and /p^h/ will be produced naturally by all children, with the parents of Hindi-learning children reinforcing the difference by their responses, leading their child to preserve and develop this distinction in the release mechanism of plosives. (By adulthood, a Hindi speaker may have redescribed the difference in terms of timing relationships, as I suggested may be the case for English and German speakers.)

[55] Except, it seems, in stress-accent languages in prominent positions (Flege and Brown 1982).

By my account, the P-centre of a foot will correspond to the peak respiratory system pulse (or perhaps the peak of perceived effort if this occurs at a different point).

6.4.2 'Syllable cut' phonology

This account also provides a phonetic motivation for Trubetzkoy's *Silbenschnittkorrelation*, by which he tried to explain the differing function of length in, for example, Germanic and Romance languages. Phonological analysis based on these ideas has seen a resurgence since Vennemann (1992/2000) and the emergence of experimental evidence for articulatory differences between German tense and lax vowels (e.g. Kroos et al. 1997; Mooshammer et al. 1999). The approach seems to have had success in explaining some historical aspects of language change (e.g. Murray 2000; Uguzzoni et al. 2003).

'Syllable cut' can be seen as the way that West Germanic languages can combine stress-accent with resonant vowels yet remain learnable. It protects child speakers from aerodynamic conditions that would otherwise threaten production during lax vowels.

6.4.3 Phonotactics

I have two short points to make under this heading. Firstly, Pickett et al. (1995:8) note an observation made by Stetson:

> "... only languages with strongly marked syllable stresses seem to make extensive use of syllable-final consonants; he cited as examples English and German which are heavily stressed and employ many syllables of CVC form as compared with French where there is less variety in syllable duration, and fewer CVC syllables."

This seems to be a natural consequence of the less energetic respiratory system pulses that French children, for example, would develop.

Secondly, Jakobson and Waugh (1987: 153) credit Martinet for the observation that highly aspirated plosives tend to co-appear with an /h/ phoneme in the sound inventories of languages. Clearly pulses of extra respiratory system activity facilitate both of these phenomena. French speakers of English, for example, often struggle with their production of /h/, and under my account this will be because they have never learnt to integrate pulsatility with EBP in their speech breathing.

6.4.4 Lengthening effects

In general, I am suggesting that 'compensatory shortening' effects are not timing phenomena *per se*, but are the result of only the resource from an invariant respiratory system pulse being available to be expended within the pulse's domain. The proposition that all the resource must be expended before the system can initiate a cycle with a new pulse may provide an explanation for word- and phrase-final lengthening effects. Alternative explanations are discussed by White (2002:282). See also Tye-Murray and Woodworth (1989:313), whose third proposal foreshadows mine, and Lieberman (1967), for another similar idea.

Applying this to the earliest vocalisations of infants (given the expected pulsatility of their SB), explains the otherwise puzzling data of Laufer (1980), who found final lengthening in the protosyllables of infants as young as 12 weeks. It also answers B.Smith's (1978:44) objection to a physiological motivation for final-syllable lengthening: that its extent varies in different languages. (Spanish, for example, showing a reduced effect compared to English.) This would be expected given the differences in the style of SB required by a child for each language[56].

6.4.5 Declination

Fowler (1996:549) reviews the experimental data on declination and asks if, "an account of declination as a dispositional consequence of respiratory changes during an expiration has been disconfirmed". She explains that in her view it has not, and continues,

> "This does not mean that declination is wholly unregulated by talkers. The fall in P_{sg} is not a simple reflection of lung deflation because, as noted, expiratory muscles are recruited increasingly during an utterance to offset the effects of the decline in the recoil force of the lungs. Apparently they often do not offset the reduction of the recoil force entirely. Why not? Possibly, they do not because talkers intend F0 to decline and that is how they implement declination. Alternatively, they may only offset effects of reduction in the recoil force on P_{sg} enough to ensure sufficient transglottal pressure for phonation out to the end of an utterance. Within that constraint, they allow P_{sg} to fall as the lungs deflate, and they allow F0 to fall with it. The latter account has the advantage of explaining why declination occurs so commonly across languages. The former account may have some validity as well however."

Fowler is discussing adult speakers when she describes relaxation pressure contribution and increasing expiratory muscle recruitment over the course of an utterance. The data

[56] See the longer discussion in Davis et al. (2000:1259).

for a young child's relaxation pressures presented in section 3.2 make it possible that he simply ignores the mechanical backdrop to his speech during the period when this is largely produced on pulsatile respiratory system activity. In other words, from a control perspective he uses the inspiratory increase in lung volume as a source of air volume, but not as a source of pressure (through involuntary relaxation forces). The magnitude of the expiratory pulse he produces is the same whatever lung volume it is made at. If this is the case, then when two or more feet are produced within one breath group, the first will be produced with a higher subglottal pressure than the next and, other things being equal, at a higher fundamental frequency.

Since the result will not be perceptually marked, there will be no reason for the child to modify this strategy until his respiratory mechanics change. At this stage, his natural tendency will have become part of his model for natural speech.

6.5 Summary

In this chapter I have considered the effect of a child deploying stress-accent on his pre-existing form of speech production. I have assumed that he implements stress-accent with a pulse of increased expiratory effort/activity created by the respiratory system musculature.

I looked in detail at three phenomena that only start to emerge at around the same time that a child learning a West Germanic language starts to deploy stress-accent (between 18 and 30 months, according to the studies reviewed in section 5.2.1).

I explained the first of these, foot level shortening (FLS), in the same way as pre-fortis clipping (PFC) in chapter 4. I argued that if weak syllables are reanalysed from the perspective of the load seen by the respiratory system, their 'vocalic' elements are not equivalent to full vowels. A foot then represents essentially the same production task in SB as syllables did earlier, although the 'frame and content' relationship between its aerodynamic and articulatory requirements now involves more complex elements.

The second phenomenon, that of tense and lax classes of vowels, was explained by examining the aerodynamic constraints of speaking with a small production apparatus. So on the one hand, to balance the various demands that this makes on him, a child warps his output in ways that result in temporal changes. On the other, the West

Germanic languages have evolved in such a way that they only allow sound sequences that a young learner will be able to produce safely. The result is a set of properties for vowels (including diphthongs and so-called reduced vowels), that are not arbitrary, as previously thought, but coherent responses to the demands of stress-accent.

That said, the vocal qualities associated with tense and lax vowels are not obviously consistent with this account. It will need to be developed further to explain this data satisfactorily.

I explained the third phenomenon, long lag VOT's, as a natural solution for young speakers of a stress-accent language to find when faced with the need to differentiate a second set of plosives. For reasons of time and space I did not give this proposal a full treatment. In future, I am confident it will be possible to demonstrate this more satisfactorily and, perhaps, to explain the fortis/lenis distinction within the framework I have described.

Finally, a number of other phenomena - including P-centres, 'syllable cut' phonology, lengthening effects and declination - all receive natural explanations within this framework.

In a fuller account I would have reviewed previous proposals that foreshadowed and informed my proposals on compression and other vowel length phenomena. Apart from the work of authors already cited, these included: E. Meyer (1903), Öhman (1967), Lieberman (1967:98), and Abercrombie (1965; 1967:36).

7 Problems with an imitative account of acquisition

It is generally believed that English speaking children acquire PFC, the WGmPh and the other phenomena I have discussed, by a process of imitation (meaning, in this case, modelling). I will shortly describe how this might occur and criticise the idea. However, in preparation for this and for Part 2, I will start with a discussion of imitation itself, defining terms and describing some mechanisms that can also lead to the replication of phenomena in social learning.

7.1 Imitation and other mechanisms which account for matching behaviour

The word 'imitation' is given many different meanings, both by ordinary speakers and in the literature. What follows is a short discussion of mechanisms that lead to matching behaviour that will establish the terminology I use and begin the development of a conceptual framework suitable for analysing the role of imitative processes in speech sound replication. In Part 2, I will develop this framework further.

There are various ways of describing the participants in social learning situations. As mentioned earlier, I will use 'A' for a (female) demonstrator, 'B' for a (male) observer, and, when it is necessary to be specific about maturity, 'C' for a (male) child.

7.1.1 Definitions

Thorndike (1898) famously defined imitation as, "learning to do something from seeing it done," and much of the imitation literature is similarly concerned with B being motivated to learn something from A. For child speech, we should keep in mind that this is not the only possible connection between the actions of the two, since a child may vocally imitate an adult with no intention of learning, but with various psychosocial motives instead.

The need for precision in the social learning literature has prompted various technical definitions of 'imitation' and related concepts. I will describe two approaches that have been taken.

In the first, Call and Carpenter (2002) take what B can obtain from observation as a starting point. They describe a multidimensional framework, where A produces three

sources of information simultaneously: about her goal (aim or intent), her actions and the results (changes in the state of the environment that are a consequence of her actions). Note that each of these may have hierarchical internal organisation, so there will be a main goal but there may also be sub-goals concerned with how to achieve it.

The three sources of information are independent, so it is not possible to draw conclusions about any one from the others. Call and Carpenter then distinguish four different activities (which could all be colloquially called 'imitation'):

1. Imitation (now understood in a technical sense): where B understands and adopts A's goal, copies her actions, and reproduces the result (a change of state in the world). If B fails to achieve the result, then we would describe this as failed imitation.
2. Goal emulation: where B understands and adopts A's goal but does not copy her actions. When successful, he reproduces the result.
3. Mimicry: where B either does not understand or does not adopt A's goal, but does copy her actions. Whether he produces the same result as her or not, is not important.
4. Emulation: where B either does not understand or does not adopt A's goal, does not (intentionally) copy her actions, but does produce the same result as her. (If B fails to achieve the result then we can either describe this as failed emulation or just no social learning at all.)

Want and Harris (2002) describe a similar taxonomy, extended to distinguish 'blind' and 'insightful' imitation. These can be mapped onto the framework by considering whether or not B understands the results he will create by his actions: if not, imitation is 'blind' (Carpenter and Call 2002:23). Heyes and Ray (2002) describe another refinement for 'imitation', between 'outcome-sensitive' imitation (where B is concerned with results but not A's goal) and 'intention-sensitive' imitation ('imitation' as described in point 1 above). However, these distinctions will not be important for my arguments, and I will use plain 'imitation' to cover all of the situations described.

A second approach is taken by Whiten and his colleagues. Whiten et al. (2004) propose that copying should be a superordinate concept within which imitation and emulation lie (see figure 7-1.) At this point, unfortunately, terminology diverges. Whiten (2002)

argues that the use of 'emulation' made by Call and Carpenter and some others perpetuates a use of the word which is at odds with Wood's (1989) introduction of the term into contemporary usage[57]. Whiten et al. use 'affordance learning' to describe what Call and Carpenter call 'emulation'.

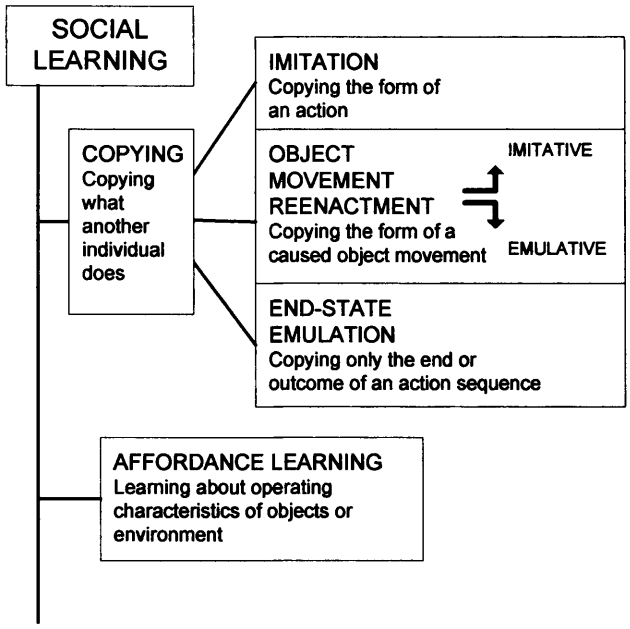

Figure 7-1. Detail from Whiten et al. (2004) Figure 1. "A taxonomy of social-learning processes ..." Original drawing includes detail to the right of and below the boxes shown.

Figure 7-1 draws attention to a problem with any idea of there being a simple imitation-emulation dichotomy. When A acts on an object, a tool for example, another source of information is created for B. He may now choose to copy the movement of the tool, rather than or in addition to the actions of A or the final change of state in the environment that her actions and the tool create. Huang and Charman (2005) recently demonstrated that object movement on its own (e.g. with the image of A edited out of an instructional video) was a necessary and sufficient source of information for young children to replicate behaviour.

---

[57] It seems to me that Call and Carpenter's use also diverges from how the term is used more generally: when we talk about George W. Bush having 'emulated' his father in becoming U.S. president (achieving the same office, but by different means), or of a piece of computer equipment being an 'emulation board' (producing the same output for a given input, but via different circuitry and programming). Here emulation is always purposeful, concerned with B either copying A's goals or the results she achieves or both. It seems odd to use the word when the same results are produced but without a copying process being required to have operated.

How can this new source of information be incorporated into each taxonomy? Whiten et al. (2004) discuss two examples, use of a hammer and use (by primates) of a rake to retrieve food. The way that B might pick up information from watching a hammer used seems likely to be similar to how he might view limb motion. So we might regard the tool as an extension of A's body, and say that B is imitating her when he copies the motion of the hammer. However, the coupling will not be so close with a rake, in which case the clear meaning for imitation in Call and Carpenter's scheme would be compromised: B's actions might well be very different from A's, and not derived from hers.

Alternatively, the motion of the hammer and rake could be seen as proximal results of A's actions, so B would be emulating A in reproducing their movements (not attending to her actions to achieve the result). Some authors take this line (e.g. Huang and Charman 2005), and there are nuances to the argument related to cognitive complexity and how representations of the object's movement are coded that are not important for my purposes but which might distinguish different tool usage (Rigamonti et al. 2005).

Whiten et al. (2004:40) argue that the most responsible solution is to describe the situation as 'object movement re-enactment' which may, "vary in the extent to which it shares characteristics with imitation."

This discussion is relevant because significant aspects of speech production are not visible to a learner. Speech sounds are intermediaries (like the motion of tools) which are meaningless in themselves but contribute to a final result that is meaningful to a listener[58]. It is common to talk about imitating or mimicking sounds, but this may obscure some subtleties of the situation. It may be preferable to say that a child is re-enacting a sound when copying its qualities (if and when this is possible).

In summary, then, I will provisionally define the four principal terms I will use as follows:

[58] I am putting to one side the question of whether speech perception is auditory or direct. If it is direct, then it might be more acceptable to use 'imitate' in connection with speech sounds.

1. **Imitation**: where B either understands and adopts A's goal, or seeks to reproduce the result of her behaviour, or both; copies some part of the form of her actions; and reproduces the result.
2. **Re-enactment**: where B either understands and adopts the model's goal, or seeks to reproduce the result of her behaviour, or both; copies some part of the form of the proximal effects she has on an object; and reproduces the final result[59].
3. **Emulation**: where B either understands and adopts A's goal, or seeks to reproduce the result of her behaviour, or both; does not (intentionally) copy her actions; but, when successful, reproduces the result[60].
4. **Mimicry**: where B is indifferent to A's goals or the outcome of her behaviour, but does copy some part of the form of those actions. Whether he produces the same result as her or not, is not important.

Of course, authors who I quote from will sometimes use 'imitate' with a broader meaning, as Thorndike defined it, for example. Also, while this taxonomy will be adequate for the discussion in this part, it will be revised and added to in Part 2 (particularly with respect to mimicry) to form the conceptual framework needed to consider the re-enactment of speech sounds.

Now, having attempted to draw some behavioural distinctions, we must recognise that in real life things are less tidy:

> "When humans imitate, we neither emulate another person's goal irrespective of the means to achieve it, nor do we mindlessly regurgitate action sequences without regard to what they accomplish. Rather, means and ends go hand in hand (so to speak) when we imitate an action." (Morrison 2002:115)

In other words social learning is undertaken for a purpose, and particular mechanisms are only employed to the extent that they further this end. This is described by Call and Carpenter (2002:219):

> "... it is a common experience among adult humans to watch someone achieve some result (e.g., with a new tool, or when learning to play a new sport or musical instrument) and then to attempt to reproduce that result oneself. If one's first attempt is

---

[59] In speech, the 'object' is the air that is imprinted with a sound pattern.

[60] So in order to transform my usage into that of Call and Carpenter, Want and Harris and some others, (1) "goal" would be inserted before each use of 'emulate' and (2) any use of 'affordance learning' ("discovering that a change of state in the world is possible" (Nehaniv and Dautenhahn 2001:22)) would be read as a plain 'emulate'.

unsuccessful, during the next demonstration one might pay more attention to the demonstrator's actions than to the end result."

Their diagram of this process (which is likely to be considerably longer and more complex in reality) is reproduced as figure 7-2. Call and Carpenter go on to ask if we should then say that B has learnt by emulation or imitation. Their proposed solution is for us to recognise that the question sets up an artificial dichotomy; instead, we should attend to the behaviour of B during each cycle of learning, breaking it down into its constituent parts of goals, actions, and results. I shall describe this overall process as 'learning in cycles', with B attending at any given moment to whatever he judges are his immediate information requirements. In one cycle he may be preoccupied with results only; in the next, attending to A's actions; then returning to the results; then considering her goal, and so on. I will use this approach when analysing how speech sounds are acquired, in Part 2.

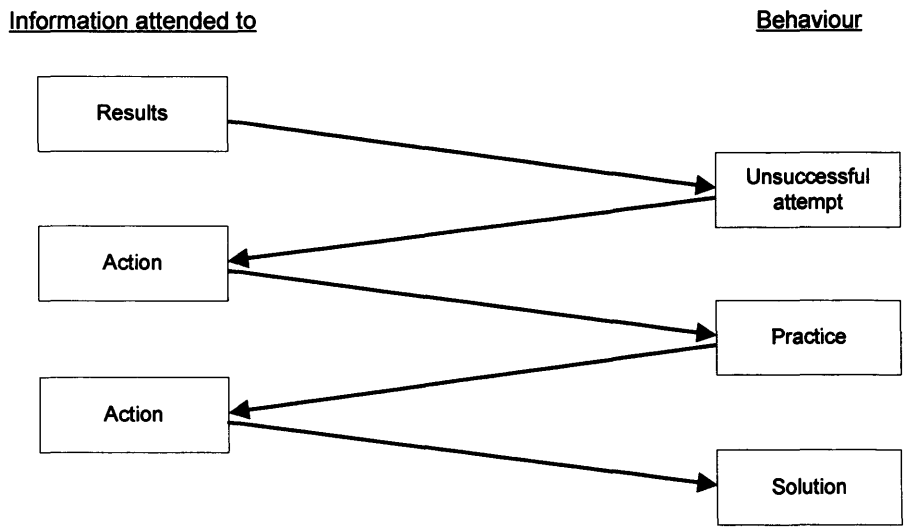

Figure 7-2. From Call and Carpenter (2002:220) Figure 9.3 "Shifting between sources of information in a social learning task."
The 'Practice' phase will include learning activity and an improved attempt.

Whiten et al. (2005) go a step further, viewing the processes involved in copying from another's actions as constituting a continuum ranging between imitation and emulation (see also Stoinski et al (2001:279) for discussion):

> "The core idea we wish to capture is that at the emulation pole of the continuum, the observer may be actively ignoring - "selecting out" - various aspects of the actions of the model, which at the imitative pole would instead be "selected in." We can envisage various contexts in which such a flexible, selective strategy could be adaptive ... For example, some actions may be seen as accidental rather than intended, or as not causally linked to an outcome of interest, and thus not copied. Or it might be that even though certain acts are perceived as intended and/or causally necessary, the observer has at his or her disposal alternative behavioural techniques for gaining the results of interest, which are preferred over those performed by the model ..."

Whiten et al. note that imitation only ever involves copying some part of the form of an act, and emulation some results of an act. Both imitation and emulation are purposive, by which I mean that at least part of B's purpose is derived in some way from A's behaviour. This would seem to be a notion that will both guide what he selects to copy (for imitation, "plausible causal actions" as Whiten et al. (2004:46) propose), and that will form a contrast with selection for mimicry. I will return to this when considering mimicry in more detail in section 13.2.

Several things are missing from my discussion so far. Most importantly there has been no mention of learning. Processes like imitation and emulation depend upon B having motor skills which can exploit the sources of information available to him from A. The fact that 'what to do' (the order or timing of events) can be learnt 'by imitation' (or, in this case, 'observational learning') sometimes obscures the fact that 'how to do' (the motor skills involved, including those for speech), cannot. There must be separate processes which achieve this (for example, in the box labelled 'Practice' in figure 7-2), and it will be important for us to understand how this learning may or may not be able to exploit what B can pick up from A. I will return to this in section 12.2, when I discuss the so-called 'correspondence problem'.

7.1.2 Other aspects of 'imitative' performance

We need to be sensitive to at least four other dimensions of B or C's behaviour during the replication of speech phenomena.

Levels

Mitchell (2002:444) describes levels of 'imitation' that can be identified, together with 'designing processes' for them. We will be most concerned with his third level which is,

> "... designed by processes of learning and memory, and has a program that recognizes resemblances and disparities between copy and model and induces more accurate

> reproductions of the model by diminishing disparities and accenting resemblances. Learning to pronounce a foreign language shows such a program in operation."

This, of course, is also how it is generally believed that children learn the timing and other phenomena I have been discussing in this part of this thesis, and speech sounds, which are the subject of Part 2.

Mitchell's earlier levels are characterised by simpler design processes (e.g. 'create a copy when you perceive a model'), more often seen in animal behaviour. Later levels include both 'pretend play' and calling attention to the fact of imitation to communicate something. For example, exaggerating actions to indicate imitation, as in satire or caricature; or when children simply imitate other children as a means of communicating mutual interest.

Both Locke (2001) and Want and Harris (2002:10) point out the value to a child of imitating its caregivers for promoting social communion/interaction with them. As I mentioned earlier, it seems that during the period of speech acquisition we must be aware that some instances of vocal imitation (in the broad sense of the word) will not be concerned with learning at all, but examples of performance with other motivations.

Time of performance

So called 'on-line' imitation occurs when there is an immediate perception-action coupling by B, mediated in some cases by (fast decaying) iconic or echoic memory. Clearly this is a feature of the early learning of speech (with the mother typically imitating the child rather than vice versa (Pawlby 1977)), and also of some language teaching situations.

There can also be a lag between observation and performance ('deferred' imitation). Vogt (2002b:531) points out that,

> "For deferred imitation, it is ... possible to assume temporally distinct stages of perceptual processing and a later "translation" of the resulting (verbal or iconic) representation into action."

Representations

Vogt (2002b:535) explains that representations which underlie the repetition of an action that B originally made in response to an action of A, may be of at least three different types:

1. a sensory image captured from A; and/or
2. a sensory image created by B of his initial or subsequent responses (i.e. auditory images in the case of speech); and/or
3. an image in motor format of B's initial or subsequent responses (that is, a representation in terms of the motor commands B generated as a response to A's action).

These images, "can play different, overlapping and similar roles in imitation tasks" (ibid). In Part 2, I will suggest that creating a temporal hierarchy by reversing the order of these representations as they are listed above, might yield an insight into the nature of underlying representations for speech sounds.

Means of performance

Finally, two aspects of the psychological processes that underlie learning during imitative episodes will also be important to my arguments.

With respect to the overall process, Moerk (1989:289) states that, "imitation is based on pattern abstraction and pattern reconstruction." I have been using the term 'modelling' to describe this.

With respect to attention, Vogt (2002:545) reports that sport scientists, "emphasize that for successful training schemes that involve modelling, attentional focusing to specific aspects of the display is essential." This takes us into the realm of motor skill learning and refinement, which is relevant to the WGmPh because the main factor suggested for the delay before a child's performance on these reaches adult values is inadequate neuromotor control. In the skill learning literature there seems to be some controversy about whether attention is better directed to movement technique or to the outcome (e.g. in a golf swing, (Wulf et al. 2000)), but not as to the necessity of attention being paid in order to improve a skill.

7.1.3 Imitation in children

In reviewing a recent collection of papers on imitation, Wilson and Woodward (2002:537) describe, "[A convergence] on the conclusion that actions are represented beyond the level of motor patterns," with both adults and children representing action in terms of its goal structure.

Experiments conducted by Bekkering and his colleagues (for example, Gattis et al. 2002) show young children attending to goals in preference to actions, with this demonstrated in children as young as 14-months by Gergely et al. (2002). In other words, if children can infer a goal, they do not copy actions for their own sake, only as a means to an end[61]. We can certainly expect this to be a general principle for children at the later age when they start to acquire the WGmPh.

Although I will shortly be criticising an imitation-based account of the appearance of the WGmPh, I recognise that imitation plays a major part in the lives of young children. In language (as opposed to speech) development, Moerk (1989:290) describes 10 subsets of imitation identified in mother and child behaviour ('identical imitations', 'reduced imitations', 'substituting imitations' and so on, all the way up to 'quotations'), and reports at least 50 examples per hour of such imitations by the children in his corpus. Bloom et al. (1974) also demonstrate the usefulness of language imitation, so that, "the mapping and coding relation between form and content [in a given situation] can be affirmed."

7.1.4 Problems of interpretation

Finally, a recurring theme in the imitation literature is the reminder that matching behaviour cannot, of itself, be taken to imply a process of either imitation or mimicry in its production.

So, for child speech it is important to note that a match in behaviour with adult speech can also be produced as a result of emulation, because children, "sharing a common if somewhat differently sized form of life tend to employ the same parts of their body [as adults] in similar ways to achieve common ends." (Wood 1989:72)

Galef (1988) describes an impressive number of non-imitative processes that have now been found to explain behaviour that was once thought to be imitative. On this theme, Noble and Todd (2002) warn that,

> "Human observers of animal and robot behaviour have a propensity to invoke mechanisms that are more complex than those strictly needed to explain the observable

[61] There is also evidence that children sometimes overcopy (e.g. Whiten et al. 2005:280), but this seems explicable as an intelligent response to the particular situation they were presented with.

> facts ... quite possibly this has something to do with the human tendency to interpret the world from an 'intentional stance' (Dennett 1987)."

(They continue with the remarks in the epigraph of chapter 8.)

7.2 The control and variability of timing in speech

The WGmPh are largely timing phenomena. At 18 months of age a child has not 'acquired' them, but he is then imagined to do so by a process of topographical learning which may take many years (until full, adult-like mastery is achieved).

To examine the plausibility of conventional accounts of how this happens, we need to consider how timing is controlled in speech, and what aspects of timing we can expect to be learnt explicitly.

Klatt (1976) is the classic reference for a view of segmental timing as a controlled parameter, which is learnt as such from infancy onwards. In Klatt's model, detailed patterns of timing are explained by the combination of (i) effects due to physiology (or phonetic 'inevitability'), (ii) effects of phonetic context, and (iii) paralinguistic effects of rate, emphatic stress etc. The model includes a,

> "... phonological component contain[ing] rules that weigh all of this information [about syntax, segmental representation and stress pattern] and specify an abstract underlying duration for each segment that is to be produced."

This approach to speech timing suggests that there is a basic durational target for a phoneme which is modified by a (large) number of factors, some of which are the result of learning, some not.

Timing effects due to physiology (as often proposed to account for the differences in VOT as a function of place of articulation) are assumed to be the result of the speaker foregoing control at this level. He chooses not to compensate for timing differences either because these anyway exist in the speech of the models around him or because they are below the level of perceptual 'just noticeable difference' limens.

On the other hand, the durational effects of phonetic/phonological context are assumed to be learnt (albeit with many researchers expressing reservations about the complexity of this in practice). The conventional accounts of the WGmPh fall into this category.

Finally, paralinguistic effects may modify speech via 'global' variables, which may or may not originate in imitative learning.

7.3 Acquiring the West Germanic phenomena via imitation (modelling) of their timing

If a child learns PFC, the WGmPh and so on through imitation, then we can adopt the approach suggested by Call and Carpenter (2002), above, and ask about the goals, actions and results of this. I will use this framework to group some critical thoughts under five question headings:

With respect to his goals,

(1) what underlies the child's topographical learning?

With respect to his actions,

(2) how does the child first identify durational targets and then use them in his speech?

(3) what evidence do we have of the acts of neuromotor learning that are supposed to be taking place?

With respect to the results,

(4) is the path to adult performance consistent with an imitative process of acquisition;

(5) why is the final result in adults so variable?

I would not claim that any of the points I can make here is conclusive. My intention is only to establish that while an imitative mechanism for the learning of the WGmPh is widely accepted, it is not unproblematic.

It is worth mentioning again that the child is able to communicate satisfactorily before his speech exhibits the WGmPh, and could probably continue to be understood satisfactorily without them.

7.3.1 Goals: what underlies the child's topographical learning?

For imitation-based change to occur, the child must (i) notice discrepancies between the durational characteristics of his speech and that of those around him, and (ii) be sufficiently dissatisfied with his own performance to be motivated to remedy this situation.

Noticing discrepancies

Contextual changes in VOT's, syllable lengths, etc, can be subtle. Weismer (1979) found reliable durational effects which are below a practical 'just noticeable difference' in speech, but for which he could find no plausible mechanical explanation. To support an imitative account, then, we have to suppose that from an early age the child is extraordinarily sensitive to fine phonetic detail (that is sometimes linguistically non-significant), in his own and others' speech.

Yet, as Menn and Stoel-Gammon (1995:350) point out, "young children generally tend to view language as a means of communication, with primary focus on content and use rather than the form of an utterance." And children often seem unaware of gross discrepancies between their speech and that of others.

Dissatisfaction with his own performance

Why should a child be dissatisfied when there are minor and unimportant timing discrepancies between his performance and that of adults?

It is possible, perhaps, that he wishes to speak like an adult for reasons similar to those that are supposed to motivate speech accommodation (Coupland and Giles 1988; West and Turner 2002), i.e. to elicit approval, maintain a positive social identity, or achieve communicative efficiency. Perhaps he aspires to participate as a 'professional' rather than an 'amateur' in a common code of communication.

I can imagine some children experiencing dissatisfaction with their speech at some times for some of these reasons; but the phenomena in question are learnt by every child. Unless they cause communicative difficulties, it is hard to see minor discrepancies in speech timing distressing all children.

And for change to continue over the many years that the learning of the WGmPh take, there must be dissatisfaction at every stage (Locke 1996), even when there is no issue of others misunderstanding the child (for example, when VOT's adequately distinguish cognate pairs, but development toward adult values continues).

Apart from asking why a child should be dissatisfied, we can also ask if there is any evidence that children are dissatisfied by their performance with respect to the WGmPh. They don't seem unduly troubled by far more significant differences in the fundamental frequency, formant patterns, etc, of their speech, so long as they are understood by their interlocutors.

7.3.2 Actions: how does the child identify and use durational targets?

Flege and Eefting (1988:730) describe the structure of imitation of timing:

> "Imitation is generally regarded as consisting of three distinct processes: perception of structural properties in the stimuli being imitated, coding and storage in memory, and regeneration in the form of a motoric code suitable for skilled movement."

With respect to the first of these, speech in natural contexts is of varying quality, and is typically uttered in conditions where the child's priorities will be to attend to its underlying message and form his response. Circumstances are not ideal for the observation of fine phonetic detail.

However, assuming that the child is motivated to attend to the relevant data and capable of this, he now has to identify its structural properties. Variability due to physiological mechanisms has to be recognised and ignored; differences due to paralinguistic factors have to be factored out; phonetic context and the interaction between phonetic contexts have to be correctly analysed; speaker variability has, perhaps, to be averaged away; variability of tokens within speakers (due, for example, to production at different lung volumes (Hoit et al. 1993)) averaged as well; and so on.

As Port (1981:262) puts it: "How can timing be an effective source of phonological information when it is subject to such a variety of overlapping distortions?"

Researchers attempting to emulate even small subsets of this feat, with consistent data sources, have failed to do so in any linguistically natural way (Gopal 1996; van Santen and Shih 2000).

7.3.3 Actions: what evidence do we have of the acts of neuromotor learning that are supposed to be taking place?

For learning motor skills, practice helps (Welford 1976). But repetition is not the same as practice. Just saying a word as part of an utterance does not improve one's pronunciation if one's attention is elsewhere. Practice in a speech context requires saying something while attending to the quality that is to be changed[62].

What evidence is there of children practising their speech with respect to PFC, the WGmPh etc, as opposed to just speaking, when their attention will always be elsewhere? On the occasions when children do practise speech on their own, I would be surprised to discover that the focus of their attention had been on any of the qualities of the WGmPh, as opposed to qualities of sounds and words that create linguistic distinctions.

7.3.4 Results: is the child's path consistent with an imitative process of acquisition?

The timing of children's speech moves towards the adult norm. We assume that the extended period that this takes must reflect development of their imitative abilities. Which of these will change with age, and thus explain the long period of apprenticeship?

1. Presumably the child's perceptual ability improves, so he can become more sensitive to the fact and nature of discrepancies.
2. His internal model of speech will improve, with the addition and analysis of extra data resulting in better targets for production.
3. His neuromotor apparatus will mature, giving greater potential control; and the opportunities for practice he has enjoyed will enable him to translate this into greater actual skill[63].

---

[62] The ineffectiveness of repetition on its own is illustrated by the many speakers of a second language who reach plateaux with respect to their pronunciation.
[63] Shattuck-Hufnagel (1986:80) gives a similar analysis.

All of these developments should assist imitation. We would therefore expect that his progress will be linear – one of continual improvement. In fact, the development data paint a different picture.

It would be impractical to review all the studies demonstrating this. However, the following quotations from B.Smith and Kenney (1999) illustrate the point. (They undertook a longitudinal study of four children of various ages, following them for between four and six years):

> "A general conclusion that has emerged from the collective results of a number of studies regarding acoustic characteristics of children's speech production development is that both the duration and temporal variability of segments, words, and phrases tend to decrease as one considers increasingly older groups of children. While this conclusion is largely appropriate when considering averages for groups of children studied across intervals of several years, the limited amount of longitudinal data that exists suggests that similar patterns may not routinely be observed in the development of individual children especially when shorter time periods are involved."
>
> "Sometimes older groups have been observed to have longer durations or greater variability than younger ones, and this is then commonly assumed to be due to a sampling problem or other methodological factor."
>
> "On the basis of the present data, it seems more reasonable to expect that a given child will commonly demonstrate **plateaus, reversals or other variations in the development of temporal parameters of speech production quite different from group trends**."
>
> "Assuming that a given child's neuromotor system generally improves with age and does not tend to regress, the present findings suggest that additional factors must be involved in accounting for the reversals observed in temporal measures of speech among individual subjects."

So one aspect of the developmental data is a variability in individual performance which is inconsistent with an account of ever improving abilities to imitate.

7.3.5 Results: why is the final result in adults so variable?

There are now many studies that show greater variability between adult speakers for durational phenomena than was once appreciated. Some recent examples include:

- Smith et al. (2003) finding ratios for PFC in native speakers ranging from just over 1.2 to just over 1.8.
- Allen et al. (2002) finding long lag VOT's ranging from 74 to 100 msecs, even after correction for relative rates of speech.

- B.Smith (2002) finding wide variation in 16 selected measures of temporal patterns as rate was varied.

Yet speakers are able to achieve impressive convergence of timing at segmental, syllabic and phrase levels if motivated to do so. Choral singing provides one example, the synchronous speech paradigm reported by Cummins (2003) is another. Given that we have this capacity to match speech timings faithfully if necessary, why do we observe so much variation in normal speech performance if it is based on imitated models? And if the speech timing phenomena in question are important enough to motivate a child's imitation of them, why do adult speakers appear to be so unaware and indifferent to the variability in timing demonstrated in the studies cited above?

I suggest that the issues I have raised under these last five headings make the acquisition of surface timing properties by imitation rather implausible. I shall weigh this against the BSD approach in the next chapter.

7.4 Changes in production strategies

In the next chapter, I will discuss the extent to which BSD play a continuing role in shaping adult speech. As a preliminary to that, I will now briefly consider the question of how timing relationships between segments are controlled.

It is possible to imagine direct control of timing in adults, without requiring that this is also true for children. Changes in a control mechanism may occur over development.

Ohala (1970:143) described two possible models for the sequential generation of syllables (which correspond to the 'comb' and 'chain' models also referred to in the literature):

> "A 'Timing-Dominant' system, i.e., a system which maintains a tight time schedule perhaps at the expense of precise and thorough accomplishment of the gestures.
>
> An 'Articulation-Dominant' system ..., i.e., a system which maintains precise and thorough performance of the gestures no matter how much time it takes."

Ohala went on to point out the possibility of hybrid models for speech production, for example where words are executed one after the other (according to the chain model)

but the gestures within words are executed according to a strict time plan (i.e. following a comb model). See also Löfqvist (1997) and De Jong (2001).

Vihman (1996:231) describes an application of the idea of two types of timing strategy to explain, in this case, "the paradoxical findings and interpretations of coarticulation studies of school age children.":

> "Hawkins (1984) outlined a U-shaped curve of this kind, involving first a shift from the 'timing-dominant' speech production basis of babbling to a segment tied 'articulatory-dominant' system, which results in slower and more variable segment production than is seen in adults, and then, at age 7 to 9 years, a return to a timing-dominant system."

Shaffer (1982:111) gives an example where a motor skill is performed by a subject in a timing-dominant fashion under one set of circumstances, and an articulatory-dominant fashion when those circumstances change.

My proposal of a frame and content relationship between respiratory system activity and actions of the upper articulators, would be an example of articulation-dominance. The question for the next chapter will be whether this persists, or becomes part of a general timing-dominant strategy of speech production.

7.5 Summary

I started this chapter by defining imitation and some other forms of social learning. I then turned to how timing is controlled in speech, and questioned the plausibility of a child acquiring the WGmPh by a modelling process. Finally, I returned to the control of timing in speech, to point out that children and adults may differ in how they achieve this.

8 Evaluating the accounts of replication, and further issues

> "It has long been recognized within fields like artificial life that complex global phenomena can arise from simple local rules, and this is precisely what we will suggest is happening in many social information processing contexts: individuals follow a simple rule (e.g., "stay close to your mother") and, in combination with some form of learning, the overall pattern of behaviour that arises makes human observers suspect complex imitative abilities."
>
> J. Noble and P.M. Todd, *Imitation or something simpler?* (2002)

In the previous chapter, I argued that an imitative account of the replication of the speech phenomena I have been discussing is implausible. A breath stream dynamic (BSD) account is not vulnerable to the criticisms I made there.

In this chapter, I will reflect on some other issues which bear on which account is more likely to be correct. I will then extend this to consider different ways in which BSD might be influential (different 'strengths' of the ideas I have presented), and how adult performance can be explained.

Finally, I will touch on some issues to be investigated in the future.

8.1 Explaining phonetic patterns in children

8.1.1 Resilience of phenomena

In section 2.1, I asked how it was that the WGmPh can be such resilient features of English, given that they make it harder to learn to speak the language, and that it could dispense with many, perhaps all, of them. If the WGmPh are temporal phenomena that are learnt by imitation, then this is puzzling. Languages can evolve away from such difficulties (Fowler 1985:197; Menn 2000:754).

However, I am arguing that the reason why the WGmPh are replicated is not because they are part of a communicative code, even though listeners will make use of the temporal cues they reveal for word recognition. Instead, the phenomena are replicated in production as unintended expressions of stress-accent, consequences of speech being

embodied. As such, their resilience is explained, as is their non-appearance in non-stress-accent varieties of English (those of Singapore, the West Indies, etc).

8.1.2 Variable paths and variable results

In the previous chapter, I pointed out that evidence of regression in the development of timing patterns is, at least on the face of it, incompatible with an imitative account. Similarly, it is odd that phenomena like PFC show such variability across speakers, given our abilities to control speech timing very precisely when necessary.

However, in the BSD account the phenomena I have been discussing arise from embodiment. I mentioned earlier that Boliek et al. (1997) favoured a dynamical systems framework for interpretation of their SB data (see also Stathopoulos 1995, 2000). Hawkins (1979) applied similar reasoning to speech timing, in particular. She favoured,

> "a theory of the development of speech timing within a polysystemic, parallel processing approach ... [where] adults and children will differ in speech production processes not so much in the nature of those processes as in their relative importance and domain of influence. The child's 'system' cannot be regarded as static at any time, but rather as reflecting the effects of several continually changing systems that replace each other during development. Changes in one subsystem may affect others, producing either progression or temporary regression."

Within this paradigm, then, a non-linear path to adult levels of performance and the variability then observed would be natural and expected, not something requiring explanation.

8.1.3 Task complexity

It must surely add to the plausibility of a BSD account that it does not require a young child to become a junior phonetician at the same time as he has so many other demands on his attention. He must become very familiar with his own production system, but beyond this he is only required to talk, to notice when his production is unbalanced, and to discover simple ways to resolve these issues.

In other fields, it has been found that the ways that people actually solve problems is not by the rational, numerical analysis that one might use if deploying the tools of an engineer, a physicist or an economist to model a system. (E.g. McLeod and Dienes (1996) on how to catch a cricket ball in the deep, and the articles in Gigerenzer and Todd (1999) on how to make decisions using fast and frugal heuristics.)

In speech, Cutler and Swinney (1987:163) endorse a similar suggestion from Bolinger with respect to accentual focus. He claimed that a child's pitch rises at points of interest for physiological reasons simply because the child is excited, not because a communicative code has been analysed, internalised and now deployed.

8.1.4 Effect of power supply

The BSD account is based on a speculative, but surely reasonable, view of speech breathing (SB) in young children. SB is a complex skill, learnt at the same time as speech itself, during a period of physical changes and changes in linguistic demands (e.g. the need for longer utterances, the adoption of stress-accent). As a result of all this, I have argued that a child must adopt a simple style of SB as a starting point for speech. This can be refined towards the smooth adult model over a period of time.

When a system has a continuous, stable power source, one has to look for explanations for variation in final output elsewhere. But a variable power source for speech, as for other systems, would have the potential to condition every stage of production that follows.

I may well have some of the details wrong in my BSD account (cf. section 6.2.5, on tense and lax voice qualities, and section 4.2, where figure 4-1 cannot be realistic), but if a child's early SB is not stable and smooth, then it will surely play more of a role in the final output than we have realised so far.

8.1.5 Opinions among researchers

As I mentioned in the previous chapter, researchers in speech development have often suggested that natural mechanisms to generate the variability they find in durational patterns must exist, without being able to specify them. Others have contended that adult phonology grows out of the child's, and not vice versa. (e.g. Ferguson 1986). A natural mechanism, such as the one I have described, is in line with both of these viewpoints.

8.1.6 What would be evidence against the BSD account?

There are some simple ways in which the BSD account would be undermined. For example, by the existence of any of the following:

- A dialect of English, German or Dutch that exhibits the WGmPh, but not stress-accent. Then these phenomena would presumably be being replicated by imitation. The converse would also be problematic; however, the dialects of English that I am aware of that lack the WGmPh - those of Singapore and the West Indies - don't use stress-accent for prominence.
- A long vowel/diphthong in a stress-accent dialect of a West Germanic language variety, which had low vocal tract resistance.
- The chronology of any child speakers showing the WGmPh emerging before stress-accent is deployed.

I am not aware of any such evidence.

8.2 Explaining phonetic patterns in adults

I have described how the BSD of a child's speech might condition his output. However, there are other ways in which the effects I have described might shape the phonetics of English.

1. In a very weak version of my theory, BSD pressures might have been completely phonologised at some point in the past, and are now learnt by imitation (cf. Anderson 1981; MacNeilage and Ladefoged 1976:94; MacNeilage and Davis 2000).

2. In a weak version, the phonologisation might have occurred, learning may be by imitation, but BSD may have a continuing influence by constraining innovation away from the WGmPh. That is, any child who tries to speak English without replicating the WGmPh will encounter difficulties that will encourage him to conform rather than to continue along an idiosyncratic path[64].

   Fischer-Jørgensen (1990:135) points out that even if differences are controlled by a speaker, one can still seek physiological or acoustic explanations for them because, "they may be embodied in the knowledge of the speaker and included in the norm of

[64] The French way of writing the numeral '7' with a horizontal bar may be an example of this type of replication. French children must be increasingly exposed to 7's written without central bars. However, the handwritten form of the numeral '1' in French is much closer to an unbarred 7 than the simple vertical stroke taught to children in many other cultures. So any French child who innovates away from the French model for a 7 risks confusion with the French style of 1. French children learn to write a 7 by imitation, but persistence of this style in the face of simpler alternatives is regulated by aspects of the system that only become apparent when an innovation away from the standard is considered or attempted.

the language because they are physiologically economical and perceptually useful."

This would explain the resilience of the WGmPh, even when replicated by imitation.

3. Among stronger versions, until now I have not properly distinguished between two possibilities. BSD might canalise the output of child speakers, but might not be a continuing pressure in adult speech. If adult speech is timing-dominant, then one might explain the persistence of PFC and WGmPh effects by suggesting that adults continue to speak as they did as children. The models of 'correctness' they developed then, continue to guide their production[65].

4. Alternatively, BSD pressures may be a continuing force in adult speech, as allowed by the pulsatility that Finnegan et al. (1999; 2000) discovered.

I suspect that the answer may lie somewhere between these last two possibilities. The question of whether or not the respiratory system is involved in routine stress production may not be the point. Sometimes it is, sometimes it isn't, but it comes to be unimportant with respect to the timing of speech.

English speakers are not trying to achieve a flat loudness contour, so it is sensible for them to retain a pulsatile capacity in the organisation of their SB. However, how they realise loudness changes may be decoupled from how they plan speech.

If styles of SB can move along the continuum I described in section 3.1, speakers may plan the articulation of a foot so that it can be executed at any point by any legitimate strategy including the highly pulsatile style of the Accent Method. The actual style used is 'decided', after planning, in some other way (presumably based on cost, effort, context, etc). So planning continues into adulthood as if SB were an articulation-dominant factor, meaning that English speech is coherent across whisper, normal speech, shouting etc.

8.3 Future work

Clearly some form of empirical investigation of these proposals is an important next step. Unfortunately, SB is hard to investigate instrumentally at the level of detail

[65] Dick Hudson pointed this out as being rather simpler than some alternatives I had been considering.

corresponding to phonetic segments. The idea that effort is being controlled further complicates the task, as this cannot be measured directly.

However, in some circumstances aerodynamic power may provide a reasonable proxy for effort. The precise measurement of pressure and airflow at appropriate points will still be challenging, particularly, of course, in young children. Pressures below the glottis can be measured directly, but only with invasive techniques. Volume change can be inferred from airflow, but if this is measured at the mouth then jaw movements and the compliance of the vocal tract create artefacts. Measurement at or below the glottis may be possible, but again this involves invasive procedures[66]. Volume change can also be inferred from the flow of the chest wall.

It may be that lesser goals are more achievable. Testing the idea that P-centres are correlated with peaks in the output of the respiratory system might be reasonably straightforward. Testing the need for the child to made some of the accommodations I have described, for example to avoid turbulence in tense vowel production, might be amenable to aerodynamic modelling.

There are other aspects of the proposals that require investigation. I have not tested them against all the dialects of English, so their compatibility with the Scottish Vowel Length Rule, for example, should certainly be looked into. Further issues for future investigation include the evidence on anacrusis and initial syllable deletion by children; the history of stress-accent; and the implication of these ideas for theories of syllabification.

There is a summary of Part 1 in chapter 16, together with the overall conclusions of the thesis.

[66] The audience of a talk at the University of Arizona in early summer 2004 suggested that children with Treacher-Collins syndrome would be minimally inconvenienced by the measurements required, given their existing treatment regime.

Part 2 – The Replication of Speech Sound Qualities

9 Introduction to Part 2

> "[Imitation] is a transposition, a translation of what the child hears, in phonetics that are familiar to him. The best way to describe a child's imitation of language is to compare it to his drawings. He transposes the model, using the symbol that he knows how to reproduce (this is a man, a head, a hand) and forever trying to reduce to a minimum the effort to copy accurately. As soon as the symbol is understood and accepted by those around him, it fulfils its functions. Similarly, at a later stage, the child (and also the adult who is not well educated) remains incapable of reproducing accurately the word he hears. He transposes it, believing he is repeating it."
>
> P. Guillaume, *Imitation in Children* (1926:50)

The first word-shapes produced by a child from an adult model are probably 'imitated' in the sense of being re-created holistically. However, the child comes to a stage where he can acquire words 'by imitation' in a different sense (and the one that concerns me in this thesis). The new process involves 'transposing' (as Guillaume uses the term) the sequence of speech sounds in a word that the child hears into a sequence he produces, where each speech sound produced is taken to be equivalent to the appropriate adult one by his listeners.

This raises the question of how a child learns the correspondence between each speech sound heard and what he needs to produce to match this. I.e., how does he learn to imitate speech sounds prior to being able to learn words 'by imitation'. Is this matching process itself achieved by a process of 'imitation' – the copying of sounds by the child through producing sounds which he judges to match those he hears – or by a different process?

In this part I will explain why I think it is unlikely that children learn speech sound qualities by copying them through acoustic matching. Instead, I will describe a child who cannot match speech sounds but who can, with help, directly associate a speech sound he hears to an action he performs (all, in fact, that is required of him). There are, I think, good reasons why he has no alternative; why the child is at least temporarily 'disabled' in comparison to how we have viewed his capabilities up until now.

To make the idea of a direct speech sound to action association and its consequences clearer I will start with a thought experiment, in which an infant has no choice but to learn to communicate this way. I hope that it will seem plausible that speech learning can still occur in these apparently adverse circumstances. I will then describe how my

main argument will unfold in this part and how an integrated picture of the development of both production and perception will emerge by the end.

9.1 Learning to speak in Ruritania

Imagine that we are in Ruritania, where a linguist-ruler in centuries past implemented a number of language reforms. So Ruritanian has a one-to-one phoneme to grapheme correspondence, and deaf Ruritanians do not use a separate sign language but a manual coding of the spoken form, called 'signed Ruritanian'. Some, at least, of the cheremes/phonemes in signed Ruritanian are realised with movements that are quite natural: the sort of movements an infant will anyway make when exploring the use of his hands, arms and other articulators at an early stage in development.

All Ruritanians are fluent in signed Ruritanian, which is learnt together with oral speech in childhood.

It happens that in one family a mother and her child have intact sight and hearing, but different disabilities that affect their speech production (figure 9-1). The mother can speak but cannot make manual gestures (although she understands signed Ruritanian). Her child can make gestures, but is unable to speak. He does, however, have some exposure to people other than his mother, who he observes signing; this is enough to provoke him into a phase of silent (manual) babbling.

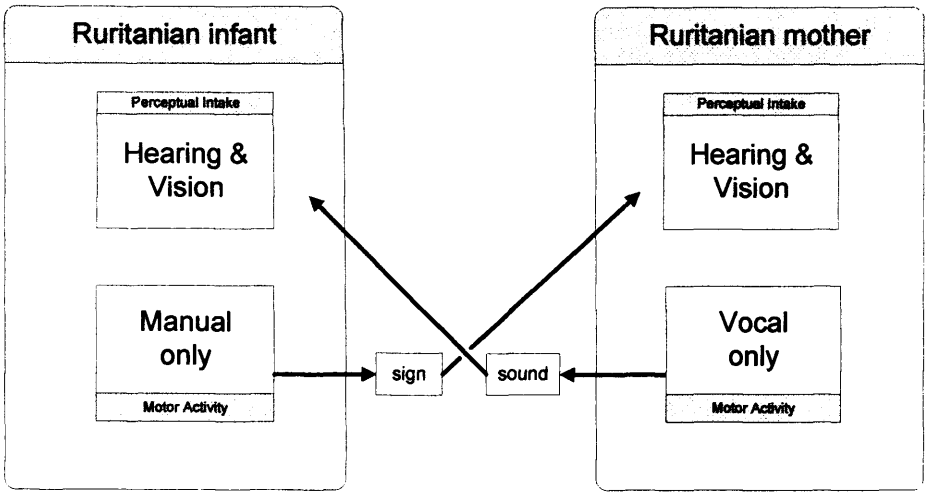

Figure 9-1. An infant who can sign but not speak, and his mother who can speak but not sign.[67]

When, by chance, the infant first makes some gestures which more or less correspond to *mama*, his mother is delighted. Aware of the effect his gestures have had, the infant might build on this by creating further signed protowords, some of which would also taken by his mother to be meaningful. Alternatively or additionally, he sees others signing in what appear to be goal-directed ways, and through all this he increasingly understands the communicative potential of manual movements. However, he does not have enough contact with signers to be able to copy any of what they do.

His mother talks to him, of course, and he begins to understand words in Ruritanian. Further, as just described, she responds to those of his signed protowords that she comes to recognise. However, if this were the extent of his linguistic interactions with the world it seems impossible that the infant could learn the signs and words he needs to express himself in signed Ruritanian. There is nowhere he could learn them from.

What, though, if his mother starts responding to his gestural babble and protowords by reformulating them into speech? So although the infant is not intentionally signing the sounds/letters which form Ruritanian words, his mother acts as if he is, judging that some of his gestures are similar to those making up the standard movements for particular sounds/letters, and transducing these to their imagined spoken counterparts.

[67] I would like to thank Ian Howard for developing this style of diagram.

Some of the segments involved will have been movements that, during manual babbling, the infant mastered to the point where they are stereotyped and repeatable. We can call these movement motor schemes. So the infant's experience will be that when he produces a movement motor scheme, on its own or embedded in a larger pattern of gestures, there will be times when,

1. it is clear that his mother is saying something she takes to be equivalent to what he has just done; and
2. on these occasions he hears back a sound (or group of sounds) that is reasonably consistent, to the extent that he can build examples into a perceptual category.

If this style of interaction happens regularly, then the child will realise that his movement motor scheme and the sound category he has developed for what he hears his mother saying are equivalent in her eyes.

When he now recognises one of these sound categories within one of her words, he can reproduce its equivalent movement. His mother recognise this as part of a signed word and responds accordingly. His active word adoption for expression has begun, and can continue in this fashion.

9.2 Structure of part 2

It may not be obvious what parallels can usefully be drawn between the Ruritanian situation and the learning of speech sounds and words by a normal infant. Why has the acoustic modality of communication shared by real mothers and their infants been replaced by an asymmetry of sound and signing? I hope to demonstrate that the situation described is actually a much better analogy for a real infant's predicament than it may at first appear.

In the next chapter, I discuss some preliminary issues.

In chapter 11, I make a short excursus into speech perception. I propose that early word recognition is the result of learning 'by acquaintance', and point out that the categorisation of speech sounds (as opposed to sounds in general) could be done on the basis of functional equivalence rather than perceptual similarity.

In chapter 12, I ask how children come to be able to adopt words from their linguistic environment, and then discuss existing accounts of speech sound development.

In chapter 13, I return to the subject of imitation itself, expanding on the treatment I gave in Part 1. This enables me to identify a number of potential problems with the accounts of speech sound learning previously described. These show why infants may indeed be unable to match speech sounds on an auditory basis.

In chapter 14, I propose an alternative mechanism. As in the Ruritanian example, the child makes use of others' judgments of speech sound similarity or equivalence. He makes the rather different judgment that he is able to do something which is taken by others to be equivalent to their output. He starts to learn this pre-linguistically, during 'imitative' interactions when his mother responds to him as if his output were well-formed in L1. She reformulates his utterances, showing him what these are equivalent to in her own speech. Thus he starts the word adoption process with the ability to match some categories of sound that he can identify in her speech with movements he is able to perform. This style of learning - a variety of mirroring - may have already been informing him about his affective state for some time before it features in vocal interactions.

In chapter 15, I explore the implications of word learning by this alternative mechanism, considering various psycholinguistic issues in speech. I demonstrate how my proposals would resolve some longstanding problems. Then I describe the integrated model of speech perception and production that emerges from them.

My full argument is summarised in section 16.2.

10 Preliminaries

There are four topics which are most conveniently dealt with prior to the main discussion: types of noticing, types of perception, conceptual units of production and perception, and vocal activity prior to first words.

10.1 Noticing

Mason (2002:33) has analysed noticing. He discriminates what he calls 'ordinary-noticing', marking and recording:

> "To notice is to make a distinction, to create foreground and background, to distinguish some 'thing' from its surroundings."
>
> "It is useful to distinguish between *ordinary-noticing*, or perceiving, in which sufficient memory is established accessibly to be jogged and reconstructed by what someone else says, and *marking*, in which not only do you notice but you are able to initiate mention of what you have noticed. Ordinary-noticing is easily lost from accessible memory. It is only available through being re-minded (literally) by someone or something else. To *mark* something is to be able to re-mark upon it later to others. *Marking* signals that there was something salient about the incident, and re-marking about it to someone else or even to yourself makes the incident more likely to be available for yet further access, reflection and re-construction in the future. Thus *marking* is a heightened form of noticing. Intentional marking involves a higher level of energy, of commitment, because it requires more than casual attention.
>
> Ordinary-noticing, or perceiving, provides the rich backdrop of experience on which learning depends, but in itself is insufficient. It is distinguished from *noticing* which refers to all aspects of moving from ordinary-noticing or perceiving, to marking and recording, and to various practices which support these. Marking provides specific data to work on. But even incidents that are marked may be overlaid by or merged into subsequent events. There is a third level of intensity or energy in noticing which fuels *recording*. The desire to make a note, to record in some way, may be a product of our written-visual culture, but it does play a significant role in personal development. Recording could use words as in a list, journal, or creative writing, but might be expressed in some other medium including performance. By making a brief-but-vivid note of some incident, you both externalise it from your immediate flow of thoughts, and you give yourself access to it at a later date, for further analysis and preparation for the future. You can note something inwardly, making a *mental note* and initiating a state in which you might choose to re-mark at some future moment, and you can outwardly note by making some record. Tying a string around your finger used to be a clichéd form of noting in order to re-member. The incident recorded becomes an object not only to analyse, but also a component in the building of rich networks of connections and meaning.
>
> Recording definitely requires motivation, for it takes extra energy beyond that required for marking."

The key distinction for my purposes will be between (i) 'ordinary-noticing', demonstrated by memory reconstruction through recognition but not, it seems, available

for phenomenal report, and (ii) marking/recording, which makes it possible to initiate mention, or remark upon, what has been noticed. It seems implicit in this latter definition that marking is a prerequisite to being able to evoke something in imaging, imagining or mental rehearsal.

10.2 Awareness of Sensation (AS) and Meaningful Perception (MP)

10.2.1 Different perspectives on perception

There is a longstanding issue in psychology concerned with correctly characterising two flows of information to conscious awareness that can be called 'sensation' and 'perception'[68]. I will argue that this issue is also important for speech acquisition.

When a door closes we can attend to the resulting acoustic signal and hear the sharp rise and swift decay of a rather 'flat' sound. Alternatively, and more usually, we use the signal to perceive the event that created it: we hear that someone has shut the door, for example.

Öhman (1975:42) explores the difference between these two modes of listening at a finer level. He characterises the former as an immediate awareness of the developing states of our auditory sense. We are experiencing the varying condition of our hearing mechanism, the changes which occur in the physical system that we ourselves are. I will call this mode of listening '(conscious) awareness of sensation' (AS). From the perspective of the stimulus, the signal is attended to for itself, without any meaningful interpretation drawn from it[69].

In contrast, Öhman characterises the second, more usual mode of listening, as an awareness of the developing states of an external physical system, the world. I will call this mode 'meaningful perception' (MP). I am adding the 'meaningful' prefix both to

---

[68] However, these terms are unsatisfactory because 'perception' has (at least) two potentially confusing meanings: 'to be in contact with' and 'to have a mental experience of something'. For this reason I will shortly introduce the names I will use instead.

[69] I am using "awareness of sensation" with some misgivings, firstly because it does not unambiguously convey, in itself, the idea that the information it supplies is meaningless and secondly because it uses words which come with a lot of baggage. Alternatives I considered were "bare perception", "meaning-free perception", "neutral perception" or "context-less perception".

indicate that this is a mental experience and to contrast that experience being meaningful to the meaningless nature of bare awareness of sensation[70].

Öhman's (1975:42) own description refers to his wife, who he can hear (but not see) moving around the kitchen, opening the refrigerator door etc.

> "In the way I am listening, I listen to these events. I do not listen to the sounds of the events. I could listen to the sounds of the events, however, if I wanted to. I would then listen to them as a sort of concrete music, disregarding their physical meaning. This latter sort of listening ... consists in an immediate awareness of the developing states of my auditory sense. As such it is a form of perception, viz. perception of the states of my own body.
>
> Thus, there are two kinds of perception, each of which consists in an immediate awareness of something. Ordinary [meaningful] perception is an immediate awareness of the developing states of an external physical system (e.g., a woman moving around in a kitchen). Perceiving in this way, my mind 'constructs' the external system-state configurations.
>
> [Awareness of] sensation, on the other hand, is an immediate awareness of the developing states of the physical system that I myself am. Perceiving in this way, my mind 'constructs' internal system state configurations."

MacKay (1987:65) describes AS as an awareness of, "the pattern of sensory stimulation."

In practical terms MP is the more important flow of information. One vital function of our senses, and those of every living being, is to deliver information about the changing state of the environment. Survival may depend upon it. In fact, the mode of perception that does not deliver anything meaningful – AS – seems hard to maintain in anything like a pure form. We may only be regularly attending in this mode in situations like eating (relishing the taste of a dish, perhaps), or listening to music (although even here MP will often create images and structure[71]).

Öhman's descriptions make clear the very different natures of AS and MP. I will now make two tentative assertions (from introspection):

---

[70] In some ways I would have preferred to use "categorising perception" for "meaningful perception", but it seemed there would then be considerable scope for confusion with "categorical perception". Öhman himself used "ordinary perception", and Neisser (1994) used "recognition/representation" for what I take to be the same basic idea. Following Jeannerod (e.g. 1994) I might have used "semantic" (and "pragmatic" for AS).

[71] As someone who is ignorant about the natural environment, the way I experience birdsong is largely through AS. Friends who are more knowledgeable might extract something meaningful from the same signal.

1. *Marking events in both modes of listening at the same time is not possible.* (Linell (1982:67) makes a similar assertion.) Thus an ephemeral signal may be amenable to being attended to either for retrieving a veridical sound image or for its 'meaning(s)', but not for both. Note, though, that this assertion does not deny that it is possible to retrieve meanings from multiple planes[72] simultaneously within MP.

2. *Switching into AS when there is meaning available in a scene may require considerable presence.* In a speech context this seems related to an observation made by Repp (1984:321):

> "The interest of [categorical perception] lies largely in subjects' strong resistance to adopt a mode of listening that enables them to detect subphonemic detail. That this resistance can be overcome by appropriate methods and training is one of the most significant findings reviewed here."

But though the resistance can be overcome, Bruner et al. (1956:50) comment on the related issue of concept attainment by pointing out how,

> "curiously difficult [it is] to recapture pre-conceptual innocence. Having learned a new language, it is almost impossible to recall the undifferentiated flow of voiced sounds that one heard before one learned to sort the flow into words and phrases. ... In short, the attainment of a concept has about it something of a quantal character."

It is straightforward for us to be in MP mode. It is our usual way of attending to signals. When we move into AS mode, we nevertheless can and usually do apply meaning-making activities to the results: activities like recognition, categorisation, comparison, and so on. These characterise MP, so most of the time AS might be better labelled as AS/MP. However, it is also possible to attend to a signal in a state of pure AS.

The AS/MP split is illustrated by the exercise of copying a line drawing upside down or right way up (Edwards 1979; see figure 10-1). When people without artistic training or aptitude (like myself) are asked to do this, the results we achieve by turning it upside down are often better than those obtained with it the right way round. When the picture is correctly oriented it seems that we find it difficult not to parse the scene for meaningful elements (i.e. to ask ourselves the question "What am I seeing?"), and then

---

[72] Werker and Curtin (2005) use "plane" rather than "level" to resist an implication that meanings need be organised hierarchically (i.e., in a speech context, that segments would have to be retrieved to build syllables, syllables to build words, etc).

to reproduce our typical representations for these elements rather than the actual instantiation of them in front of us. Inverting the page allows us to operate without the distracting presence of meaning and its associated preconceptions, and so to attend more closely to lines and shapes as they really are. The final (re-inverted) result then resembles the original more closely than the projection of our 'object hypotheses' onto the page would (Gregory 1970:115).

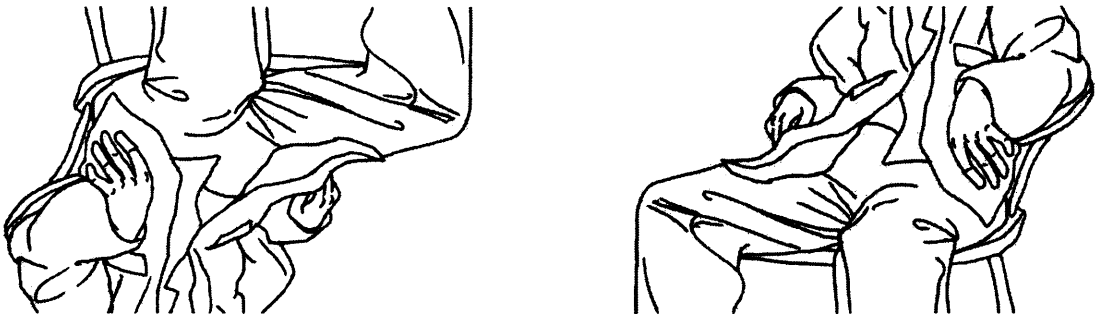

Figure 10-1. A picture is attended to differently for copying if it appears meaningless (left) or contains meaningful elements (right)

The distinction between these situations is sometimes described as drawing 'what you see' versus 'what you know'. For both adults and children, task and context are among the determiners of which mode predominates (Vinter 1999).

Clearly what I am aware of when the picture is upside down is not the most basic level of sensory input to my system. Integration of signals from each eye, for example, and some other low level visual processing will have occurred. Nevertheless, the image is free of meaning: nothing in it contributes to any cognitive operations.

A third assertion follows on from this:
'Surface-level copying' or mimicry (which contrasts with the reproduction of a meaningful or purposive act) is an AS function, in both vision and hearing. The criterion of resemblance involved is most insightfully viewed as similarities in the varying states of the physical systems that make up our visual and auditory senses[73]. Any instance of mimicry may, of course, be interpreted by an observer through MP, but it will be successful as mimicry only to the extent that non-meaningful resemblances are recognised.

[73] Vogt (2002a; 2002b) makes an apparently similar suggestion for what he calls parametric and action imitation.

I return to this in section 13.2.1, but in preliminary support I would observe that since it is possible to attend to a signal in AS mode and be consciously aware of its effects, and it is possible to make comparisons between such experiences and to decide that they are more or less similar, then it would be surprising if such an activity had not been recognised and named. I am suggesting that it is mimicry, and that the other main imitative processes (those which are 'purposive') operate principally on the results of attending to signals in MP mode.

Analogous ideas

A number of theoretical notions seem to be consistent with, or an alternative expression of, something like an awareness of sensation (AS) vs. meaningful perception (MP) framework:

1. There are developmental models that distinguish between stages of sensorial and perceptive coding of events, e.g. that of Vinter (1986) that leads to sensorimotor and perceptuomotor 'coordinations'.

2. Griffiths and Warren (2004) describe the alternative possibilities of an acoustic 'event' (i.e. the signal) or a sound's 'source' being the auditory object of attention during listening. Put otherwise, the 'proximal stimulation' versus the 'distal source'.

   Note that the source need not only be the physical apparatus generating the signal. It might also be the thought that composes the production. In a piece of orchestral music, then, the sources could reveal themselves as any of (i) the instruments (via the structure they impart to the signal), (ii) the composer (via, for example, the melody), and (iii) the conductor and players (via the structure they add through interpretation). In speech, similarly, the vocal tract, the word spoken and the idea behind the utterance would all be 'sources' (cf. Vallabha and Vallabha, in preparation).

   The identification of phonemes, syllables, onsets and rhymes, and so on, is also an

example of sources being revealed by the MP rather than the AS mode[74].

3. In audition, dichotomies of cognitive processing have been used to account for experimental data on discrimination. Durlach and Braida (1969) described two memory modes in a model of sound intensity discrimination that they called 'sensory-trace mode' and 'context-coding mode', and this framework was applied by Macmillan et al. (1988) to speech sounds. The modes seem to equate to what Pisoni (1973) called 'auditory mode' and 'phonetic mode'. He suggested that the former holds a fast fading but reasonably veridical acoustic image; the latter a classification of the signal that is more stable.

Dorsal and ventral streams

A different light on the AS/MP framework may be shed by recent research into brain function. Jeannerod and Jacob (2005:301) start their review of the so-called 'two visual systems model' as follows:

> "Although seeing is commonly experienced as a unitary activity, the scientific understanding of human vision resists such a simple view. Both psychologists and neuroscientists consider that the processing of visual information is distributed across several different routes which eventually reach different functional outcomes, and that these processing routes can be mapped onto well-identified anatomical subdivisions of the visual system."

The concept and techniques for investigation of dissociable streams have been extended from vision to hearing. Middlebrooks (2002) says that, "[a] prevalent working model for auditory cortical research," is that of dual streams of processing in the brain corresponding to object localisation and object identification. These are a dorsal (posterior or caudal) stream which is often described as the 'where' stream, and a ventral (anterior or rostral) stream that is described as the 'what' stream (Arnott et al. 2004).

My definitions of AS and MP are phenomenological, but as Neisser (1994) and Norman (2002) demonstrate, it is possible to connect perceptual experience to the dual streams evidence. Their proposals have a broad scope but my requirements are more limited and the review below reflects this. I am principally concerned with the learning of a motor

---

[74] Norman (2002:96) notes a personal communication from Neisser that, "the dorsal system [AS] ... does not categorise, not even on the basic level."

skill, so my AS and MP describe conscious awareness of the products of brain activity (though not awareness that can necessarily be verbalized)[75]. There will, of course, be much activity in both streams that occurs without conscious awareness.

Mechanisms for dissociated 'where' and 'what' streams of perceptual processing were first described for vision. The original function of the dorsal 'where' stream was taken to be the mapping of the location of objects. Milner and Goodale (1995), working with a subject, DF, whose brain lesions allowed her to act on the orientation of objects in the world but not to recognise these orientations, redefined the localisation role of the dorsal stream to one of guidance/control of motor behaviour. (In this sense, Scott (2003:101) says that the dorsal stream can be considered as a 'how' pathway.)

Thus the visual control of behaviours like grasping and walking is in some sense extra-perceptual, independent of visual experience, or operating with little or no conscious awareness (Westwood and Goodale 2001); along, in fact, with much other work done by the visual system including the synchronisation of circadian rhythms with the light-dark cycle and the visual control of posture (Goodale and Humphrey (2001:312). So the visual processes necessary for the 'how' of guiding action could be described as visuo-motor transformations (Gentillucci and Negrotti (1994) quoted in Norman (2002:81)).

Note that Westwood and Goodale also point out that the ventral stream will sometimes play a role in the planning and control of action. For example, where an action requires the identification of an object (in order to access stored information needed such as, say, its probable weight), then both streams will be involved in fulfilling the task. Jeannerod and Jacob (2005) also argue for a richer characterisation of the model, particularly with respect to dorsal stream processing.

As I did when discussing AS and MP, we can ask where the various types of imitation might fit into this scheme. Since in Goodale and Milner's conception, the dorsal stream is responsible for visuomotor control, it is the natural place to locate the observational component of simple mimicry (surface level imitation), including the copying of an upside-down drawing[76]. On the other hand, the reproduction of any act based on its

---

[75] In section 12.2, I refer to the evidence for some effect of so-called implicit learning on how to act.
[76] In March 2006, at the Institute of Cognitive Neuroscience, London, David Milner presented drawings of an apple, a book and a yacht first copied and then drawn from memory by patient DF. DF has no

meaning (whether it is executed with actions that resemble those of the model or not), will presumably principally involve the ventral stream. See Vogt (2002a, 2002b) for a properly argued exposition of this idea.

Wise et al. (2001) postulate the existence of an auditory 'how' pathway as part of, or perhaps a distinct subsystem within, the dorsal, 'where' region (Scott 2005:199). This would parallel Milner and Goodale's findings for vision. Wise et al. report results compatible with a hypothesis that "the posterior superior temporal cortex is specialised for processes involved in the mimicry of sounds, including repetition," (i.e. sounds being used to direct articulatory muscles: sound-to-articulation rather than sound-to-meaning (Scott 2003:420)). This region is active during both silent articulation and normal speech production. See also Scott and Wise (2004:27) for discussion, Hickok and Poeppel (2004:86-90) for a not-dissimilar account of auditory-motor integration in the same dorsal stream, and Warren et al. (2005) for an argument to characterise the dorsal auditory stream as a 'do' pathway[77].

Jones and Munhall (2000:1250) describe evidence in Houde et al. (2000) suggesting that the perceptual system which controls action is separate from the one which is used for category judgments (deduced by comparing (i) auditory cortex response to subjects hearing their own speech while it was being produced, to (ii) the response when it was played back later from a tape recording). This may explain how we can respond to bite block perturbations within one vocal fold cycle – well before we are aware of the auditory disturbance caused.

Whether a brain organisation of dissociated streams in visual and auditory perceptual processing is innate or emergent is an open question (van der Kamp and Savelsburgh 2000:240; see also Norman (2002:128) and the commentaries he discusses), but it seems reasonable to assume that by the time an infant starts word learning, some dissociation will be operational.

The simple model that I am working from, then, has the functions associated with processing an unenriched signal situated in the dorsal stream (e.g. object location

---

ventral stream function so she could not 'see' either the targets or her own drawings. Nevertheless, some features of the targets were clearly capable of being copied by dorsal stream visual activity alone.

[77] In addition, I have recently come across Kraus and Nicol's (2005) view of the 'what' and 'where' pathways that seems very different from what I have described.

through comparison of intakes at each eye or ear, and visuo/audio-motor mapping including mimicry). The functions associated with enrichment of the signal through previous experience to create mental percepts are situated in the ventral stream.

Details of this model may be wrong without affecting my general argument. The limited point I need to make is that a phenomenological division of hearing into two modes of perception finds a place among similar proposals with many other bases, and its implications for speech sound development are therefore worth exploring.

10.2.2 Consequences of meaningful perception

Having described two different modes of perception, I will now investigate the flexibility we have to move between the two; in particular, to move from meaningful perception to awareness of sensation when listening to speech[78].

For most of our conscious interactions with the world, MP will be the more important and customary mode, since it supports the skill of recognition of the source of a signal. We are generally involved in a "quest for meaning" (Norman 2002:89) at all levels from the mundane upwards. Indeed, consciously attending to the bare results of AS may be quite rare.

However, I have proposed copying a line drawing as one example of the conscious use of AS (and upside-down copying for the unskilled), and another example comes from speech, when we attempt to pronounce a new word that we do not expect to be able to reproduce within a phonological system we control. So for 'foreign' words – the name of a person or place heard in a stream of otherwise recognisable speech, perhaps – the following steps may occur. (I have put the account into the first person since it is based on introspection.)

1. I ask for the word to be repeated. (I probably do not have what I consider to be an adequate image to copy from available from my normal listening.)
2. In most circumstances I can be confident that the word will, indeed, be repeated so I can release myself from the need to verify this (which I would do by listening for recognition in the MP mode).

---

[78] I am grateful to Rachel Smith for a conversation about some of the issues in this section which led to considerable revision of an earlier version.

3. I am also aware that I will have to go outside the faculty within MP that would retrieve phonemes or other phonetic categories for this task if the target word was in English. So I put myself into a different perceptual set, that of AS.
4. I capture an image of the word to the best of my ability, aware that my representation is neither durable nor, perhaps, trustworthy.
5. So allowing myself only a short delay, I re-create it with whatever resource I can bring to bear. I 'keep in mind' the previously captured target image (or, as Öhman might phrase it, 'attend to a remembered image of the previous development of my auditory sense') using it to drive whatever facility I have for sound re-creation. I then compare what I hear of myself ('the developing state of my auditory sense') to the target image.
6. I may be successful to a greater or lesser degree, but note that in recreating a sound pattern this way, using the target image in order to release whatever behaviours will create a sensory match to it, I do not learn <u>how</u> I am producing the word. Some time later, asked to say it again without the support of a fresh model, I would struggle with the best of what my auditory memory could supply to me as a new target for re-performing step 5.
7. On the other hand, if I have subverted the process in step 5 to some extent, and allocated some of my attention to watching my articulatory system at work, then I will have some resource to draw on to reproduce the word later.

I discuss the later stages of this process in section 13.3, but for present purposes what I find striking is the preparation of myself to capture a sound image in step 3, and how watchful I need to be to maintain that perceptual set. Along similar lines I described the difficulty of copying a picture, where for the graphically unskilled our predisposition to discover and reproduce meaning prevents us from 'drawing what we see' (i.e., for us MP blocks AS). Vinter (1986:341) describes a similar situation with respect to how Chinese characters can initially be seen as shapes but how this facility is lost when learners begin to understand them, and both Jenkins (1980:225) and Gregory (1970:117) point out the need for extensive "training of the eye" if novices are to become artists who can perceive what they see rather than what they know.

Phoneticians, too, require training to cultivate the ability to hear their own language objectively. In the absence of this, most people find it hard to transcend (i) the alignment of discrimination to the ambient language that has started to take place at the

end of the first year of childhood, and (ii) the "sympathetic reconstruction" of what they hear others say (Nathan 1999:313), which means that they don't normally perceive allophonic variation, elisions and so on. Meaning derived from MP dominates over AS in both cases. Similarly, McMurray et al. (2000:19) describe how in many experimental situations the actual VOT of an individual stimulus appears to be discarded, with only category membership (the result of MP) remaining in the percept. To keep one's attention with the result of AS when meaning is present to be attended to is hard.

While in AS mode we do not, by the definition given earlier, recover any meaning. I would suggest that in MP mode we do not recover bare sound. That is, to the extent that my attentional set is to seek meaning I will not perceive words or the sounds within words with their veridical acoustic qualities.[79]

In some models of speech, auditory processing has been pictured hierarchically. The raw signal is coded into phonemes (or some other phonetic categories), these are recoded into words, words to meaning and so on. Subordinate representations to the one specifically attended to are 'thrown away'. I prefer the accounts of word recognition in Polysp (Hawkins 2003) or Öhman (1975:39), for example, where the acoustic signal can inform any level of attention to meaning without it necessarily moving through a hierarchy.

However, the important point for what will follow is that attention to the basic signal is a qualitatively different process from this. Fast switching of one's attention between the two may be possible, but perhaps only after substantial training/experience (because there is normally no reason to dwell in AS), and only when any tasks of recognition are easily accomplished[80].

It may be that the strict consequences of this are tempered by the operation of sensory (echoic) memory. If we do quickly switch from MP to AS after recognising an

---

[79] Linell (1982:67) makes a similar assertion: "In normal speech perception, we do not focally attend to the sounds at all. Rather we use fragmentary information at the phonetic level to recognize words, phrases or even 'meanings' directly. In fact, **it seems impossible to search for 'meanings' and at the same time attend to the phonetic properties themselves.**"
And on the other hand: "**As soon as we make ourselves focally aware of the phonetics of the utterance we lose the integrative, higher-level perspective** (cf. Polanyi 1969); afterwards we would have great difficulties to report the semantic content of the text we heard."

[80] A more rigorous description of this would have to explain some of the data reported by Gerrits and Schouten (2004).

anomalous sound (or failing to recognise a word) then there may be sufficient persistence of the stimulus for us to perform a rudimentary analysis of what has occurred. I may be able to say that "something like an /x/" was produced, for example.

Several lines of argument that I will pursue later will rely on a set of working assumptions that can now be summarised as follows:

- We have two modes of perception, AS and MP.
- At any one moment, we can be consciously attending to an event in one mode or the other.
- When meaning emerges out of sensory input it normally draws our attention away from the bare signal, i.e. from AS to MP.
- Moving in the other direction, from perceiving meaning with MP to attending to a signal objectively with AS, may take practice. However, the difficulty of doing so varies with experience and context, and with training we can certainly make such a process fast and smooth.
- In adults, at least, sensory memory may allow some access to the results of AS after a stimulus has been understood through MP. However, what remains is unlikely to be good enough to use as a basis for re-creating the signal. If this is desired, adults will request a 'good instance' of the target, adjust their attentional set into AS mode, and track what they hear.

Figure 10-2 shows the framework into which I will be fitting speech processes in chapter 15, when I describe an integrated developmental model of the production and perception of speech.

Following the discussion above, the two modes of perception are shown separately, flanking production which is differentiated into babbling, protowords (L0) and the child's first language (L1).

Time moves from the top to the bottom of the page. The results of the activities to be depicted will be four separate processes (in dashed boxes) giving the child the abilities to (1) re-create sounds and sound patterns, (2) reproduce words, (3) shadow speech, and (4) recognise words.

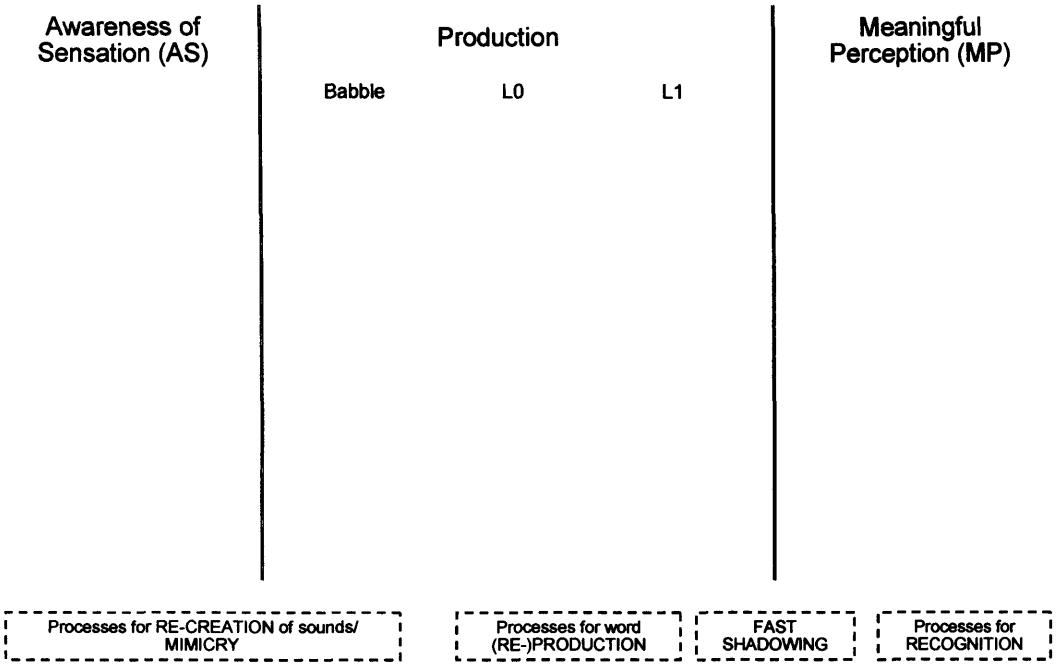

Figure 10-2. Structure for the portrayal of speech processes and their results. (The central area of this diagram will be filled in chapter 15.)

10.3 Terminology for perception and production

Units of speech

It seems unlikely that children learning to speak operate at a phonemic level of analysis (Ladefoged 1984). The smallest unit of production, after all, is a syllable, and MacNeilage and Davis suggest that apparent segmental independence within the syllables of babbling is illusory. (This is discussed more fully in the next section.)

I will use the terms 'speech sound', 'soneme' and 'movement' to describe the child's discrete units of sound, their psychological representation and the motor activities that produce them, respectively. To expand a little on these definitions:

1. The first follows Guenther et al. (2006:283):

   "A 'speech sound' is defined here as a phoneme, syllable, word, or short phrase that is frequently encountered in the native language and therefore has associated with it a stored motor program for its production. For example, we expect all phonemes[81] and frequent syllables of a language to be represented by unique speech sound map cells."

   They explain the link to Levelt's mental syllabary as follows:

[81] But phonemes are likely to be a later development than syllables.

"The model's speech sound map cells can be interpreted as forming a 'mental syllabary' as described by Levelt and colleagues. Levelt et al. (1999) describe the syllabary as a 'repository of gestural scores for the frequently used syllables of the language.'"

2. A 'soneme' will be potentially larger than a phoneme, but shares the characteristics it has when the latter is conceived of as a psychological entity. Thus a soneme is a conceptual category realised by phonetically different speech sounds that speakers classify as members of the same category[82].

3. To differentiate between the acoustic output and the way it is produced, I will use 'movement' to mean a unit of speech sound motor production[83]. Movements will be labelled with Greek letters.

I will use slanted // and square [] brackets to indicate sonemes and speech sounds as they would phonemes and allophones. So at a given moment in development the infant might have developed sonemes including /i/ /pi/ and /pit/ into which he classifies the speech sounds [i] [ɨ] and [i:], [pi] [p^hi] and [pi:], [pit] [p^hit] and [pi:?] produced by his mother.

In my diagrams, letters from the end of the alphabet, 'x y z', are used as algebraic variables rather than particular IPA symbols. They are written in lower case when used for adult sonemes, but in upper case for the child ones. I make this distinction to try to keep some symbolic correspondence between child and adult categories (for ease of understanding) while recognising that developing categories need not be identical to or even a subset of what will be the mature category. So the use of an upper case letter in /X/, for example, indicates that this category will evolve into the soneme /x/ in the mature speaker. Thus the perceptual category where the majority of the speech sounds of the mother's /x/ are likely to be assigned by the child will be labelled /X/ although the possibility (and likelihood) is that some of these sounds are not initially recognised to be in this category, and that some other sounds are initially assigned to it even though they will eventually be recognised as belonging elsewhere. /X/, then, is likely to be similar but not identical to /x/.

---

[82] According to Donegan (2002:6), Baudouin de Courtenay spoke of a phoneme as, "the psychological equivalent of a speech sound," so the use of a term similar to this seems appropriate.
[83] In fact I will use it rather more loosely than this, since it will not be necessary to explore the implications of sub-syllabic versus syllabic constellations of gestures.

Input and output

In MacKain's (1982) discussion of 'linguistic experience', she points out that mere exposure to a stream of speech does not necessarily mean that any particular linguistic or other structure will be experienced (perceived):

> "... the speech environment as perceived by the infant has been regarded as having a static structure; that is, it does not change as the infant develops, and speech input correspondingly has been specified independently of the way that infants perceptually structure that speech input. Given the possibility of developmental change in processing capacities for speech, we would expect changes to occur in the way(s) that infants structure speech input during the course of perceptual development, and descriptions of this input would not necessarily be synonymous with adult descriptions."

I will try to be sensitive to this issue by using terminology that distinguishes between what is presented to a listener's ear and what he can or does perceive. There is a similar issue in production, between what a speaker produces and what others hear him to have produced. Thus for an infant listener/speaker, I will use the following terms:

input	what his mother produces
intake	what the child perceives of this
output	what the child produces
outcome	what his mother perceives this to be

I will use 'intake', then, to mean that part of the input that the child is presumed to become aware of perceptually. There may be several reasons for differences between the input and the intake. First, the physiological characteristic of the infant's auditory apparatus may impair processing. Second, the brain processes (including cortical processes) that transform the earliest representation of the signal into higher order stimuli, develop over time. For example, the well-known 'reorganisation' of consonant perception at 10-12 months (and vowels perhaps earlier), results in infants not discriminating differences they previously heard. Thus at least some older English-speaking children are unable to distinguish between French words like *dessus* and *dessous*. (If we choose to, we can make ourselves sensitive to such differences again, of course.)

The notion of the input and the output are familiar. In the speech development literature they usually do not refer to the signal *per se*, but to what an idealised listener would be capable of hearing in the signal (i.e. the intake of a sensitive adult).

10.4 Production prior to the transition to words

In this section I will discuss five aspects of production that predate first words but which are believed to play a part in the performance and/or the choice of what a child comes to say.

10.4.1 Motor, auditory and proprioceptive (MAP) information

Menn et al. (1993:424) provide a useful summary of the sources of information available to an infant during the learning of speech, starting with the adult-generated input/intake, which will be auditory and (for a seeing child) visual.

Self-generated sources include (i) the motor commands the child makes to articulators, (ii) the sounds he creates, and (iii) proprioceptive feedback from a range of non-auditory sensations,

> "... that go along with vocal production but which are generally overlooked: sensations of motion of tongue, lips, jaw, and chest; sensations of airflow through mouth, nose, and trachea; sensations of air pressure; sensations of muscular tension; senses of position of articulators; and sensations of contact (lip against lip or front teeth; different areas of the tongue against various parts of the mouth etc.), and the rates of change of all the above."

Clearly this proprioceptive feedback will play a role in informing the child about the actions he has taken. The auditory feedback will also play that role, which is independent of any use of it the child makes to compare his output to that of others.

Furthermore, when the child uses auditory feedback and proprioceptive feedback to monitor the consistency of his output, there is no need for the auditory feedback to be equivalent to the sidetone (i.e., what others will hear him producing). That is, while the second potential use of auditory feedback (to facilitate copying activity) requires that what he hears of himself is similar to what others hear in order to be successful, the first use requires only that it be consistent over time. (It is no different, in this regard, from the other proprioceptive feedback, which plays no role in informing the child about what others can perceive of his activity, just that it was similar or different in certain ways to what he has done before.)

The hearing child is therefore in a privileged position in a number of ways as compared to a deaf child, one being that his ear can be fashioned into a sensitive tool to reveal

regularity and inconsistency in his articulation. The apparently more important advantage of hearing his own output of speech sounds in order to match it to that of others, is not necessarily the reason why a hearing child's pronunciation is usually superior.

10.4.2 Inverse/forward models and passing control to the ear

The human and robotic motor control literature has developed the terminology of inverse and forward models to relate action and the results of action (see Wolpert et al. 2001). In brief, an inverse model (IM) provides the motor command which will cause a desired change in state of a system. A forward model predicts the next state, given the current state and the motor command.

Figure 10-3 illustrates the setting up of the inverse and forward models. As alternative terminology (which may be more transparent), some authors use 'emulator' for forward model (e.g. Grush 2004) and 'behaviour' (loosely) for inverse model (Demiris and Hayes 2002).

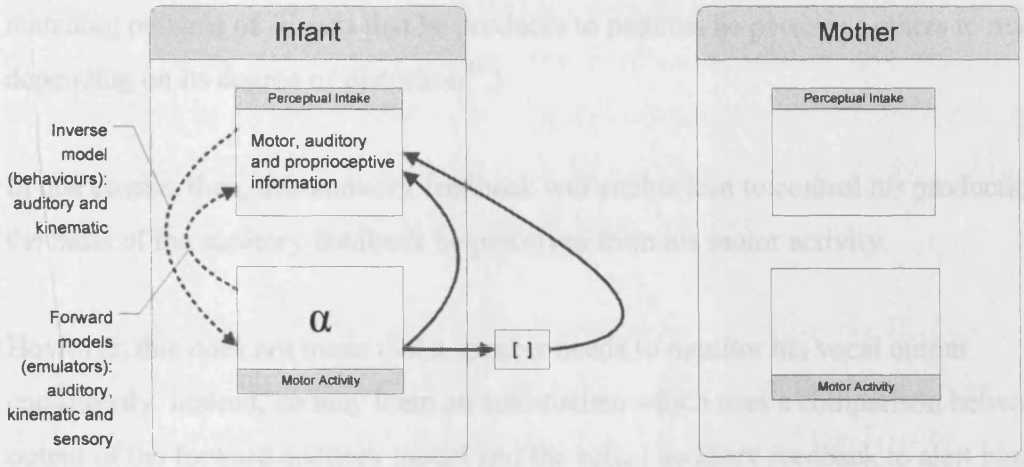

Figure 10-3. The creation of inverse and forward models. The source of a dashed line generates the information or activity that the line indicates.

In Menn et al.'s summary of motor learning quoted from above, they further point out that,

> "What one learns in learning to speak includes the simultaneous and sequential correlations between motor commands and this rather large set of [proprioceptive] sensations, most of which probably lie below the threshold of consciousness[84]." (p.424)

As I have just explained, similar correlations are learnt between motor commands and auditory feedback which can be thought of as inverse and forward models. Of the former, the auditory IM and the kinematic IM are most important (specifying the actions needed to create particular auditory feedback and particular changes in the position/state of the articulators, respectively). The multiple forward sensory models can be conceptually simplified into three: a forward auditory model (the expected consequences of motor actions on what the infant hears of himself), a forward kinematic model (the expected consequences of motor actions on the position/state of his articulators), and a forward sensory model (the same for all other proprioceptive feedback).

Note, again, that the inverse and forward auditory models will both be calibrated on what the infant hears of his own production, which may or may not be a good match to what others hear of this. (This may, therefore, be more or less useful to the infant in matching patterns of sounds that he produces to patterns he perceives others to make, depending on its degree of distortion[85].)

In due course, then, this auditory feedback will enable him to control his production on the basis of the auditory feedback he perceives from his motor activity.

However, this does not mean that a speaker needs to monitor his vocal output consciously. Instead, he may learn an automatism which uses a comparison between the output of the forward auditory model and the actual auditory feedback to alert him only when a significant error is made. ('Significance' being a learned criterion.) The other forward models will be used in the same way.

---

[84] The fact that they lie below the threshold of consciousness in adults is no reason to believe that this is true for young children, however. See Gattegno (1973:6) on somatic consciousness, and Gattegno (1973:63-74) on the earliest stages of talking.

[85] The same issue obtains in vision. What I see of my own actions – arm waves, beckoning gestures, body attitudes – is perceived through the same modality as I perceive others, but always from a different point of view. This may create few problems if I am trying to match an arm wave, but what about a curtsy (Heyes 1993)? For speech there are reasons to believe that this issue may be more significant, explored in section 13.4.

The auditory forward model gains particular importance when the goals of articulation are conceptually re-represented from being motor to being auditory targets. Control 'passes from the throat to the ear' (Gattegno 1973): a 'structural displacement' of the resources used for task performance (Ivry 1996:286). Ejiri (1998) presents evidence of this occurring for shaking a rattle, at the onset of canonical babbling, but notice that the use of general sound imagery this way may be completely independent from the ability to use speech imagery for it. I return to this in section 13.4.1.

10.4.3 Babbling

Is babbling drift evidence of vocal mimicry by infants prior to first words?

Babbling drift is the name given to a supposed change in the phonetic character of utterances made by infants during babbling: a move towards the model provided by the ambient environment. An alternative view is that while changes are certainly seen in infants' output over the course of babbling, that these are universal tendencies.

Kuhl (2000; Doupe and Kuhl 1999) uses the evidence for babbling drift put forward by de Boysson-Bardies and her colleagues in the 1980's and 1990's (summarised in de Boysson-Bardies (1999)) to support the idea that infants have started imitating adult speech sounds before they start learning words. Kuhl and Meltzoff (1996) is taken as evidence that they have this capacity from at least 6 months.

However, the reality of babbling drift is not accepted by all authorities (e.g Locke 1993:166; Oller et al. 1997:423). Oller and Eilers (1998) question the methodologies used in studies which find support for it. Engstrand et al. (2003) carefully review previous studies as well as presenting new evidence, and do not find support for a strong or universally applicable version of the hypothesis.

On the other hand, Mattock et al. (2005) recently found some differences in F1 and F2 production between French- and English- learning infants.

At present, then, the extent to which pre-L1 imitation plays a significant role in the development of an auditory IM that will be used to produce words is unclear.

However, I will propose that there is a difference between the mimicry of sounds and the re-enactment of speech sounds. Babbling drift would probably fall into the former category, in which case while it might help with the re-creation of the first adopted words as whole shapes, it will not contribute directly to the ultimately preferred strategy of word adoption through speech sounds. (Some indirect benefit would be gained to the extent that a mother may reformulate more of an infant's output if it is closer to L1, though.) Locke (2001:295, 304-305) makes a point complementary to this.

An embodiment perspective on babbling

In a series of publications, MacNeilage and Davis have developed a performance-related (or motor-related) approach to the understanding of babbling and first word production that is consistent with the proposals I will make[86]. They invoke the motor aspects of speech production to explain data that others explain through perception (the child copying what he hears), 'phonological processes' (MacNeilage and Davis 1990:61), or genetic determination via Universal Grammar (MacNeilage et al. 2000).

With respect to copying (which would presumably include babbling drift), Davis et al. (2002:101) argue that:

> "Perceptual factors do not appear to play a very important role in determining the overall form of the output patterns, [and there is] no obvious basis in perceptual facilitation for any of the preferences for sounds and sound patterns observed in the babbling and first word stages" (p.101)

Instead they propose, "a phonetic interpretation of first word patterns [whose] predominant causality is motor." With respect to 'cognitive' mechanisms they contend that,

> "Few would dispute that a satisfactory conception of modern adult speech production must include a hierarchy of organizational components including both a motor level, and premotor levels that would be more readily characterized as conceptual. In contrast, our conception of the earliest stages of speech acquisition is, with one exception, basically confined to the motor level. From our standpoint the evidence regarding what actually occurs during babbling and first word production does not necessitate positing a dominant level of organization at these stages that is basically conceptual." (p.102)

[86] Faber and Best (1994) is a set of proposals with a similar orientation, though dealing with later stages of development.

MacNeilage and Davis support this with data from the serial organisation of speech in babbling and early words. From this, they find,

> "... very little evidence that infants have developed any segmental independence by the end of the single-word stage. Instead, what we found can be summarized by the term 'Frame Dominance'. Most of the variance in vocal output from 7 to 18 months of age can be attributed to mandibular oscillation alone, with very little evidence that any of the other articulators – lips, tongue, soft palate – are moved independently during vocal episodes." (2000:286)

This would seem to conflict with transcriptions that, at first sight, show segments within a syllable combining relatively freely. However, the inaction of articulators other than the mandible during a frame can result,

> "... for example, in the illusory impression that there are two gestures of the tongue apex itself in [tata] just as there would be in adults while in fact the tongue was preset in the apical position and simply maintained in that throughout the utterance. The impression that the [t] was produced twice may thus be due to the fact that neither the position of the tongue nor the amplitude of mandibular oscillation changes during an utterance." (1990:60)

In fact they propose that both reduplicated and variegated babbling can be understood best from a performance perspective. Thus,

> "... for normally developing infants, the main source of variance in reduplicated babbling is uniform amplitude of successive cycles of oscillation of the mandible. The main source of variation in variegated babbling is non-uniform amplitude in successive cycles of oscillation of the mandible, affecting either the closing (consonantal) phase or the opening (vocalic) phase, or both. This finding suggests that even in many cases in which transcription suggests that the tongue occupies a non-resting position in the front-back dimension during a babbled utterance, even a multisyllabic one, its position may not change during the utterance. Thus, an utterance such as [dædædæ] might only involve vertical tongue movement produced by the mandible. The forward movement required may occur before the acoustic onset of the utterance. One way to summarize the results obtained in this study is with the phrase 'Frame Dominance.'" (Davis and MacNeilage 1995:1208)

Davis and MacNeilage (1995:1200) explain the associated slogan, 'Frames, then Content',

> "... as a metaphor to describe spatio-temporal and biomechanical characteristics of babbling and changes during early speech. The term Frame applies to the regularity of mandibular oscillation resulting in listener perception of syllable-like and therefore speech-like output. It is claimed that close and open phases of the cycle often may have no associated neuromuscular activity other than movement of the mandible and consequently may have no subsyllabic organization or Content. In this view, the syllabic Frame thus constitutes the earliest temporal envelope within which segment-

> specific Content elements develop as the child gains increasing independence of control over speech articulators in speech movement sequences."

I have already made use of this idea, in section 3.5.3, to describe the relationship between speech breathing and the activity of the upper articulators.

10.4.4 Vocal motor schemes (VMS)

McCune and Vihman (2001:671) investigated the phonetics of babble and early speech in order to better understand the development of the single-word lexicon. One focus was on, "well practiced and longitudinally stable vocal productions as distinct from more sporadic occurrences." They took the former to be indirect evidence of a capacity for consistent phonetic patterning, which they term Vocal Motor Schemes:

> "[McCune and Vihman (1987)] proposed that early articulatory skill could be characterized as the learning of 'vocal motor schemes' (VMS), generalized action patterns that yield consistent phonetic forms. In a similar vein, Thelen, Corbetta, and Spencer (1996) demonstrated that children's successive reaches toward an object at 6 months of age showed random variation in trajectory, whereas by 8 months each child showed a consistent trajectory in repeated reaches, presumably based on a consistent and repeatedly utilized 'reaching motor scheme.' The vocal motor scheme concept is applicable to any consistently occurring phonetic pattern developed, in theory, by repeated and regularized child action." (p.673)

Strikingly, McCune and Vihman found that, "children based virtually all stable words on their own specific VMS consonants (92% vs. 8% for other consonants)," with six out of their nine subjects basing all stable words on VMS consonants or [h].

Vihman uses the VMS concept to support her model of early word learning, and I shall make use of it in mine.

10.4.5 Protowords

In the diagram that will portray my model of speech production (see figure 10-2) I have divided the child's production into babble, protowords (forming his 'L0') and words (forms that are associated with the conventional forms found in the first language that he is learning).

I am using the term 'protoword' in the same way as Vihman (1996:130), to refer to, "relatively stable child forms with relatively consistent use which lack any clear connection with the form + meaning unit of a conventional adult model." Vihman points out that this is different from how Menn (1976) used the term. She also discusses

some of the alternative terminology that appears in the literature, including 'vocables', 'phonetically consistent forms', 'quasi-words' and 'acts of meaning'.

Halliday (1975) argues that protowords give the infant expressive vocal objects which are associated with a wide range of functions. Blake and Fink (1987:230) document an apparently extensive use of protowords by their infant subjects.

10.5 Summary

In this chapter I have discussed some topics that will play a part in the proposals that follow. First, I described Mason's distinction between ordinary-noticing and marking; the former creating the possibility of recognition, the latter the possibility of initiating a 're-mark' to others and the possibility of an evocation to oneself.

I then described two way of attending to sensory input:

- 'Awareness of sensation' (AS): seeing/hearing with nothing external to the scene brought to bear and no meaning extracted or inferred; I asserted that this perceptual mode is used for the purpose of mimicry or re-creation of surface level characteristics, where aspects of sensory patterning are being matched. I am assuming that processing in the dorsal/'how' stream (supported by ventral stream activity in a way that will be described later) is among the equivalent ways of describing this.
- 'Meaningful perception' (MP): the way we normally relate to a scene, at the same time identifying meaningful elements and seeking overall meaning. This mode would be used when perceptuomotor transformation is intended[87], and is assumed to be the result of processing taking place in the ventral/'what' stream.

I defined how I will use 'speech sound', 'soneme' and 'movement' for units of production and perception at the scale at which a child meets speech. I also distinguished the input to the child from his intake, and his output from the outcome.

[87] In case this is unclear, 'perceptuomotor transformation' would mean producing sounds that 'mean' the same as the target sounds. In some situations 'mean' will equate to belonging to the same phonetic category. This is then opposed to 'sensorimotor transformation' which, in a speech context, means producing sounds that resemble or can be taken for target sounds.

Finally, I discussed five aspects of early production: the sources of information a child has about his performance, how he creates inverse and forward models, the role and nature of babbling, vocal motor schemes and protowords.

11 Early speech perception

In this chapter I will treat some issues in speech perception that can be dealt with independently from production and prior to considering it. One purpose is to propose 'acquaintance' as an overarching characterisation of early word recognition.

Then I will look at similarity and categorisation behaviour in the context of infant speech development. Most accounts of speech sound development assume that the determination of equivalence between an adult and a child's production of a speech sound is made by a child as a result of a judgment of similarity made by him, but the determination of equivalence need not be the result of such a process. I will point out that functional equivalence provides an alternative.

11.1 Early word recognition: acquaintance

An infant must develop many skills in order to understand words. He must learn, for example, to hear speech sounds spoken in different voices as equivalent, to segment words in the stream of speech, and to pair form and meaning. Some of this work has been characterised as 'statistical learning' (for a recent review, see Saffran 2003), which shares many characteristics with so-called 'implicit learning' (Perruchet and Pacton 2005).

However, it seems to me that prior to production, the overall context in which these skills develop is one in which there is no great benefit to the infant in learning to recognise words. He will gain a significant benefit only once he can use words to express himself. At this point, recognising words becomes a route to acquiring them for this purpose and hence for reciprocal communication. Prior to this, words are said <u>at</u> him, but not to any advantage <u>for</u> him[88].

This having been said, his mother knows that recognising words is going to be of benefit to the child in the future and may well be of benefit to her in the present. So she rewards him in play activities that use word recognition ("Where's the ball?" "Well

[88] I imagine that a similar perspective underlies this observation by Davis and MacNeilage (2000:230): "The classical approach to speech perception in human infants does not acknowledge that, in the most general sense, speech perception is in the service of speech production. The goal of the normal hearer is universally to become a speaker."

done!"). From time to time, then, some motivation to learn a word or a reward for demonstrating past learning will arise.

In general, though, prior to being speakers infants have no motive to be word learners. One sign of this may be that they do not pay particular attention to the speech around them. Huttenlocher (1974:339) describes how difficult it can be to get a young child to pay attention to speech, even speech addressed directly to it.

An absence of motivation does not mean, though, that an infant will not learn to recognise words. But it will not be through a route that demands a motivated participant.

Gattegno (1987:55ff.) describes familiar ways of knowing such as perception, action, analysis and synthesis. He also describes less well studied ones, including contemplation, intuition and, in particular, acquaintance.

The normal meaning of 'an acquaintance' reveals some characteristics of this way of knowing. Through no more than contact with a person and being prepared to yield to what they present to us, we end up with knowledge about him or her. No desire to get to know them is required. If we have extended contact with someone, as with a colleague at work for example, then we might end up with very considerable knowledge about him or her this way.

As another example of acquaintance, I might be the passenger on a long car journey during which the driver plays country and western music throughout. Despite my lack of interest in the genre, I will end up discovering things about it. There will certainly be times when nothing else is occupying me and I will listen to the music, noticing recurring patterns, drawing distinctions, finding questions about it occurring to me, and so on. In fact, if it is a long journey I may end up knowing quite a lot about country and western by the end. But apart from curiosity aroused during the experience, I have no motivation to achieve this.

Infants are on a very long 'journey', during much of which they are immersed in speech[89]. The conjunction of this extended contact and the personal attributes that

[89] Locke (2001:294) quotes statistics from Hart and Risley (1999) indicating that Midwestern American parents of 11-month-olds produce about 300 child-directed utterances per hour. Another 500 utterances

children bring to learning is sufficient to lead to word recognition beginning.

Describing early word learning as happening 'by acquaintance' is not, then, to deny any of the skilful things an infant has to do for this to be possible. Segmentation of the speech stream, normalisation of voices and so on must take place.

It is possible that the infant will mark some of what he notices. Curiosity will provoke questions which he may answer by quite active exploration and theorising about how sounds are being used by his mother. These activities may benefit from him making mental notes in some form. However, the overall context will be one where he is just in contact with speech rather than actively trying to acquire understanding of it. So most of his learning will happen by ordinary-noticing, and contrary to what is sometimes imagined, his 'recognition lexicon' is not being acquired in a way that will assist production.

11.2 Infant categorisation (equivalence classification) of sounds

Discrimination, categorisation and judgment of similarity are as fundamental to speech learning as they are to cognitive development in general. Kuhl and her colleagues have presented a body of evidence dealing with these matters with respect to infant speech perception which I will briefly summarise. Then I will describe an analogy with vision that has helped me to understand the issues involved and which I will refer back to at later points. Finally, I will note the relationship between similarity and equivalence that my main proposal in this Part rests upon.

Kuhl's experiments on speech categorising by infants

As Kuhl (1987:337) points out, categorisation requires more than can be demonstrated by evidence of 'categorical perception'. For within-category contrasts that are not responded to differentially, it is possible for categorical perception to be based not on a recognition of category equivalence but on a subject being incapable of discriminating the stimuli presented. Categorisation, on the other hand,

---

per hour, by the parents and other family members, also fall on infant ears, for an average of one utterance every 4.5 seconds.

> "... requires that discriminably different stimuli be perceived as equivalent. In speech, this means that infants must recognize the similarity among phonetically equivalent events that are represented by very diverse acoustic cues. We want to know whether infants recognize phonetic equivalence in the sounds produced by different talkers, and if they recognize a phonetic unit's identity when it occurs in different contexts. This necessitates two things: (a) that infants recognize phonetic equivalence when the values of the critical dimensions underlying the perception of the phonetic unit are altered; and (b) when additional dimensions that are acoustically prominent, but irrelevant to the task of phonetic categorization, are introduced.
>
> When we ask whether infants can categorize we want to know if they can sort a variety of instances into "Type A" and "Type B" events, even though the various A's (or B's) are clearly differentiable. To perform such a task requires that infants recognize the similarity among discriminably different instances representing a category, while at the same time recognizing the essential difference that separates the two categories. Thus, categorization requires a process in which the perceiver recognizes equivalence."

Kuhl and her colleagues have shown that infants can correctly categorise stimuli designed to contrast vowels, consonants and distinctive features across distracting dimensions of talker identity, context and pitch movement. Thus the vowels spoken by different people can be correctly sorted by infants (tested with spectrally similar and spectrally dissimilar vowel categories); as can syllables that share initial consonants or just a phonetic feature (e.g. the set of stop plosives or nasals); and vowels spoken with different pitch contours and by male/female speakers. (See Kuhl (1987:337-355) for a summary; also Kuhl (1991) for a report on the most demanding conditions applied.)

However, the basis for these judgments is unresolved. Do the results imply a segmental level of analysis and representation? Are infants abstracting phonetic concepts like distinctive features, or making their judgments on the basis of patterning in the acoustic waveform that is just indexical of these phonetic concepts?

Kuhl (1987:351-355) uses the neutral terminology of 'portions' and 'parts' of the acoustic signal to discuss these issues. The data are consistent with the claim that infants have access to a segment-level analysis, but they do not demand this explanation. Similarly they demonstrate that infants must be able to break syllables down into parts that might be phonetic segments or phonetic features, but might equally well be "something else".

Visual analogy

These are quite subtle issues, and it has helped me to think in terms of a visual analogy. Consider an observer who has never constructed anything at all and who is now presented with a variety of physical barriers for inspection. For him, the most obvious

characteristics of a wall might be its overall dimensions. Its opaqueness and relatively smooth surface might then distinguish it from fences of various types. (Remember that our observer has never constructed a barrier of any kind; he will not be perceiving anything other than the visible characteristics of the structures and will not be aware of the functional connotations to distinctive terms such as 'wall' and 'fence' even if the labels themselves are known.)

In comparing walls of similar dimensions, the observer can compare the patterning, colour and texture of the repeating rectangular elements. Similarly, he can notice the colour, size and shape of the strips that divide these elements.

As people who do understand wall construction we know that the rectangular elements are bricks and stones whose sizes, colours and textures can reveal information about their hardness, durability and porosity, and whose patterning (in a stretcher bond, for example) reveals something about the strength of the overall construction. We also know that the dividing strips are mortar, whose colour, thickness and pointing reveals further characteristics of the wall.

The important thing is that the purely visual inspection of the barrier reveals detail which co-occurs with what a functional parsing reveals but is not the same as it. If asked to distinguish one wall from another, then, our observer can do so using visual cues without ever representing them in terms of the literal building blocks which underlie them. Similarly he can put walls into categories based on the same cues, in a way that would resemble the categorisation of a builder doing the same task, but one who would be basing his judgment on the construction materials and type.

This is the reason why it has proved impossible to determine whether infants categorise sounds based on the recognition of linguistic elements like phonemes and distinctive features, or whether they are just noticing the similarities and differences in the acoustic waveform that these elements generate. In the latter case they would be examining "portions" and "parts" of the waveform[90].

---

[90] In Studdert-Kennedy's (1986) review of this topic, he says:
"Consider, next, the evidence that infants can form 'phonetic' categories across a variety of acoustic contexts. Here again the data are overinterpreted. Since every phonetic contrast is marked by an acoustic contrast ... phonetic and auditory perception cannot be dissociated in the infant (though they can be in the adult...)."

The analogy with speech is improved if we imagine our construction-naive observer looking at barriers from a train window. He will see them for short periods, at varying speeds, under different lighting, with the barriers closer or further away, and so on.[91] Yet as proposed in the last section, if these scenes were incessantly presented to him, then he would surely learn quite a lot about barriers (including, perhaps, some fine detail) whether or not he had any particular interest in doing so. I.e., he would surely start to discriminate and recognise types of barriers just as a result of his acquaintance with them.)

Similarity and categorisation

In general research on categorisation, one important issue is whether categories are created from items which are inherently similar, or whether the notion of similarity between category members emerges out of categories formed on an independent basis. The present balance of opinion seems to be in favour of the latter view (e.g. Sloman and Rips 1998) but with an acknowledgment that categorisation is too large and complex a subject for it to be likely that a simple answer be possible.

Mompeán-González (2004) reviews the issues and applies current ideas to speech.

The working assumption that I will use is that perceptual similarity comes to be a convenient classifying heuristic for mature speakers in language environments they know well, but that the results of a judgment made on this basis are not conclusive. At root, linguistic/phonetic categorisation is functional: if two sounds do the same linguistic work then we assign them to the same category.

For example, the qualities of the vowel in *cat* may be very different in two dialects of English (they may be clearly dissimilar when presented in isolation), but once we understand how each is used we happily accept both as realisations of /æ/. Similarly I may adjust my category boundaries if I am listening to a learner of English, whether an infant or a non-native speaker. As Mompeán González (2004:433) points out, "Similarity could be a by-product of conceptual coherence rather than its determinant." With respect to speech sounds, 'similarity' may emerge from what listeners judge to be equivalent as much as being based on inherent properties of the acoustic signal.

---

[91] I would like to thank Gautam Vallabha for prompting me to pursue this.

11.3 Summary

In this chapter I argued that there is no reason to impute to an infant the desire to learn words. That motivation will come later, when he discovers what he himself can do with them, but in the beginning there is no real benefit to him to learn to understand what is being said to him. Nevertheless, his predicament means that he learns words anyway: by 'acquaintance'. For this, only ordinary-noticing is required, which means that the process builds a lexicon but not one whose entries are structured to help with production.

I also pointed out that similarity is not necessarily the basis on which speech sound categories are formed. Equivalence can be defined functionally. A sense of similarity would then be derived from category membership, and play the useful role of a guide to this for experienced listeners.

There are many other issues which a full treatment of early speech perception and word recognition would cover, including how categories for speech sounds are formed, more about talker variability, the amount of detail captured by infants in their representations, and how they make use of what they capture. Recent reviews include Houston (2005), Walley (2005), Werker and Yeung (2005) and Gerken (2002).

At the heart of some of the discussion in the literature, although not always obviously so, is an assumption that in order to be able to reproduce words using speech sounds (rather than holistically), children must recover units of speech production as part of their process of word recognition. This suggests that at least one facet of a word's perceptual representation must be structured into soneme units at a fairly early stage. Opposing this view is evidence that young children fail to make the detailed distinctions that such an organisation of the lexicon would seem to demand, even after being speakers for several years.

My proposals on a non-imitative mechanism for speech sound development will resolve the apparent contradiction and I will return to this subject in chapter 15.

12 Learning to imitate

> "All this means that a child is certainly not born with a fully fledged power of imitation. Children have to learn to imitate. It is a slow and laborious process. A child learns to imitate in the same way as he learns many other things in his progress towards language: by a combination of what he himself does and the responses of others towards what he does."
>
> M.M. Lewis, *How Children Learn to Speak* (1957:49)

12.1 Introduction

The next three chapters are concerned with how children come to be able to pronounce the words that they hear. In due course this becomes straightforward and we say that a child can learn a word's pronunciation by imitation using speech sounds rather than by re-creation of the whole-word shape. In this chapter I describe previous proposals for how children put that ability in place and in chapter 13 some potential problems which might tell against these accounts. In chapter 14 I will propose a mechanism that is not vulnerable to these particular criticisms.

Children come to be able to transpose or transduce speech sounds so it is reasonable to ask how and when they learn to do this. The context in which the ability mainly develops is one where they have a superordinate goal of being able to say words, and this must then be the source of the continuing motivation for them to improve their pronunciation. However, some learning prior to word adoption might provide an entry point or bootstrap into the process, even if motivated for reasons other than word production. Locke has consistently drawn attention to this point (e.g. 1996, 2001). Similarly, some learning apart from word production itself is presumably important: for example, various authors have documented children appearing to work independently on their articulation of particular speech sounds (e.g. Weir 1962; Menn 1983:38).

There is a danger in the (convenient) practice of talking about children "learning speech sounds" since one sense of the term "learning" would imply that children are actively copying what they perceive in the environment. To conclude this would be premature, since there are several alternatives which would first have to be rejected. One is that they discover in the environment things that they can already do being done by others, enabling them to match this behaviour. The more neutral sense of the phrase "learning speech sounds" that I will have in mind when I use it, therefore, is simply for it to mean

how the sound qualities of children's pronunciation becomes increasingly acceptable to those around them.

The question of how children learn speech sounds is an issue of practical as well as theoretical interest. The pedagogies by which millions of children and adults are taught the pronunciation of languages and by which many receive therapeutic treatment for disorders of speech are very often based upon and/or justified by what is presumed to be 'natural' in the learning of young children. The assumption almost universally made is that of a copying mechanism. In fact this is so uncontroversial that Skoyles (1998) could point out that neither 'imitation' nor any of its synonyms merit an entry in either of two contemporary multi-volume encyclopaedias of language and linguistics, while 'bats' and 'dance notation' do.

Some comments from authors who have asked how it happens indicate that the question is not yet resolved:

1 Studdert-Kennedy (1986:60) described imitation as the "central problem of early speech development", continuing:

> "We have been easily diverted because it seems natural that, if an adult speaks a word or grasps the air with a hand, a young child can repeat the word or imitate the hand movements. But how, in fact, does the child do this?"

2 After pointing out that vowels are not easy for children to acquire, Davis and MacNeilage (1990:16) opined that until the neglect of correct vowel acquisition as a research topic is rectified, "it is doubtful that any major issue in child phonology can be satisfactorily addressed."

3 Lewis (1951:99) observed that, "for a very long time the forms used by the child in imitation of adult language consist of his own familiar sounds spoken as approximations to those that he hears," and then considered, "[one] remain[ing] point of great difficulty: how does the child pass from his familiar sounds to unfamiliar ones?"

> "This is a difficulty which has impressed one psychologist after another ... we must be content to leave the question open for further evidence." (1951:100)

4 Over 50 years later, Kent (2004:12) observed that,

"Because infants often spontaneously imitate speech sounds produced by adults, it is sometimes supposed that vocal imitation is a route to the learning of speech.

Curiously, vocal imitation by infants has not been systematically or widely studied."

12.2 Learning words by imitation vs. learning to imitate sounds

A first issue to address when considering how words can be imitated is whether imitation *per se* is a form of learning. Young (2000) gives an example which makes it clear that for one sense of 'imitation' the two are completely distinct:

"When I was a child, I learnt to walk a tightrope. So if I installed a rope between this rooftop and that one, and told you, "Now off we go. Just do like me." would you try? Of course not, because you know as well as I do that you can't imitate me in this. You have to develop the sensitivity to your centre of gravity and all the other technical skills and the muscular power in your feet and in your abdomen which will allow you to do it.

This is an extreme example, but a little thought shows that in all circumstances, without exception, it is only possible to imitate what one can already do. If I don't already possess the gesture I can't imitate someone else doing it. If imitation were part of the learning process we could all be champions in any discipline we wanted to. Just watch and do[92].

Imitation exists of course, but when someone is imitating they are not faced with the unknown. They are using skills which they already possess."[93]

---

[92] Young's point is not contradicted by recent evidence on motor learning through observation, but it indicates that 'watching' fellow learners can help.
Mattar and Gribble (2005:153) asked, "can information specifying 'how' to make movements at the level of motor execution (e.g., novel patterns of muscle forces) be conveyed through observation?"
They report that, "...by observing another individual learning to move accurately in a novel mechanical environment, observers move more accurately themselves. Subjects can acquire neural representations of novel force environments on the basis of visual information." (p.157)
A further finding of interest was the possibility that, "motor learning by observing may occur unbeknownst to the subject." (p.158)
Note that their subjects observed fellow learners. Mattar and Gribble presume that, "observing the actions of skilled individuals (after learning has already occurred) would not lead to motor learning in the observer." And the observational learning did not mean that subjects fully learnt how to move in the new environment; additional experience of performance was required.

[93] I now realise that a BBC television programme called "The Generation Game" which was very popular when I was a child richly illustrated these and other themes in imitation and memory. Contestants were supposed to "watch and do" in order to imitate an expert in some activity (throwing a pot is an example which comes to mind). They quickly discovered that this was not possible: the skill element of the task could not be acquired this way although some idea about the sequencing of the subskills needed could. Instead contestants had two minutes to teach themselves how to accomplish the task and impressively compressed learning was sometimes seen.
At the end of the programme the winners kept whatever prizes they could recall from a collection of items that had passed before them on a conveyor belt. Unfortunately they were not then asked whether they had achieved this through subvocal name rehearsal or by the recall of visual images...
Perhaps the programme was devised by a cognitive psychologist. Its name appeared to me at the time to refer to the fact that contestants were drawn from different generations of families, but it was more cleverly punning than that.

So if there is a subskill missing when we are attempting to learn something 'by imitation' then the gap must be bridged by a learning process that is separate from the imitation attempt (which has only provoked the awareness that the learning is needed). Mastery of the subskill will then give us the ability to imitate. The "by/to" distinction is described more fully by Parton (1976:14) (with the author's original emphasis in italics, mine in bold):

> "Learning **to** imitate is a topic which must be distinguished from learning **by** imitation. In the vast majority of imitation studies the subject observes behaviours composed of molecular responses that the subject has previously acquired. Consider, for example, a preschool child who observes a model put on a hat, walk across the room, and pick up a book. Much earlier in life the child acquired the motor skill of walking and the visually guided reaching and grasping required for putting on hats and picking up objects such as books. Thus, the typical *learning-by-imitation* study assumes that the child **has previously learned the molecular responses** exhibited by the model ... [and] what the child learns from the model includes ... the sequence of the performance ..."
>
> "Learning **to** imitate, on the other hand, focuses on the infant acquiring the capacity, upon hearing or seeing a molecular behaviour, to innervate the particular motor response which is similar to the observed behaviour. Any imitative response reflects perceptual-motor isomorphism; the perception of the behaviour has become associated with the motor efferents which produce a response judged to be similar to the modeled response. Even when the infant is capable of exhibiting any of a great many responses, the development of imitation requires that the infant become able to exhibit precisely the response that will match, and the development of this capacity comprises learning to imitate."

Alissandrakis et al. (2002:485) similarly describe learning to imitate as, "the study for specifying the necessary mechanisms by which observed and executed actions are matched, so that the agent can use imitation to learn how to perform useful behaviour." Nehaniv and Dautenhahn (2001:32) call the same process, "trying to imitate", where the task is to achieve a correspondence between the effects generated by another's actions and those effects generated by one's own. In the subsequent phase of learning by imitation, Byrne (2003) calls the observation of the sequence of molecular actions making up a performance 'string parsing', a terminology I will adopt.

As part of her Associative Sequence Learning model, Heyes has introduced some simple terminology which helps to keep the "by/to" distinction clear as well as illuminating the mechanisms by which learning to imitate might occur. The following brief description draws on Heyes and Ray (2000), Heyes (2005) and Hoppitt and Laland (2002).

Heyes takes a typical case of imitation to involve a number of action-units with sensory and motor components, which combine to form an action sequence (see figure 12-1). She defines two axes as shown. If we imagine, for the sake of an example, that a child would imitate *gruffalo* by breaking it into three syllables, then these would map onto 'Sensory$_1$' to 'Sensory$_3$' and the 'horizontal' links would define the sequence in which these have been heard and will need to be performed.

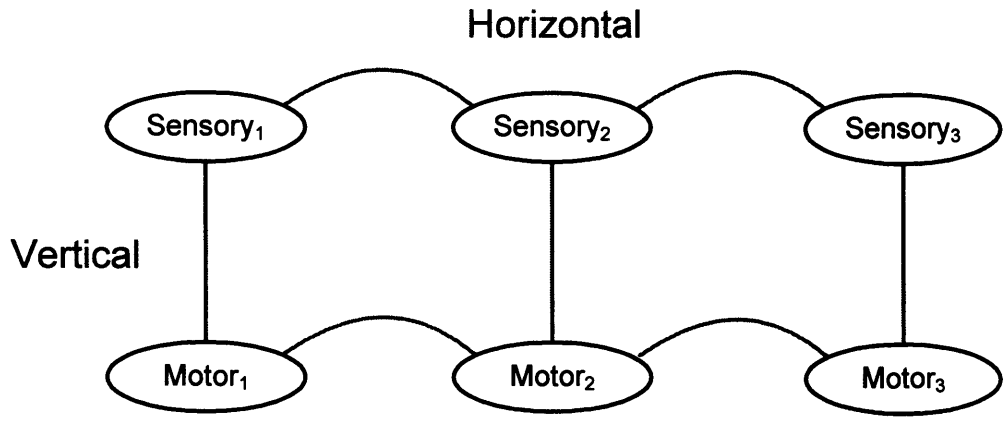

Figure 12-1 (adapted from Heyes 2001). Action units are combined to form an action sequence. Parsing the input creates a horizontal specification, but the sensory and motor components of action units must have been matched before reproduction is possible.

The correspondence problem, learning to imitate, is represented by the 'vertical' links which match sensory and motor boxes. These need to be in place before a successful imitation is possible. Some theories of imitation, for example Meltzoff and Moore's Active Intermodal Mapping, propose that matching in action units can be innate, without a process of learning being required. However, in the absence of such a mechanism, the means by which these links can be developed will vary depending upon the sensory capabilities of the observer (B) relative to the model's (A's) actions and his own.

The degree to which the sensory experience of observing another individual performing an action overlaps with the sensory experience of observing oneself performing that action is called perceptual opacity. For example, it is widely believed that vocal actions are perceptually transparent, since hearing oneself produce a sound and hearing another individual produce the 'same' sound are believed to be identical experiences. (This may be true for some birds (Zentall 2004:16) but in the next section I question it for humans with respect to speech sounds.) Facial gestures, on the other hand, are perceptually

opaque, since one cannot see one's own face; one can only feel what one is doing with it. Some actions will be intermediate, such as those involving one's hands or legs. Making a ring between thumb and forefinger is perceptually highly transparent, while kicking a football is more opaque because of the different angles of viewing involved.

Forming a vertical association for perceptually transparent actions appears straightforward: B needs to generate variant actions, compare the sensory feedback with a sensory representation of A's actions, and then select/adjust his actions to minimise the discrepancy he identifies. I will shortly describe this process as 're-enactment' when applied to speech sounds. Note that B makes a judgment of similarity to establish correspondence.

Perceptually opaque actions do not yield comparable sensory experiences (except with support from a device such as a mirror or a tape recorder). However, feedback to B about his action can nevertheless lead to a vertical association being formed. For example, imitation of B by A (note the direction of imitation here) is one form of behavioural synchrony that provokes concurrent activation of sensory and motor representations of the same movement. After doing something, B is shown by A what he just did. In essence, A acts as a 'mirror' for B, and this metaphor can extend to the mirroring of internal states as well as surface behaviour, as I will describe in chapter 14.

Since we are considering how speech sounds are learnt, our concern is with learning to imitate:

- solving the correspondence problem;
- in Parton's terminology, the child acquiring the perceptuo-motor isomorphism linking what he hears others produce to the molecular motor behaviours underlying his actions;
- in Heyes' terminology, forming 'vertical associations'.

This is what will allow a child to acquire words by imitation (after parsing them for their constituent speech sounds), in the efficient way that is observed during the period of rapid vocabulary acquisition and thereafter.

Note that I am taking this last point to be uncontroversial: that once we have solved the correspondence problem for speech sounds, the process of learning to produce words by imitation is not one of recreating a sound image holistically. Instead we parse the sound

image to retrieve a string of elements for which we have equivalent motor routines. The acoustic similarity of the resulting sequence of speech sounds that we produce is not the determining factor for success. For speech to be comprehensible, the sequence of speech sounds needs to be taken as equivalent to their own by listeners. In practice, apparent similarity will be a normal by-product of the process where the dialects of model and observer match.

As a summary of some of the points just made:

- The next three chapters are concerned with how children come to be able to imitate speech sounds and hence to learn words by imitation.
- Imitation is not the same as learning, although things can be learnt from the results of an imitative attempt and the activity that follows it.
- If speech is perceptually transparent, a child may develop the qualities of speech sounds by a matching-to-target process (a cycle of imitative attempts guided by self-made judgments of similarity).
- But if speech is instead perceptually opaque, an alternative could be for him to be informed of the adequacy of his attempts at speech sound reproduction and word imitation by others' responses, creating a 'mirror' for his performance.

12.3 Previous accounts

There are a number of substantially different accounts of how children learn to imitate speech sounds. I will start by describing what seems to be the most natural of these, a copying account which corresponds to lay beliefs about the process. I will then discuss alternatives to this, organised around the answers they give to two questions:

1. What is the basis of the child's production? Is it
   (a) developed from the start to copy what he finds in others' speech; or
   (b) developed independently and then found to be present in others' speech? (At which point it can be deployed with the new purpose of imitation.)

2. What is being matched?
   (a) The acoustic signal; or
   (b) articulatory gestures?

Not all accounts are amenable to being classified in this way. I will finish this section by looking at an approach to solving the correspondence problem that draw on ideas from neuroscience.

Later, my own proposal (and that of Yoshikawa et al. (2003)) will differ in another way from the leading accounts described. These all assume that the child makes judgments[94] of similarity between his own and his mother's acoustic output to establish their equivalence. For this reason I will call these 'similarity-based equivalence' (SBE) accounts. In contrast, I will describe a mechanism in which it is the mother who judges equivalence, with this judgment recognised and acted upon by the child. I will call this a 'mirrored equivalence' (ME) account. (Figure 14-2 in section 14.3 may help with the location of some accounts in this scheme.)

12.3.1 The child copies an acoustic model (mainstream accounts)

Assuming speech sounds are perceptually transparent, then solving the correspondence problem for them is not, in principle, difficult. The following steps in a repeated 'matching-to-target' process are required (the first two occurring in either order):

1. The young child perceives and holds in memory (captures) target values for a speech sound from adult speech;
2. He produces an attempt at a match, capturing the acoustic result;
3. He judges the similarity and differences between his attempt and his target(s);
4. Any error indicates how he can produce a better attempt next time.

One constraint on progress is that articulation is a motor skill which has to be executed fast and precisely. A child is likely to make many motor errors, which might partly explain the long time it takes to master speaking. Another factor in this might be the cognitive 'overload' induced by the many tasks that must be executed in parallel as part of speech. But the specifically imitative aspect of the process seems to be reasonably straightforward. These steps can all be imagined to be within a young child's capacity; even step 3, for which the problem of the mismatch in vocal tract sizes has often been pointed out. This means that a child can never produce a copy of adult speech with the

[94] I use 'judgment' in a broad sense here and elsewhere. It does not seem important to my argument whether this is a consciously aware activity or, perhaps, some implicit matching process.

same acoustic structure as the original. However, since Kuhl and her colleagues (e.g. Kuhl 1991) have shown that infants of just 6 months can categorise 'speech sounds' across male, female and child speakers (i.e. that they can 'normalise' the signal or, alternatively, abstract those features that characterise speech sounds) it seems reasonable to imagine that young children can deploy this skill in a comparison involving their own speech. Kuhl (1987:341) explains why talker normalisation is a prerequisite for imitation, yielding a "pattern" from vowels spoken by different talkers which is the target that the infant mimics (rather than the target being the absolute formant frequencies of the adult).

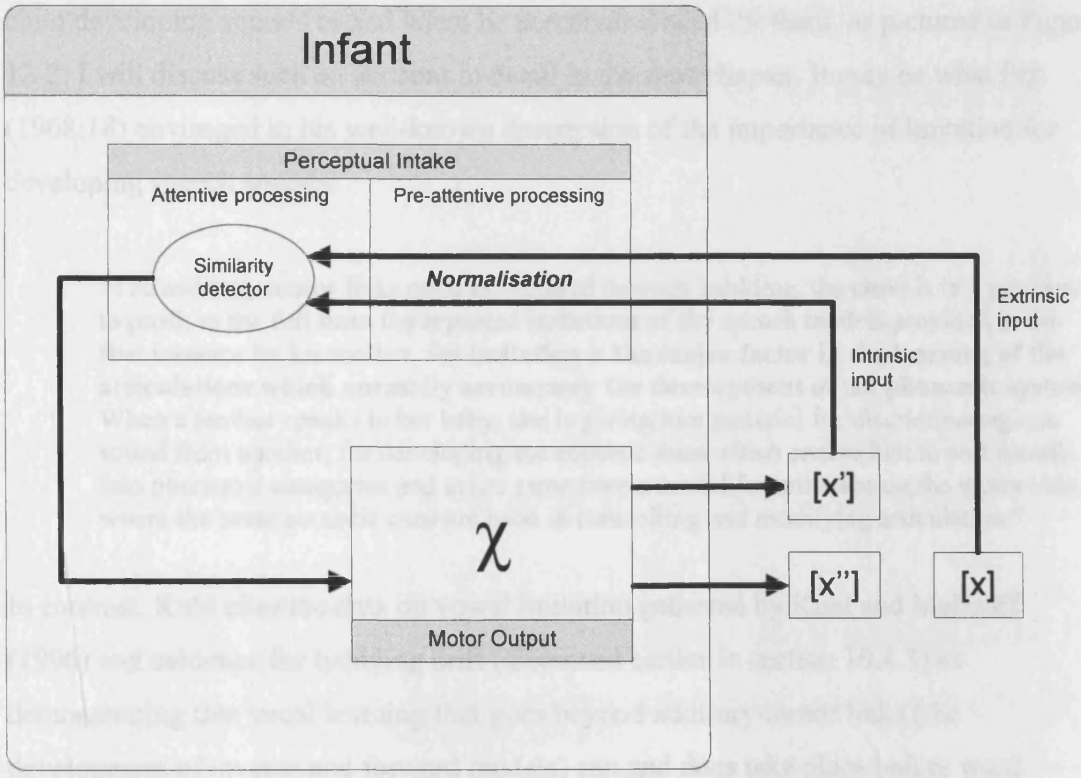

Figure 12-2. A simple copying mechanism for speech sounds would have extrinsic ([x]) and intrinsic ([x']) inputs compared (post-normalisation); judgments of similarity and discrepancy drawn by the infant; informing his next attempt at production ([x"])

Kuhl's account of speech development largely confirms mainstream beliefs:

> "Infants learn to produce sounds by imitating those produced by another and imitation depends upon the ability to equate the sounds produced by others with ones infants themselves produce." (Kuhl 1987:351)

> "Homo sapiens is the only mammal that displays vocal learning, the tendency to acquire the species-typical vocal repertoire by hearing the vocalizations of adults and mimicking them." (Kuhl and Meltzoff 1996:2425)

> "Infants not only learn the perceptual characteristics of their language, they become native speakers, which requires imitation of the patterns of speech they hear others produce. Vocal learning critically depends on hearing the vocalizations of others and hearing oneself produce sound. ...
>
> Imitation forges this early link between perception and production. By 1 year of age infants' spontaneous utterances reflect their imitation of ambient language patterns." (Kuhl 2000)

What I will call a 'mainstream' or 'conventional' view of the process would have a child developing sounds as and when he perceives a need for them, as pictured in Figure 12-2. I will discuss such an account in detail in the next chapter. It may be what Fry (1968:18) envisaged in his well-known description of the importance of imitation for developing speech sounds:

> "The auditory-motor links once established through babbling, the child is in a position to profit to the full from the repeated imitations of the speech models provided in the first instance by his mother, **for imitation is the major factor in the learning of the articulations which normally accompany the development of the phonemic system**. When a mother speaks to her baby, she is giving him material for discriminating one sound from another, for developing the acoustic cues which enable him to sort sounds into phonemic categories and at the same time a model for imitation on the motor side, where the same acoustic cues are used in controlling and modifying articulation."

In contrast, Kuhl cites the data on vowel imitation gathered by Kuhl and Meltzoff (1996) and evidence for babbling drift (discussed earlier in section 10.4.3) as demonstrating that vocal learning that goes beyond auditory-motor links (the development of inverse and forward models) can and does take place before word learning. She proposes that representations of speech sounds are stored in memory for use in perception early in life, and that it is these that subsequently guide the development of speech production rather than adult tokens captured during the period of word learning (Kuhl 2000:11854). See figure 12-3 for a modification of figure 12-2 which portrays this two-stage process[95].

[95] Analogously, the classic model of birdsong development posits an auditory template learnt during a memorisation phase, which is later matched in a motor phase of learning (Shettleworth 1998:457).

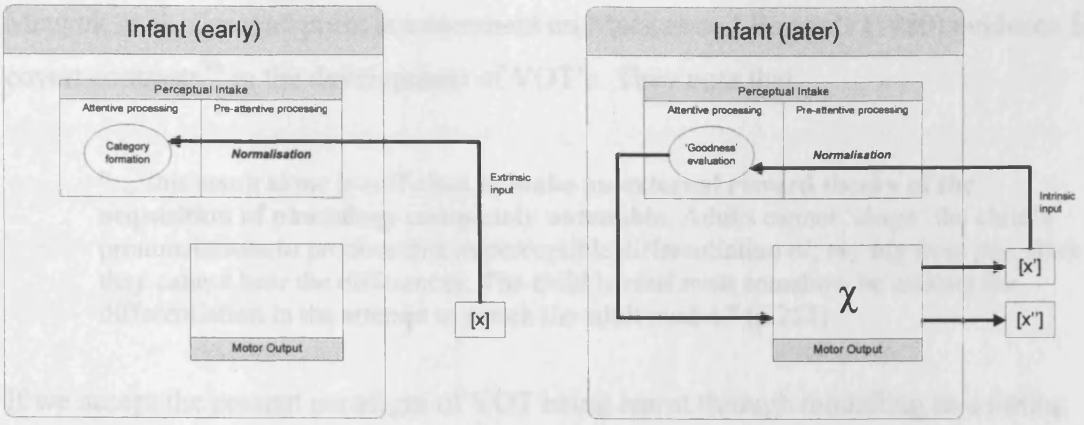

Figure 12-3. Kuhl's two stage process for learning speech sounds: on the right the infant compares his output against the model set up earlier for speech perception

Returning to the mainstream view, Jones and Munhall (2000:1247) describe how this is applied in models of speech development:

> "In formal models of acoustic-articulatory mappings, acoustic feedback plays a number of possible roles: (1) For speech sound development in children and adults and for learning new vocal tract arrangements, acoustic feedback provides the primary information about target achievement and thus is the vehicle for learning."

Among similar accounts of vocal learning, Menyuk et al. (1986:209) describe the first three steps of a "rough" sequential production task analysis for a child as being,

> "(i) to learn to produce a variety of vocal sounds
> (ii) to learn to produce vocal sound patterns so that they more or less match sounds which are heard (imitation)
> (iii) to learn to remember certain sound patterns well enough to produce them without just having heard them (delayed imitation)."

They then make two noteworthy points. Firstly, that

> "... the existence of such [transduction] routines on level (ii) (imitation) carries no implication that any general routines have been developed on [level (iii) (delayed imitation)]. That is, a child who can imitate new words with a particular degree of skill may be unable to produce them as well after a time lapse."

This draws attention to what may be two distinct forms of 'imitative' performance. It seems to me that their 'imitation' and 'delayed imitation' will often correspond to what I will describe using the terminology 're-creation' and 'reproduction'.

Menyuk et al.'s second point is a comment on Macken and Barton's (1980) evidence for covert contrasts[96] in the development of VOT's. They note that

> "... this result alone is sufficient **to make an external reward theory of the acquisition of phonology completely untenable**. Adults cannot 'shape' the child's pronunciations to produce this imperceptible differentiation of, say *big* from *pig*, since they cannot hear the differences. The child herself must somehow be making the differentiation in the attempt to match the adult model." (p.214)

If we accept the present paradigm of VOT being learnt through modelling as a timing phenomenon, then this argument against reinforcement learning mechanisms in child speech is logically undeniable. Together with the evidence for the early appearance of pre-fortis clipping, it helps to paint a picture of even the very young child as a "junior phonetician", noting fine temporal detail in adult speech to reproduce it in his own. This image is then easily extended to imagining children replicating other phenomena, including speech sounds, by imitation.

However if, as I argued in Part 1, the timing aspect of VOT is actually epiphenomenal in child speech, and the observed differences in English, for example, reflect the strength of stress pulses and aspiration rather than any temporal targets, then this argument loses its force. The differentiation described above would not be an intentional contrast and at least a supporting role for reinforcement in the "acquisition of phonology" may be allowed.

12.3.2 The child copies a gestural model

While most accounts of speech reproduction view the acoustic signal as the mediator in word imitation, Motor Theory has inspired gestural accounts of how children learn to pronounce speech sounds.

Thus Goldstein and Fowler (2003) argue that Articulatory Phonology provides a foundation for understanding a core requirement in language, that there must be a common phonological 'currency' among producers and perceivers (Liberman and Whalen's (2000:189) 'parity requirement'). Infants such as those tested by Kuhl and Meltzoff (1996) are not in a position to rely on acoustic similarity between their vowels and the model's to verify that they are imitating successfully. Instead, they achieve this

---

[96] A covert contrast (Scobbie et al. 1997) is a reliable distinction in production which adults cannot hear. Some may be revealed by instrumental analysis.

by perceiving actions of the vocal tract, attending to the distal event rather than the proximal signal (as, it is claimed, all infants do in all forms of perception). Infants are able to detect correspondences between what they perceive to be the actions of adults and what they themselves can articulate. However, additional mechanisms do operate: within-organ contrasts emerge through the acoustic medium, via a 'mutual attunement' mechanism (p.192).

Other accounts have followed motor theory in postulating that specific neural mechanisms are involved. These lead to perception being shaped by the gestural requirements of production:

- The 'perceptuomotor account' held that imitation is possible because audition structures the signal in articulatory terms (MacKain 1988:64; Studdert-Kennedy 1987).
- Skoyles (1998, 2002a, 2002b) proposes that phones have been selected (through evolutionary pressures) to be a replication code, i.e. that their imitability is their key characteristic. His view on how they are imitated does not seem to be essential to this argument, but he describes a "fluent, automatic and prelinguistic ability to transfer audition to vocal motor programming" (Skoyles 1998), not distinguishing mimicry from reproduction in this regard. He gives various reasons for believing that "vocal imitation concerns the copying of articulations through the copying of their motor goals" (2002a) and posits mirror neurons as the mechanism by which the brain achieves this conversion.
- Wilson (2001) discusses a wide range of 'imitatible' stimuli, arguing that these are "parsed and encoded in terms of articulatory gestures" and exploring the explanatory potential of this claim for speech, memory, perception and so on. She suggests that the connections between perception and imitative motor commands are "innately specified" (p.548).
- Studdert-Kennedy (2002:219) has also explored a role for mirror neurons to solve the correspondence problem, explaining speech sound development by linking them with Meltzoff and Moore's (1997) Active Intermodal Mapping, which posits an innate mechanism by which perceptual patterns are transformed to motor patterns[97].

[97] AIM, mirror neurons or both may turn out to be the basis for the learning of speech sounds. However, it seems premature to rely on either of these mechanisms yet. The criticisms of the tongue protrusion/head turn experiments which are an important support for AIM (e.g. Jones 1996; Anisfeld 2002) do not seem to

12.3.3 The child discovers the (speech) sounds he makes already being used linguistically by others

The child might use a judgment of acoustic similarity in a different way from that described in mainstream accounts. His own production of something that will be a speech sound but is initially developed independently of any input, might prime his perception so that the sound's salience and familiarity in the speech of others is increased. In this way the child would be able to pick out parts of words that he is able to at least partially reproduce. He would discover that sounds he already makes are of linguistic significance.

Similar ideas to this are found in recent general work on imitation. For example, Byrne's (2003) "Behaviour Parsing" model makes it possible for novel, complex behaviours to be acquired by observation. Byrne describes an essential preliminary stage to this of segmenting observed behaviour into a vocabulary of elements discerned which, if they are both to be 'seen' and used as building blocks in effective planning, must meet one simple principle: that each element should already be within the repertoire of the observer. He applies this conceptual approach to the preparation of bundles of nettle leaves for eating by mountain gorillas, but it maps as appropriately onto learning the articulatory elements of words and their sequencing.

In Wolpert et al.'s (2003) MOSAIC model of human motor control, 'imitation' occurs via multiple controller-predictor pairs that formulate the expected consequences of an attempt at an imitative action in advance of performance. The pair whose prediction best matches the outcome are then chosen (or assigned 'high responsibility') for subsequent production. Again, in contrast to copying approaches (where the target action is observed, definitional characteristics noted and then attempts made to reproduce these) this starts with the observer having actions in his repertoire, and then looking for whether their consequences are also found in the target. The 'active imitation' of Demiris and Hayes (2002) functions similarly, and in machine learning the "mixture of experts" approach works on similar principles.

---

have been refuted by its supporters. Nor is there comparable evidence that the mechanism operates beyond early infancy. Similarly, the role of mirror neurons in auditory-motor mapping is still unclear for adults and presumably may be different in children.
Nevertheless, I add my own speculation about mirror neurons to the pot in section 15.2.1.

Returning to speech, the corollary of this would be if the infant notices that a sound in his mother's speech is similar to one he makes himself, and therefore that he can reproduce it. I will call this 'echo re-enactment', because the question being figuratively asked by the infant could be phrased as, "is the sound my mother makes like the echo of, or similar to, a sound I can make?"

Figure 12-4 is another modification of figure 12-2, which portrays the two stages of echo re-enactment.

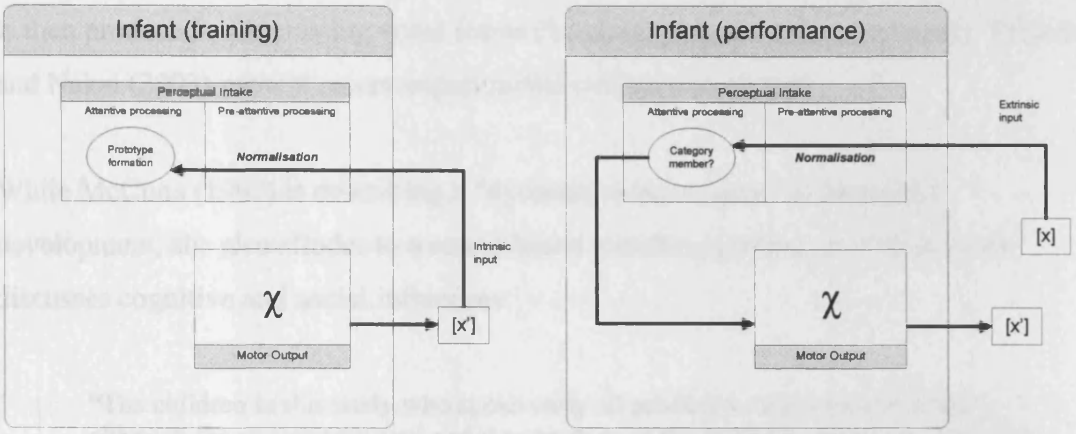

Figure 12-4. The two stage process of echo re-enactment: in the second stage, the infant judges the similarity of the extrinsic input to the sound produced by his VMS. A match can be reproduced with the VMS.

Similar mechanisms to this have been described by a number of authors, prominent among whom are Vihman and McCune, who build (as I do) on their notion of a vocal motor scheme (VMS), as discussed in section 10.4.4.

Vihman describes an 'articulatory filter' guiding the infant towards the production of words that are within his grasp:

> "It is through the mechanism of attention to his or her own babbling, or vocal exploration, that the child discovers the link between the phonetic gestures underlying speech and the acoustic patterns that accompany them (Fry 1968; Netsell 1981). This "discovery" gives rise to what we may term an "articulatory filter", a phonetic template (unique to each child) which renders similar patterns in adult speech unusually salient or memorable; in particular, **the filter picks out patterns for which the child has already established a "motor plan"** (Locke 1986) or "gestural score" (Browman and

> Goldstein 1992; Kent 1992) **specifying the articulatory implementation which will result in the particular sound pattern** (Vihman 1991)." (Vihman 1993:74)

> "Deployment of vocal motor schemes reflects (1) the exercise of emergent neuromuscular control within the framework of a changing vocal tract, (2) implicit learning of the distributional patterns in the input, such as the vowel space characteristic of the particular ambient language, and (3) the 'reinforcing' effect of **the auditory match of the child's own vocalizations to input patterns – yielding proprioceptive knowledge of how specific sound patterns may be produced**. We term the process of matching own vocal patterns to the input the articulatory filter." (Vihman and Velleman 2000:312)

This seems to describe a speech sound matching mechanism, but Vihman's current view of the articulatory filter has it highlighting a word with the sound of a VMS in it, which is then produced with existing vocal forms (Vihman, personal communication). Vihman and Nakai (2003) present recent experimental evidence in support.

While McCune (1992) is describing a "dynamic systems view" of first word development, she also alludes to a sound-based matching process (p.330) and then discusses cognitive and social influences:

> "The children in this study who spoke early all produced a repertoire of VMS's, although the consonants used and the structure of the resultant utterances differed by child. **The availability of VMS's provides an opportunity for the parents to respond to the child's consistent vocal forms as if they were words**, and to attribute meaning to them in relation to objects or events that appear to be the focus of the child's interest. For example, an infant who has the organic capacity and behavioural propensity to say [ba] is in an excellent position to produce words such as "ball" and "block" if his mother presents and names such objects, then welcomes [ba] as a name for the object. In this case aspects of the social and non-social context interact with developmental characteristics to initiate a phase-shift from "babble" to "word." A child who infrequently produces [b], or whose mother neither expects words nor presents them for imitation is unlikely to exhibit first words in this manner. This system can be considered self-organizing because the elements might interact in this way with no prior intentionality or direction. The parent who labels objects for the child may initially do so to enhance interest, with little expectation that the child is ready to speak. The child who emits a familiar repeated vocalization under appropriate circumstances may initially do so with no intention or awareness of the goal of "speech." Finally, the parent who accepts a child vocalization as a word may believe the child has somehow learned the word, although production of the pattern may be only accidentally related to the apparently appropriate context. In this example systematic relationships between child organic status and elements of context interact, leading to (apparent) word production (Veneziano 1988). Even without the knowing intention of both parties, **this situation tends toward development of words**." (p.331)

This may very well be correct. However, I will propose that the parents' response to a child's consistent vocal forms as if they were words may be even more significant at an earlier stage, in that it also tends towards development of speech sounds.

It seems to me that Davis and MacNeilage (2000) make a proposal that is close to echo re-enactment as part of their embodiment perspective on the development of speech perception. They consider what the effect on perceptual organisation would be, "if it was strongly influenced by the characteristics of the child's own behavioural repertoire," (p.230) which generates 'intrinsic' input to its perceptual system. They argue that the effect of embodied production types seen in babbling being part of the words of a language is that, "a perceptual task for the infant in correctly producing a word with a favoured CV co-occurrence pattern is not to generate a correct perceptual representation, but to match an *already existing* intrinsic pattern with extrinsic patterns from the environment." (p.237)

MacNeilage et al. (2000:161) discuss how basic patterns give the infant, "a ready-made initial access to the ambient language," reducing his initial task to, "one of fitting specific available output patterns to adult words that have similar patterns." Lewis (1951:99) makes a similar suggestion in the quotation I reproduced earlier.

Finally, Gattegno (1973:50) suggests that a discovery style of mechanism is something we shouldn't be surprised to find. Discussing actions like crawling, standing and walking he says:

> "Again, the process is entirely an inner one. **Only when a baby knows how to perform certain acts can he find that a similar act is being performed by others** – similar, not the same, for the action he sees uses another soma [body] and is perceived from outside while he knows his own action from inside."

One of his last discussions of speech development comes in Chapter 13 of *The Science of Education* (1985). I have reproduced pp. 11-21 of this in Appendix C. On page 20 he says:

> "Babies spend time now peeling words out of voices and when they meet sounds they themselves can make they keep them within one category: that of those 'common to me and them.' It is the time thus spent which will become a solid bridge, but is also a beachhead for the conquest of 'speaking.'"

One of the attractive features of an echo re-enactment account would be that the work the infant must do is distributed in time. Learning articulation to create VMS's would happen before the speech sounds they create are recognised in the speech of others and

then deployed by the child, rather than being part of that latter process. This will also be true of other 'discovery' accounts.

12.3.4 The child discovers gestures he can already make being used linguistically by others

There is a final combination of answers to my two questions, described in the title of this subsection. To my knowledge, no account has explicitly proposed this, although with one important modification it would describe the proposal I will make in chapter 14. However, in parts of Locke's writing it seems to me that he is proposing something along these lines, although I am uncertain about the exact meaning of the passages I quote below. He described how there might be, "a motor basis to speech perception ... and internal representations" (1986:245), and expresses this idea, I think, in a number of ways. I have highlighted what seem to be some relevant sections in passages below:

> "At this point, I believe something can be said of the child's phonological debut; how he breaks into a phonological system, and why his system initially is different from his parents'. It appears that as the child reaches that point in his social and cognitive life in which it is both desirable and possible to designate things with strings of sound, the **child reaches – as it were – into his collection of readily available articulations**. The available articulations, at this point, are the segments of his babbling repertoire. It is a foregone conclusion that **these articulations will be projected**, much as they are in the case of "invented words" in which there is no identifiable adult model. Since the infant at 12-18 months has a fairly well developed perceptual system, in projecting his available articulations the infant will produce a number of "hits." Many of his [d]s and [b]s will land on lexically standard /d/s and /b/s. A number also will land elsewhere, on /D/s and /v/s, because the child (a) may not notice that those sounds differ from [d] and [b]; (b) may not know that the difference matters, in some sense; or (c) could not make the necessary articulatory adjustments even if he did notice or respect the difference between the adult form and his own." (Locke 1983:83)

> "How well does the child have to be able to perceive speech – others' and his own – to develop his first words? Not well. At or before 12 months a number of children have been observed to say [da]- or [dæ]-like syllables in reference to 'daddy'. Since [d] is of considerably greater frequency in the babbling of 11-12 month olds than the other stops, one might hold that all the children have to perceive is [+stop]; if they were to produce their most frequent stop their replicas would 'automatically' be alveolar and voiced. In other words, I am not sure one can credit the child's perceptual system for all the matching features in the child's response if some can be motivated extra-perceptually.
>
> If infants are predisposed to project innate movement patterns onto standard words (as they perceive them), only *contradictory* information may be needed from perceptual analyses for the child to avoid unintelligible results. A fairly crude analysis of his own sounds in relation to environmental sounds would be sufficient, and **there would be, therefore, a motor basis to speech perception – and internal representations – from the very start**.
>
> Since I argued earlier that the infant's babbling has a largely motor or sensorimotor basis (recall that deaf infants have a similar consonantal repertoire), and as there is

> considerable overlap between the phonetic features of babbling and of speaking, it follows that when the young child draws upon his phonetic experience he is drawing largely upon his articulatory motor experience. The infant learns that his pre-lexical [dædæ]'s and [mæmæ]'s are admissible as representations of standard daddy and mommy. **Though they will become auditorily enriched through experience, the earliest representations or phonetic plans of children may be primarily motoric in nature.**" (Locke 1986:245)
>
> "When the approximate form of [the units which permit lexical communication] is known, the child in turn knows which of his currently available phonetic segments are lexically admissible, and in what sequence they ought to be deployed. It follows from this that the early internal representation of words may contain less phonetic detail than the child's speech might suggest, and that **the early phonetic plans may contain a fair amount of motoric information.**" (Locke 1986:250)

12.3.5 Neural mechanisms

Various people have suggested that some form of 'embrainment' (Mareschal et al. 2007) holds the key to understanding how the correspondence problem is solved. Earlier I mentioned some 'gestural' accounts which would at least partly transfer the challenge of creating equivalence classes from the psychological to the neurological domain. Similarly Lieberman (1980) suggested that talker normalisation is achieved by a specific, innate brain module, which would enable a straightforward copying mechanism along the lines I described first.

Westermann and Miranda (2002; 2004) present a computational model of the production of vowel sounds by an infant, in which a neural mechanism generates a coupling between perception and action. This doesn't fit easily into my classificatory framework for two reasons. Firstly because it is, as the authors point out, only a suggestion for a way forward, which has been modelled without taking account of the differences between the outputs of child and adult vocal tracts. Secondly, because once this simplification of reality has been made, no explicit judgment of similarity is needed for the infant agent to learn: 'imitation' of speech sounds is automatic.

For some speech sounds, small changes in articulation lead to only small changes in the output, meaning that these sounds can be produced reliably. In Westermann and Miranda's model, Hebbian activation between units on motor and auditory maps leads to the concurrent emergence of preferred articulations and auditory prototypes corresponding to these sounds. The existence of an auditory prototype then induces perceptual changes by displacing ("pulling") inputs towards it, thus warping the infant's perceptual space.

This learning mechanism would bypass the conventional problems of imitation:

> "After the model has developed the sensorimotor coupling [based on its own 'babble'], it can utilize this coupling to imitate heard sounds. A heard sound evokes a response on the auditory map, and through the Hebbian connections, an associated motor parameter set is activated that can then be used to produce the sound. **Imitation here is not the accurate re-play of the heard sound, but instead an interpretation of this sound within the developed sensorimotor framework.**" (2004:396)

Two simulations are reported. In the first, the agent hears only its own output. Then it hears its own output mixed with sounds from an L1. In the latter case the majority of the preferred articulations and auditory prototypes that the model develops correspond to those in the target language.

However, the agent's production and perception in this simulation operate exclusively with adult sounds, rather than with a realistic mixture of infant and adult sounds. These would occupy non-overlapping areas of the formant space and require some form of transforming algorithm between them.

The model turns the learning of speech sounds from being an imitative problem to being an automatic interpretative process. Westermann and Miranda point out its further explanatory potential, including the following:

- The model shows how perceptual events could affect motor representations: simple exposure alters representations towards the vowels of a language. "This mechanism suggests that the adaptation of infant babbling to the ambient language might not depend on the explicit [and immediate] imitation of sounds by the infant (Kuhl 2000), but that instead imitation is made possible by the reinforcement-based adaptation." (2004:398)
- Mirror neurons develop naturally from the correlated coupling of actions and their perceptual consequences, but are an emergent property of this coupling and not a mechanism for the imitation of sounds.
- The model may help to explain why it is difficult to learn a new phonology when learning a second language, because when new sounds are perceived as if they were known sounds there is no adaptation because there is nothing new to learn.

However, the model does not address the normalisation problem, and it would be vulnerable to some of the potential problems I will describe in chapter 13. While it appears to have been devised to explain babbling drift, it does not appear to be limited to this stage of infant development and I think it could be more profitably applied to later stages. If this model can be developed further, then it would seem to offer an alternative way for speech sounds to emerge without being imitated per se.

13 Potential problems with some models of speech sound development

13.1 Introduction

In this chapter, I will complete my consideration of imitative processes in speech, building on what I have already described in chapters 7 and 12. I will identify two senses in which we might say that a child learns the qualities of speech sounds 'by imitation':

- through mimicry, in a continuation of the development of the auditory (sensorimotor) inverse model (IM) which began with early vocal play;
- through re-enactment, by which a new 'speech sound' (perceptuomotor) IM is developed.

The latter is generally considered to be how children learn the pronunciation of speech sounds. Both child and adult learners are commonly believed to judge the similarity of their speech sound output against an externally-derived standard, in a matching-to-target process of increasingly correct approximation.

From section 13.3 onwards, I discuss the potential problems with each approach and I make a case against the conventional account of the development of speech sounds being possible. Some of the problems with this that I identify are novel (to the best of my knowledge), but none are yet supported well enough to be fatal to any account.

Up to now I have only described children learning L1, but now it seems appropriate to also consider the position of students of foreign languages. Pronunciation teaching is not very effective, and conventional approaches to it, based on beliefs about childhood learning, may be part of the cause[98].

[98] Of course, the two situations are not identical. For example, an infant learning L1 is believed to have an advantage over older learners of L2 whose hearing is so effectively adapted for efficient L1 perception, a task different from their new one. On the other hand, the older learners may have an advantage in their ability to engineer the presentation of target sounds from a teacher and other sources.

13.2 Concepts in social learning

In earlier chapters I began development of a conceptual framework for social learning adapted to the question of how children learn to speak. This need only include a subset of all the many social learning mechanisms, for which there are several reviews (e.g. Galef 1988; Heyes 1994; Nicol 1995) and also reviews focussed more particularly on man and other primates (e.g. Want and Harris 2002; Whiten et al. 2004). Here I will complete this framework.

It will be important to be sensitive to three properties of a motor act. In the previous chapter, I introduced the first of these, an act's perceptual transparency/opacity. In this chapter, I will discuss two possible mechanisms of 'imitation' that rely on speech signals being wholly transparent, as generally believed. However, I will then argue against transparency in the particular case of speech sounds, and in the next chapter these will be taken to be at least partially opaque.

The second property is the availability of an act's results. These may persist only as long as actions are being performed, and I will describe the signal produced in this case as ephemeral. Dance and speech are examples. Other results persist in time beyond performance of the act, such as the trajectory of a ball that has been kicked or the marks left on a page by handwriting.

Thirdly, it will also sometimes be important to distinguish three aspects of a motor act:

1. The **act**: the external, completed view of the event (the 'container', or the deed).
2. The **action**: the internal, progressive view of the event (the 'contents', or the process of doing something).
3. The **form of the action**: the information about the event typically revealed by a light or a sound signal.

Finally, I should note that my use of the terms 're-create', 're-enact' and 'reproduce' makes distinctions between them that they do not normally carry. I use 're-create' to describe the process of mimicry (where there may not be an external target, but the auditory IM is still being driven by a sound image from memory), 're-enact' as defined in the next subsection, and 'reproduce' for production from a perceptuomotor IM which has not been based on self-made judgments of similarity between extrinsic and intrinsic inputs.

13.2.1 Distinguishing mimicry, pantomime and 'purposive' copying

Mimicry is not regarded as a particularly significant mechanism of social learning in non-human animals, but is clearly a more prominent activity among humans and one which certainly plays a part in speech development. In section 7.1, I noted that I would want to revise the definition of it given by Call and Carpenter (2002), who described mimicry as B copying the form of A's actions without adopting her goal or intending to achieve the results she obtains. The first part of this suggests there is a closeness between mimicry and imitation, and a distance between their operation and that of emulation, for example. In contrast, I would like to emphasize the difference between mimicry on the one hand and all forms of purposive copying on the other, based on the mode of perception adopted by B in each case.

To start with, I would contend that Call and Carpenter's definition wrongly identifies mimicry with events in the world rather than events in an observer. As I suggested in section 10.2.1, mimicking someone or something should instead be defined as producing a signal which perturbs an observer's sensory apparatus in a way that resembles the perturbation caused by the signal from the target behaviour[99]. This is something only the observer can judge. Mimicry happens in humans (and rarely in animals) because we can consciously attend to a signal in our AS mode of perception in preference to the MP mode. Having the capacity for this, we can also recognise earlier experience in the present experience on the basis of their resemblance. The action that leads to this recognition is called mimicry.

'Resemblance', then, is a matter of subjective judgment rather than an objective property of a behaviour or thing. This distinction is not always clear or important to make in practice. With respect to visual mimicry, it is often convenient for a mimic to achieve resemblance simply by reproducing A's actions; so it is not surprising that a workable definition of mimicry can be based on this. However, by considering sound mimicry we can see that similarity of actions is not necessary, since I can mimic a motor bike with loud, exaggerated humming and a radio technician can mimic a horse walking along a street using coconut shells. The important thing in these cases is the resemblance between the sound signal produced by the mimic and the target signal,

[99] Hence 'impressionist' as a synonym for 'mimic'.

which is to say the resemblance in the responses that our hearing apparatus makes to the perturbations created by each sound.

In contrast, it seems to me that if there is no (ostensible) purpose to the result of copying forms of actions, then this copying is better described as a form of pantomime. This is defined by Duffy and Duffy (1975) as, "pretended deliberate motor sequences using no object", and by Westwood et al. (2000) as actions directed towards remembered rather than real targets.

The literature contains different definitions of both pantomime and mimicry. Adjusting an observation of Arbib (2005:115) to my preferred scheme, however, we can say that mimicry is performed with the intention of getting the observer to think of a specific action or event, and is communicative in its nature rather than purposeful in itself. While an imitator observes in order to act, the mimic intends to be observed.

As an example to clarify these distinctions, consider A cutting a piece of paper with a pair of scissors. If all the props are removed, then I would describe B as pantomiming the behaviour if he made similar actions with his hand as A had made, operating imaginary handles with his thumb and first finger. However, to mimic A he would extend his two first fingers and close his other fingers and thumb into his palm. Then by opening and closing the extended fingers the motion of the blades would be reproduced, and could release in an observer an appropriate visual effect of a jointed pair of articulators opening and closing.

I would therefore amend Call and Carpenter's scheme by giving a form of pantomime (as yet unnamed) the description they use for 'mimicry', and then adding mimicry as a process based on a subjective judgment of resemblance.

Figure 13-1 summarises some of the distinctions I have drawn thus far, updating the discussion of section 7.1.1 by focussing on the different modes of perception used by B, which distinguish mimicry from the final three processes of purposive copying.

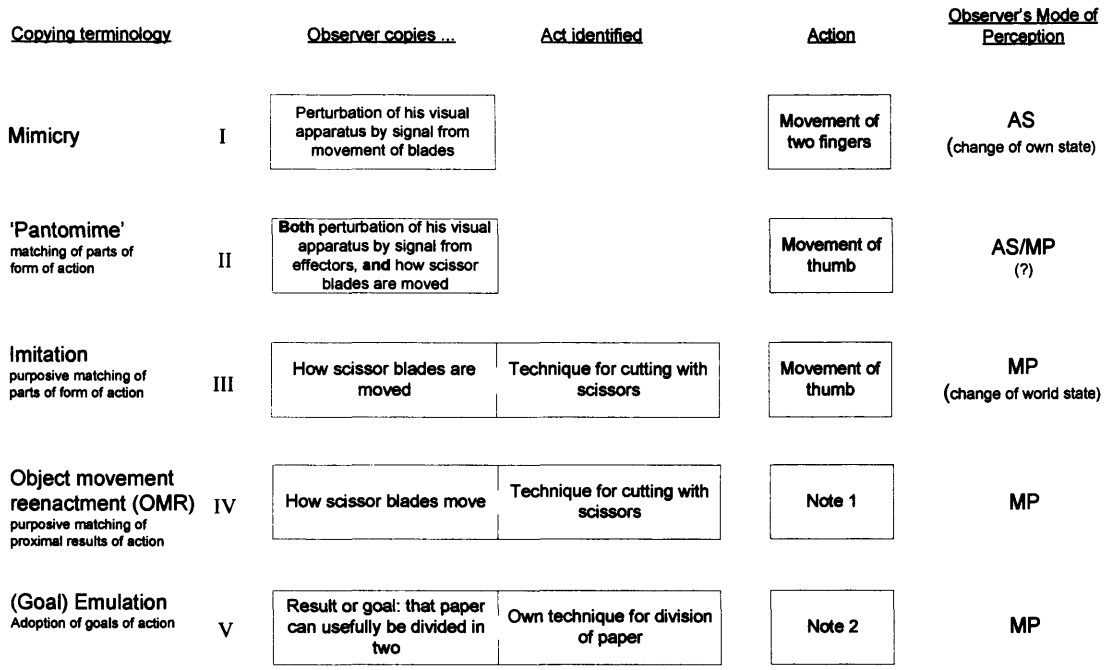

Figure 13-1. Copying processes and modes of perception. The demonstrator is using a pair of scissors.

The neurological data supports the dissociation of meaningless and meaningful actions. Vogt (2002b:540) summarises this as follows:

> "Rothi, Mack, and Heilman (1986) reported patients who, despite being unable to comprehend or discriminate gestures, could nevertheless imitate these gestures - possibly as if these were a series of meaningless movements. Rothi, Ochipa, and Heilman (1997) suggested that these patients bypassed an impaired lexical (i.e., semantic) mediation route and instead used a nonlexical, possibly iconic route that is normally only used for the imitation of meaningless gestures and may be spared in those patients with ideomotor apraxia who show improvements in imitation tasks. A similar argument for a direct, nonsemantic route from vision to motor control was made by Goldenberg and Hagmann (1997), based on a selective deficit of imitation of meaningless gestures found in two apraxic patients, which indicated a selective impairment of this route."

I mentioned some similar data for speech in my discussion of AS and MP in section 10.2.

Call and Carpenter's (2002) typology of imitative activity is based on the three sources of potential information revealed by a demonstrator. Their system is flexible enough to accommodate further development, for example, 'blind' and 'insightful' imitation

(Carpenter and Call 2002). In section 7.1, I added speech sound re-enactment (SSR) as a type of object movement re-enactment (OMR), in recognition of the nature of speech sounds as inherently meaningless, tool-like intermediaries. This means that mimicry and SSR are the two conventionally recognised copying mechanisms for speech, as shown in figure 13-2. ('Re-enactment' can be read as 'imitation' if the terminology is found unhelpful.)

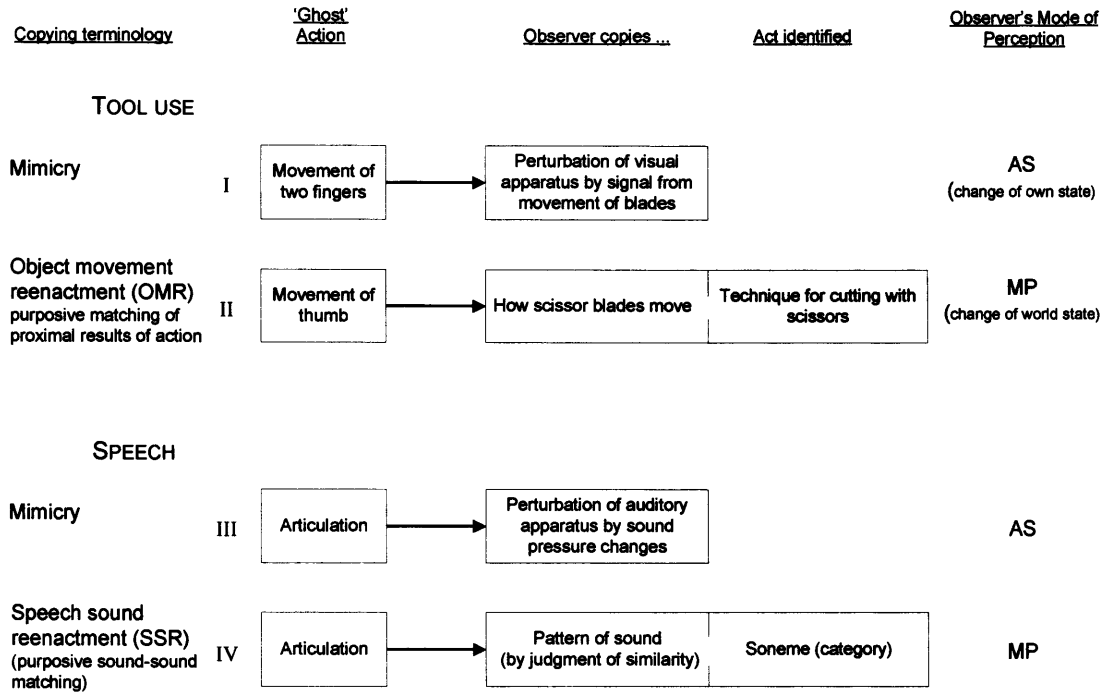

Figure 13-2. Words and speech sounds seen as tools. Below: two processes which might be believed to support the learning of speech sounds (the demonstrator's forms of actions not being apparent to the observer). Above: equivalent processes available for copying tool use (a pair of scissors) when this is under the control of a 'ghost'.

(In the next chapter a form of emulation will be added to these copying mechanisms. For purposes of comparison, it would appear below speech sound re-enactment and be described as 'purposive soneme-soneme matching'. Its boxed attributes would be, reading from left to right, 'Articulation', 'Identity of sound (by knowledge of equivalence)' and 'Soneme'.)

Call and Carpenter also pointed out how learning is typically a cyclic activity, and that in real life an observer ('B' or, if specifically a child, 'C') will be adaptive, attending to whatever source of information best meets what he judges to be his immediate needs. So in one cycle he might be imitating A, in the next emulating her, in the next, perhaps,

attending to a combination of sources of information in a way that does not yet have a label. Any attempt to characterise B's behaviour in the real world as just imitation or just emulation may be too reductionist. I will use the style of diagram they developed to illustrate this, for example in figure 13-3[100].

13.2.2 Two types of 'imitation': how a model might be used to solve the correspondence problem when signals are transparent

I can now ask about how a word's form might be adopted from ambient speech. There would seem to be four distinct possibilities.

1 The word might be re-created by the auditory (sensorimotor) IM as a whole-word shape. Infants may adopt their first words in this fashion, and I will return to this issue in section 14.3.2.

However, among supporters of this view it is uncontroversial that over time there is a movement away from this and towards (1) identifying recurring segments within word forms and (2) changing the basis of production to one of concatenating the production of these segments (e.g. Ferguson and Farwell 1975:422; Locke and Kutz 1975:185; Nazzi and Bentoncini 2003). The sequence of events and reasons for this may be as Kent (1981:179) describes:

> "A child's first words, then, may be motoric units; they may not have phonetic components that are easily transferred to other phonetic contexts. Motor control that is adapted to the production of phonetic segments in a variety of phonetic contexts in different words perhaps comes about as the child discards the principle of preparing word-sized motor sequences for each word in his or her lexicon. That is, **the child is forced to a segmental (phonetic) motor organization through sheer force of economy and manageability**."

If the preparation of 'word-sized motor sequences for each word' happens at the time of production, through evocation of a sound image from memory to drive the auditory IM, then the mental effort required may be the problem. Even when evoked weakly (outside

[100] The style of diagram is helpful, but I actually prefer Whiten et al.'s (2005) conceptual view. This describes imitation and emulation lying on a continuum of behaviours subsumed under the concept of (purposive) copying, with the question for researchers then becoming how B selects-in and selects-out information that he could potentially attend to.

conscious awareness), the need to have something "in mind" demands some attentional resource, while speech production actually becomes automatic.

2 Might these segments be re-created by the auditory IM, in a form of self-mimicry based on stored sound images? At first glance it seems unlikely: use of the auditory IM would require these sound images to be evoked from memory, to at least some minimal extent. But if segments initially produced this way become overlearned and free of the need for any attentional resource, then it seems to me that a part of the auditory IM would have become specialised and, for practical purposes, to have evolved into the mechanism described next.

3 The young child might, as generally believed, develop a specialised 'speech sound' (perceptuomotor) IM. Thus he would recognise a segment in a target word "as a" speech sound (via some form of inner representation that is, of course, controversial) and would then reproduce this. His recognition of an act performed by his mother would lead to equivalent action on his part. If he does this on the basis of the similarity of the sound images produced, which relies on the speech signal being transparent, then I am calling this speech sound re-enactment (SSR).

What learning would contribute to this speech sound IM?

Speech sounds are ephemeral, so attempts to imitate, in themselves, would not be sufficient. To improve his production, the child needs both knowledge of the action he performs (not just knowledge of the act) and knowledge of its results. He will be able to attend to only one of these while speaking and in an attempt at copying this would have to be knowledge of results (his output). Thus the imitative episode will be one of performance but not one of articulatory learning. It will be useful, nevertheless, as a test of the results of whatever nascent vocal motor scheme (regularised articulation) he performs. This can inform him of the need for differentiation or refinement of a speech sound.

To move his production forward, however, he can develop his VMS in side episodes of discovery learning (trial and error), and then test these as just described. So he would first enhance his IM and then see if the result is satisfactory. The learning of motor routines would be separate from their evaluation.

(In fact, given that a sound signal is ephemeral, the auditory IM also has to be developed in this way. By contrast, learning to write by copying an alphabetic letter is a situation where the results of actions persist. While using an IM to produce the letter from a target, the child can attend to the action he is performing because when he has finished he can "step back" and examine the results. In this case an imitative episode can be a source of learning about his motor performance.

Returning to sound, the simultaneous production and prolongation of sounds by model and learner should, in principle, allow a learner to both compare his output to the target and attend to his articulation, by moving his attention between the two. Thus he might learn from an imitative episode. For speech, however, prolonging a sound, where possible, is more likely to be of benefit during a side episode of learning where the speaker enhances his IM on his own in the absence of an external target.)

4 Finally, it is possible that the speech signal is not transparent, at least for speech sounds. I will describe how a perceptuomotor IM could be developed for opaque speech sounds in the next chapter.

With respect to the first three of these possibilities, it may be helpful to consider the handwriting of alphabetic letters further[101].

Before a child starts to form letters, he will have had some experience in drawing, transforming visual and graphic images into his own marks on a page. (Zesiger et al. (1997) call this a, "2-dimensional trajectory generator.") So he can initially copy letter shapes by drawing them, and his facility will increase with practice. It is possible to imagine writing as simply a highly practised development of this.

Forming a letter from a teacher's example by drawing it is, in general terms, a process of mimicry, where B re-creates a graphic image using a sensorimotor IM, judging the

[101] Handwriting is a linguistic skill whose output is certainly transparent but we should keep in mind that there are very significant differences between learning to handwrite and learning to pronounce, including (1) that the marks produced by handwriting persist, removing any need to capture the results of actions in memory, allowing straightforward inspection of results post-production, and enabling straightforward comparison with any model available; (2) that there is no normalisation problem, since a child can produce letters of the same size and shape as an adult; and (3) that handwriting is based on letters from its early stages, while the speech sound units of pronunciation are only likely to reduce to a phonemic level at the end of a long process (if at all, in normal speech production).

output in AS mode based on its resemblance to the original. The sensory image will be of A's output, as transformed by B through some greater or lesser processes of abstraction.

In time, however, does a child develop a distinct skill of writing, where he perceives a letter as an act which is associated with the action needed to reproduce it without an intermediate sensory image? The perceptuomotor IM to achieve this could be created by a process of signal re-enactment as I have defined it[102].

If we restrict consideration to just the forming of letters, then it seems to be widely believed that a distinct skill of writing does emerge out of an early use of drawing mechanisms (Levin and Bus 2003); i.e. that forming letters becomes something different from highly practised mimicry. A difference in the kinematics of drawing and writing similar forms, for example, is detectable at age 6 but not age 4, with specialisation increasing thereafter (Adi-Japha and Freeman 2001). Dissociation patterns in adult neurological patients also suggest that drawing and writing are distinct in various respects (e.g. Zesiger et al. 1997).

Why the change? Presumably the most significant advantage of a distinct production system for writing is its automaticity. There is no need for a sensory image to be 'in mind', so the child's attentional resource is freed. On a continuum from tracing to drawing to writing, the least (or no) attention will be required for the act of writing, even if some attention is required for monitoring written output (Gowen and Miall 2006; Tucha et al 2006).

13.3 'Imitating' speech sounds (1): problems with learning to imitate by mimicry

In the previous section, I described two ways in which B can learn to 'imitate' A's speech sounds:

1. by mimicry or re-creation (B attending to A's output in AS mode, to produce output that resembles it via an auditory IM); or,

[102] There may be other ways of developing a distinct skill of writing. For now, the important question is not how such a skill develops (which is outside the scope of this thesis) but whether it does.

2. by re-enactment (B developing a VMS whose output he judges to be a match for A's output, informing a new, speech sound IM).

People often describe vocal learning, including the learning of the qualities of speech sounds, as occurring by mimicry, so I have felt the need to define and discuss this possibility[103]. But simple mimicry is unlikely to be the path to learning the qualities of speech sounds for the reasons I have given related to the ephemerality of the signal (which I elaborate slightly in the next paragraphs). It may be that people are using 'mimicry' to describe the process that I have defined as SSR, which I will take to be the generally preferred account and discuss in some detail in the following sections.

Let me return to the question of whether an episode of mimicry is a way, in itself, of learning to imitate ephemeral speech sounds.

Recall Call and Carpenter's (2002) Figure 9.3 (reproduced earlier as my figure 7-2), which starts (top left) with B noticing the results of A's behaviour. Using the skills he already has at his disposal he emulates this, but the attempt is unsuccessful. As a result he pays attention to A's actions, and imitates these during a period of practice and in his subsequent, successful attempt.

Figure 13-3 and the figures of this style which follow in this chapter use this format but are now specifically concerned with speech. They all start with B hearing a word spoken by A. B parses this in the light of his set of sound-action correspondences (inverse model), identifying a string of speech sound categories (acts) through MP. There are correspondences of some kind between each act identified and actions which he uses to attempt to reproduce the speech sounds. However, imagine that the result for a particular speech sound is unsuccessful, for reasons other than poor motor control. He now has to work out how to improve his production of the sound.

---

[103] I should note that I am not concerned with a sustained form of mimicry where B applies a global transformation to his speech, perhaps via a change to his articulatory setting, in order to transform a complete utterance. For example, it seems that as a speaker of British English I can mimic an Australian accent by keeping my teeth closer together than normal when speaking.
(Whether this single adjustment works by genuinely distorting the action of lips, tongue etc in a way that resembles Australian practice, or just by acting as a mental trigger for a set of independent adjustments would be interesting to investigate. As far as I know the learning of dialects is not explained in terms of an overall articulatory setting for each, but I see no reason why it should not be.)

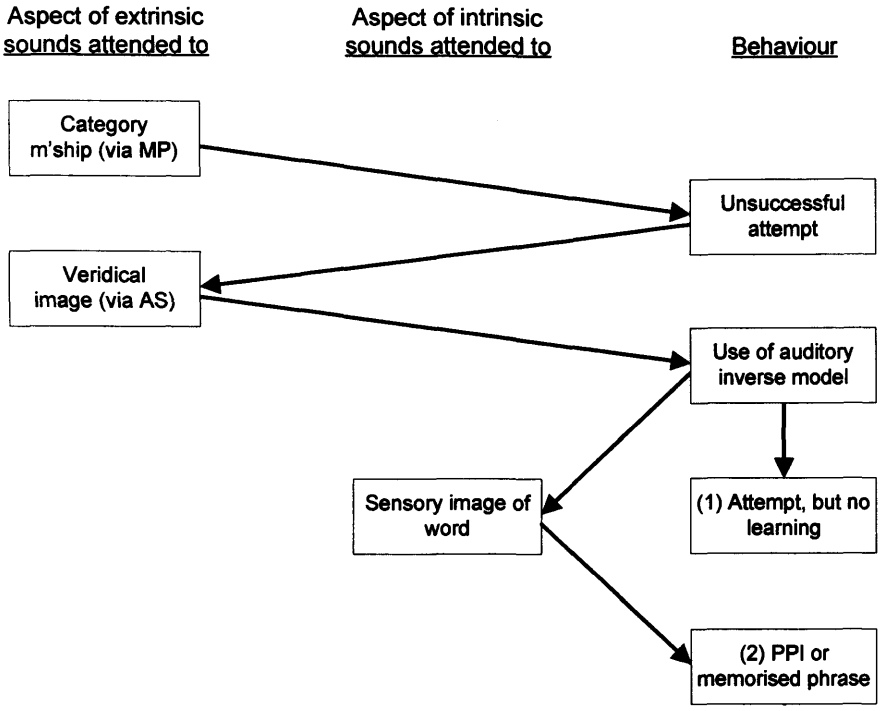

Figure 13-3. Alternative results of vocal mimicry. Tracking of a veridical (but normalised) image drives the learner's auditory inverse model.
Either (1) the attempt terminates the episode with no learning; or (2) a sensory image of the auditory feedback to the learner is retained as a means of recreating the behaviour in future. The result of (2) may be, for a child, a progressive phonological idiom (PPI), or, for an adult, a phrase 'run off' from memory as a sound sequence.

Figure 13-3 illustrates what might happen if he decides to mimic it. He switches his mode of listening to AS and attempts to capture a veridical image of the sound or word (from a subsequent performance by A).

He uses his image of A's output to drive his auditory IM (after some process of normalisation). At the same time, he must monitor the effect of his output on his auditory sense apparatus and compare this to the effect that A's output made. He is not, therefore, attending to the activities of his articulators during this process, which is what Gattegno argued he needs to do for him to learn something that he can re-produce later in the absence of sensory support. For this, he needs to know what he did with himself in order to achieve the match. Without it, his inverse model cannot be enhanced. He has performed but not practised: the first outcome in figure 13-3.

(B's monitoring of his own production may leave him with a sensory image of his own performance, or he may retain something of A's output in auditory memory. If so, he

can drive his auditory IM from the image. Providing his listeners find the result acceptable it appears that he has met the challenge of saying the word or phrase. In children such mimicry appears as progressive phonological idioms and in adults, for example, as L2 phrases that are run off from auditory memory. I have noted this as the second outcome in the figure.)

Learning new sounds

I have said that no motor learning takes place as a direct result of mimicry because the tracking of his output draws B's attention away from his articulators. What happens, then, in conventional foreign language classes where students are asked to 'listen and repeat' to improve their pronunciation of speech sounds? (They may be following the teacher directly or following a recording of some kind.) These instructions demand immediate performance and provoke a mimicked response from students. How, then, do students improve their performance?

My experience is that learners who follow the instructions they are given do not improve their pronunciation. Those that do improve, subvert the process; their teacher only requires them to mimic the sound, but they, in fact, initiate separate learning episodes which lead to them learning despite the instructions they receive.

Young (1995) describes how a foreign language learner <u>can</u> acquire a new sound (by means of what would be a side episode in a class being asked to 'listen and repeat'):

> "First of all, learning a new sound requires that the student realize that there is in fact a new sound to learn. He can then try to create the sound. In this case he is dealing with two independent but closely related systems, the mouth and the ear. Only one of these systems, the mouth, can be controlled voluntarily. All the muscles of the ear are involuntary muscles. The student can only modify the voluntary system. With his mouth he produces a sound which he guesses might be as close as possible to the sound he is aiming for. He hears the sound with his ears. Since he produced it with his own mouth, he knows that, muscularly speaking, his mouth was used in a new or special way and consequently he knows he should listen for a sound which is different from what he usually hears[104]. He can probably predict at least to some extent in what ways the sound will be different from what he usually produces. He speaks here with the deliberate intention of hearing something unusual and he listens to the result with the specific intention of hearing this unusual sound he has produced, creating a double feedback loop. He has feedback from his mouth telling him what it is doing and his ears give him feedback about what changes they detect as a result. Gattegno proposes that this is the process we all use to learn to produce new sounds.

[104] Coen (2006) notes Aristotle in *De Anima* saying that differences in the world are only detectable because different senses perceive the same world events differently.

> Once the student has managed to produce the sound to his satisfaction[105], he must practise it in a wide variety of different situations and contexts until he is completely at ease with the sound. He then reaches a stage where the sound has become completely automatised and the learning process for that particular sound is over."

Unfortunately, the alternative to 'listen and repeat' that Gattegno devised as part of the Silent Way is not widely known about. Nor is it a panacea, since it demands that a teacher develops both more expertise in the practice of pronunciation and greater sensitivity to learners and their learning than that required for conventional exercises. However, in my experience as a student and a teacher I have found that it is effective. (I say a little more about Gattegno's work in chapter 16 and appendices B and C.)

13.4 'Imitating' speech sounds (2): problems with re-enactment based on judgments of similarity

The second way that B might use A's output to copy a speech sound is by learning to re-enact it, in a self-supervised process that develops a new speech sound inverse model.

This is the mainstream view, as I described in section 12.3.1 where I explained what is thought to account for the time it takes and why the normalisation issue (from differing vocal tract sizes) is not thought to be fatal despite its mechanics not at present being understood. I said that learning to re-enact a speech sound is apparently straightforward and within a child's grasp, only involving the following steps:

1. The young child perceives and holds in memory (he 'captures') a target from adult speech (extrinsic input);
2. He produces an attempt at a match, capturing the acoustic result (intrinsic input);
3. He judges the similarity and differences between his attempt and the target;
4. Any error indicates how he can produce a better attempt next time.

The steps correspond to the numbered boxes in figure 13-4. Box 5 is similar to box 2 except that B is now performing a novel action. This may positively affect his ability to perceive the veridical acoustic effect of his articulation (as discussed below).

---

[105] On this point I am going to claim that his satisfaction will be based (at least initially) not on the sound he makes being similar to the target in his judgment, but being similar (or, more generally, equivalent) in the judgment of his listeners (and in a classroom, in the judgment of his teacher).

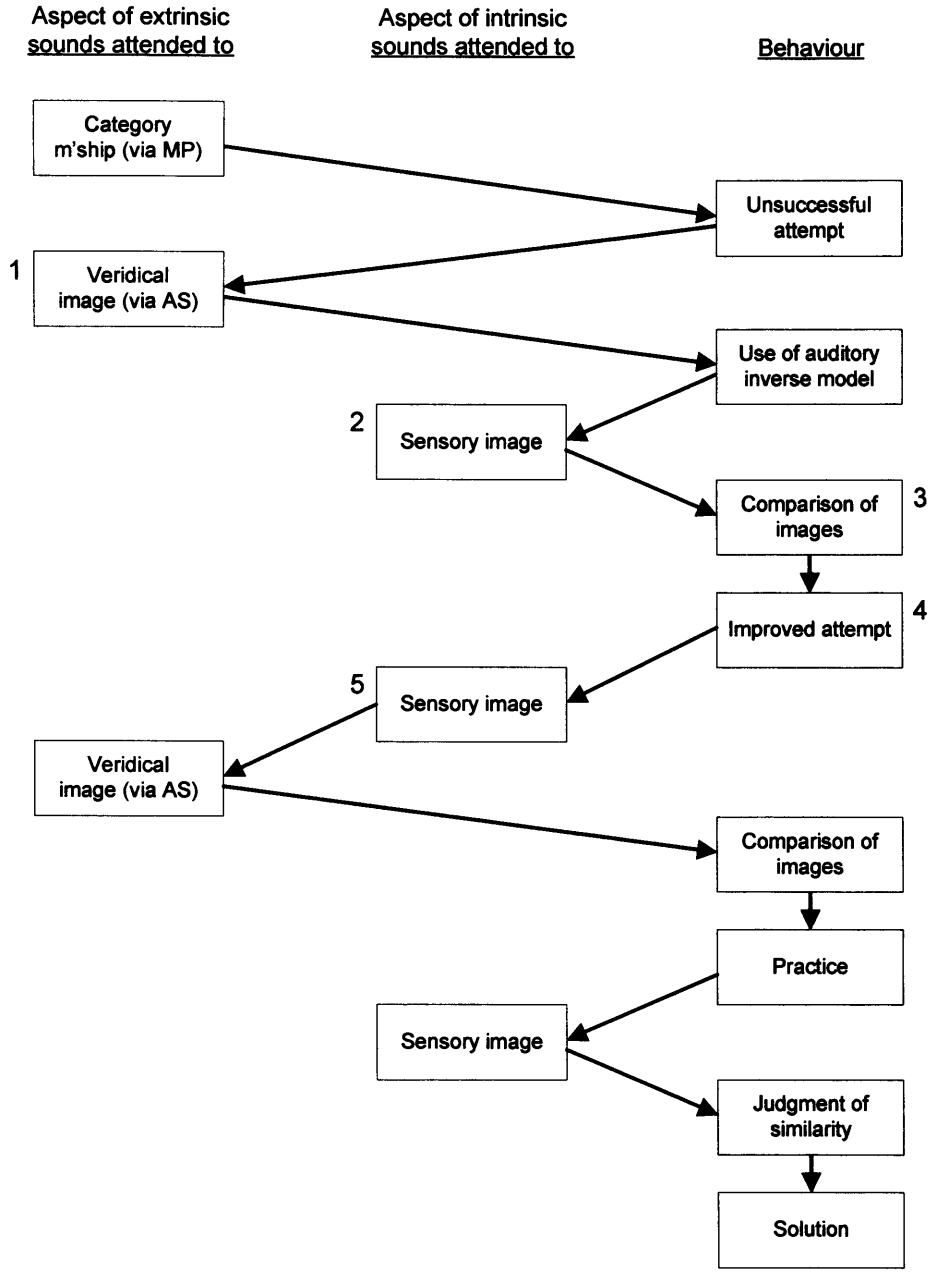

Figure 13-4. Possible sequence of learning in a conventional imitation account. After a first, unsuccessful attempt, B starts to explore new production possibilities, evaluating them against models from A.
Box 1, capturing extrinsic sounds; box 2, capturing intrinsic sounds; box 3, comparing images; box 4, using an error signal; box 5, capturing intrinsic sounds during performance of a novel behaviour.

In the following subsections, I raise various problems with the requirements of the first three stages of a re-enactment account:

- Can the child capture extrinsic speech sounds in form that is useful for improving his pronunciation?
- Can he capture his own output in a similarly useful form?

- Can he make the comparison and judgments on the two signals that would also be required?

My arguments differ in their strength and persuasiveness. None is backed with the evidence required for it to be conclusive. However, only one need be correct for any account that relies on similarity based judgments of equivalence (made by the child himself) to be undermined.

13.4.1 Capturing extrinsic sound(s)

In this subsection I discuss three issues related to capturing extrinsic speech sounds in a form that will be useful to improve pronunciation. The first is no more than an airing of some issues I find puzzling but feel may be relevant. The second is the main argument: that our two modes of perception mean that once a child has started to attend to words as being meaningful he cannot capture the veridical signal needed to copy pronunciation from them. Finally, I ask if some form of evocation from longer term memory could instead substitute for an inability to capture speech sounds from running speech.

How do we come to perceive something new?*

In due course, a listener is able to identify all the speech sounds of his language. He has formed a concept for each one and has 'attained' that concept, i.e. he has criteria for category membership (Bruner et al. 1956). This, at least in rudimentary form, is a prerequisite for the conventional account in which the qualities of a speech sound are copied for production. It accords with the notion that 'perception precedes production'[106].

As evidence that a child can identify an individual speech sound it is not enough that he is able to discriminate it from others.

Nor is it sufficient that he can recognise two words that form a minimal pair based on this sound and another. He may be able to perform this task without having marked the features that will enable him to copy the segment. For example he may be using the dynamic information available for perception, while his eventual reproduction will

[106] Shvachkin (1948/1973:113-120) describes a more complex relationship for child speech, and Sheldon and Strange (1982) demonstrate that something more subtle than this slogan implies must be possible. My arguments will lead to the conclusion that one form of perception does, indeed, precede production, but that this doesn't actually help production.

involve the holding of a well-defined steady state (unlike in highly coordinated adult production, Kent 1981:170). The steady state criteria may be something he has yet to learn for either perception or production, just like a foreign language learner (many of whom understand a language satisfactorily without properly perceiving or producing all of its vowels, for example.)

Signals we receive potentially inform us about many events which we do not, in fact, perceive. Lacking a theory or expectation about what we might see or hear we attend to other parts of the signal or fail to resolve information we do examine into an appropriate form. Examples include watching magic tricks, Huygens not seeing Saturn's rings (Gregory 1970:119), the artist's 'eye' (Jenkins 1980:225), 6-month-olds not attending to formant energy above 1000 Hz (Lacerda 1993), and generally any situation viewed by a novice as compared to an expert.

It is, of course, possible to move from not-seeing to seeing. One way is for an expert to direct one's attention appropriately. Another is to act in ways that reveal new information about the situation, thereby educating one's perception through action. A third is to reflect upon what one is (not) perceiving and to move one's focus accordingly.

With respect to speech sounds, it is not clear to me whether the vowel in *cut* is different from that in *cat* in a way that would it make it 'new' in the sense I am describing, or, alternatively, if the differences are just variations on a theme that a young listener would be familiar with: concentrations of energy at different points in the spectrum creating distinctive patterns.

Adult experience of L2 may not be a good guide to the experience of the child, since the older listener has trained his perceptual system to automatically not discriminate those sounds in L2 that do not exist in L1. On the other hand, there may be a second reason why adult listeners struggle to perceive L2 differences – that they are not yet sensitive to significant aspects of the signal – and this problem may be shared with younger learners of L1.

Another shared problem may be with identifying distinct short-lived elements in a speech string. Warren (1999:172) summarises experiments with arbitrary sounds

(hisses, buzzes, etc, making up 'words') which showed that while sequences of these which minimally differ by one segment can be discriminated (they form 'temporal compounds' similar to syllables), their component elements are too short-lived to be individually identified.

I have not expressed a satisfactory argument here with respect to whether or not there is a problem of perception of this kind for a child learning pronunciation within the conventional paradigm. Nevertheless, this seems at least possible[107].

If a child must recognise words, can he also retrieve veridical images?

There is another contrast in the experiences of younger and older language learners. An adult who wishes to reproduce a word containing at least one speech sound that is unfamiliar can ask for that sound to be repeated, either in isolation or as part of the word. In other words, he can engineer a good instance of the sound. Knowing what is to come, he can set himself into AS mode and do his best to retrieve a veridical, echoic image to make a copy from, in the belief that this will be helpful. (See the longer discussion of learners of L2 in section 10.2.2.)

Faced with a similar task a child does not seem to have the power to arrange matters in the same way. If not, then he cannot dispense with the step of recognising the word. He must identify what is said as the required target, and this is achieved via MP (operating, perhaps on the whole word). This must happen before he can, perhaps, use AS to attend to the (decaying) effects the word or sound has had on his auditory system.

So because meaning must mediate the process, the capture of a sound or word for copying from (as conventionally assumed) must be either considerably harder for an infant than for of an adult, or even not possible at all[108].

---

[107] Many people have puzzled over the nature of perception, mainly in relation to vision. I would like to acknowledge Harris (1999) as an energising discussion, and Neisser's (1976:20) aphorism, "[W]e can only see what we know how to look for," as a source of considerable, though not ultimately productive, thought.

[108] Against this, people have pointed out to me that a child might have various strategies that could engineer repetitions by adults. Saying "No!", for example, might be productive on some occasions. It is also possible that only very few good instances are required by a child, and that one way or another these will present themselves at times when he knows he can attend to them with AS and put aside the issue of recognition.

Note that we can recognise more than words in MP mode. We can identify phonetic categories (speech sounds) down to the level of phonemes, so if appropriate categories exist then the phonological string for a word can be retrieved this way. However, MP does not deliver a veridical phonetic image, the effect that the sound has on the speaker. So for copying the characteristics of a speech sound it is not helpful[109].

If this argument is correct, then the speech signal is opaque rather than transparent for the purpose of improving the pronunciation of speech sounds.

An evocation instead of the capture of a contiguous model

So far I have been making a case against the idea that young children can capture an image in echoic memory which they can use to guide their production (in the way that older learners are often taught foreign language pronunciation). An alternative is that mental activity of some kind (evocation) generates such an image, perhaps created from inner speech, perhaps from stored representations[110], from anticipatory perceptual schemata or from some other source.

Various objections can be made to this idea. Some authors have questioned whether a young child could identify a suitable prototype. So, in arguing against a single phonological lexicon, Straight (1980:57) makes the point that the templates that are developed for word recognition might not be precise enough to guide production:

> "... what of the auditory feedback expectations ...? Could these, at least, represent a unity of perceptual and productional units, just as Sapir suggested? To some extent, yes, perhaps: But it seems that **the auditory target that might guide production would have to be far more specific and constraining than is the auditory representation that underlies perception.** The range of sensitivity and tenability that characterizes our abilities to recognize speech accurately despite noise, dialect differences, speaker idiosyncrasies, and so on, requires a set of flexible categories for vowel perception

---

[109] For what it is worth, I will add my own experience of how AS and MP operate. I have some knowledge of French. When I listen to it spoken on the radio, indexical information about the speaker (dialect, gender, age, etc) seems to present itself quite independently of the message being uttered. Putting that to one side, if I am listening to the stream of speech for its overall meaning then individual words are lost very quickly. However, if I choose a word to attend to then there seems to be just sufficient echo of it that if I quickly move my attention there – losing the meaning of the continuing speech stream – I can retrieve the sound categories that make it up.
Contemplating them, they seem to have a prototypical pronunciation that I have associated with each category in the past. I don't feel that I have retrieved how the speaker actually pronounced each sound on this particular occasion. I don't feel, for example, that if I heard a series of vowel productions after hearing the vowels in an utterance that I could pick out ones that had been extracted from a recording of the original speaker (except on the basis of the indexical information mentioned earlier).

[110] As with songs learnt by birds (Beecher and Burt 2004:224; Moore 2004:315).

> unlike the set needed to account for our demonstrably very stable idiolectal vowel-production patterns."

Locke (1983:84) makes a similar point, asking how a young child could extract a model for his pronunciation from the variety of sounds he is actually presented with. He quotes Peters (1974:97) on the nature of what could be derived from adult speech:

> "... what is this? The last instance he heard? An average of the last five instances he heard? (Whether slow, fast, kidding, angry, sloppy). What if the mother and the father have different "averages". ... If a child is aiming at a constant target then why do we observe that [his] pronunciation of a specific word may improve right after an adult has pronounced the word - and then revert to [his] current level of accuracy?"

To enlarge on this objection, consider an analogy with handwriting. I learnt to write Roman script as a child and a few Chinese characters as an adult learning Japanese. In both cases I could recognise letters/characters before producing them, but when I came to writing I needed to have a model present which I could refer to. I can clearly remember the adult experience of repeatedly moving my eyes back and forth to check that I was reproducing the appropriate detail in even simple *kanji*. What I had used for recognition was certainly not adequate on its own to guide my production.

If this analogy holds for speech sounds then the child would need to create the image for a sound or word to act as a standard of comparison for his production. Where could this model come from?

One candidate is inner speech, which certainly generates an experience which resembles hearing in some ways[111]. However, at least two arguments against this would have to be overcome. First, in a careful analysis of the phenomenon, MacKay (1992) reasons that inner speech is not auditory imaging but requires a separate description; he calls it 'speech imaging'. The experience it generates is circumscribed in a number of ways, including by the speaker's ability to articulate speech sounds. It is unable, then, to generate something that is novel to the speaker.

---

[111] Studdert-Kennedy (1987) quotes Welty (1983:12) on inner speech: "Ever since I... started to read . . . there has never been a line that I didn't hear. As my eyes followed the sentence, a voice was saying it silently to me. It isn't my mother's voice, or the voice of any person I can identify, certainly not my own. It is human, and it is inwardly that I listen to it. "
See Linell (1982) for a discussion of the phenomenology.

Secondly, it is not clear that infants have inner speech. Vygotsky described silent, inner speech as evolving from private speech at a much later time in development. Flavell et al. (1997) found that 4-year-olds do not report knowledge of inner speech, and the task which led the researchers to conclude that – notwithstanding this – the children must have been silently verbalizing could, it seems to me, have been performed with a motoric form of representation[112]. Another line of evidence for early inner speech had come from the performance of young children in memory tasks. This seemed to require silent rehearsal but now this conclusion seems less compelling (Henry et al. 2000).

If inner speech does not generate a model to replace an echoic image for the copying task, then might auditory imaging be an alternative candidate? A key requirement would be that the mental image contains detail that is novel with respect to the child's existing production, so that he can notice discrepancies between his output and the model.

Imaging is a contentious subject (see Harris (1999) for a review), and the study of auditory imagery is not well developed (Neath and Surprenant 2003:261). However, there is some evidence against an image resembling raw sensory input as opposed to a percept (Wilson 2004), both in the visual and auditory domains (Reisberg et al. 1989). If, "images contain what the imager put there," (Chambers and Reisberg 1985) then the child would have to come to production with a sense of what is perceptually significant for production as well as what enables him to recognise sounds, which may not be the same (and certainly wasn't in the case of my handwriting example).

A further concern with an imaging account is raised by Faw (1997), who describes a subset of the adult population who are "non-imagers". Do we have to suppose that this group learnt to pronounce in a different way from other children, or that they have lost an ability they had as young children? (However, Thomas (2001) questions the existence of such a group.)

Against these arguments, if speech sound categories are not like Straight's 'templates' but are instead based on prototypes, then the evocation of such a token might, conceivably, contain <u>all</u> the detail a child would need to evaluate his own output. As

---

[112] As I will describe later, a number of authors have argued for a motoric representation for the earliest forms of speech; perhaps this is true for longer than we have imagined. The main reason for imagining a change has, I suspect, been the assumption that the qualities of speech sounds in production are copied from auditory models, which I am arguing is not necessarily the case.

mentioned earlier, Kuhl (2000:11854) describes speech sound categories this way, suggesting that starting early in life infants develop representations of speech sounds as prototypes for perceptual purposes, and that these guide the development of their utterances when this latter function is required.

Without knowing more about how sounds are identified in childhood, it doesn't seem possible to resolve these issues at present. Children at a later stage may fail to discriminate some minimal pair words and their perception clearly develops in sophistication over a long period (e.g. Simon and Fourcin 1978). I would be surprised, therefore, if their recognition criteria are so highly developed at an early age that they can adequately guide their own production, but this may be wrong.

13.4.2 Capturing intrinsic sound(s)

A second ability required for self supervised learning is that the speaker should be able to capture his own sounds. A hearing infant must have some ability to do this, but in this subsection I ask about whether the extent to which this is true is likely to be sufficient to allow an adequate comparison with target speech sounds or a set of target criteria.

A copying account would require that a young child be capable of hearing fine detail in his own output, between the qualities, for example, that distinguish the vowels in *cat* and *cut*. However, the relationship of older speakers with their own voices provides some *prima facie* evidence against an ability to hear oneself accurately. For example: most people are shocked by how they sound the first time they hear recordings of themselves (as an 8-year-old I was horrified); people sing flat despite being able to recognise when others do this; speak with unpleasant voice qualities despite being able to recognise this in others; and adults, it seems, can't even mimic themselves (Vallabha and Tuller 2004).

There are at least two ways that these phenomena might be the result of not hearing ourselves well, despite the impression we may have that we do. I start with ideas developed from first principles by Howell which led to him questioning whether singers could control their voices by hearing. Then I will consider evidence independent from this which also suggests that we 'hear' something different from what we say, generated in this case from the expectations we have of our output.

Can a child hear himself adequately (1)? Howell's proposals on bone conduction

Howell's views are now part of a wider proposal he has made about timekeeping in speech. The part relevant to my argument is summarised in Howell and Sackin (2002), who describe some of the problems identified by Howell and his colleagues in the 1980's when they questioned the view that auditory feedback is sufficient on its own to control speech:

> "The auditory system would need to supply the speaker with a veridical record of what was produced; otherwise, establishing if and what error has occurred with the intention of correcting it would not be possible. However, it is not clear that the representation of articulatory output provided by the auditory system is veridical of the intended message. This is because the auditory representation the speaker receives while speaking is affected by internal noise. The noise that is present then affects the information that can be recovered from the acoustic output. The main source of internal noise originates in vibrations of the articulatory structures that are transmitted to the cochlea through bone. This bone-conducted sound is delivered to the cochlea at about the same time as the acoustic output from the vocal tract. Bone-conducted sound during vocalization is loud enough to make its effects significant. Von Békésy (1960), for instance, estimated that bone- and air-conducted components are at approximately the same level. The airborne sound contains sufficient information to decode a speaker's intention (other people listening to the speech understand the message). The bone-conducted sound, on the other hand, is dominated by the voice fundamental; formant structure is heavily attenuated and resonances of body structures extraneous to vocalization (such as the skull) affect this component (Howell and Powell 1984). Consequently, the bone-conducted sound contains limited information about articulation. **The degraded bone-conducted sound would also mask out the formant information in the air-conducted sound**. Such masking would reduce the ability of a speaker to retrieve information about the articulation from the air-conducted feedback." (p.2842)

While Howell and Sackin caution that, "this argument relies heavily on the evidence presented by Howell and Powell (1984)," they point out that, "if future work confirms that the auditory feedback signal is restricted in the information it provides about articulation, models that assume feedback is used to compute a precise correction needed to obviate an error will need revision." As I described earlier, the models of speech sound development in young children that assume similarity based judgments of equivalence all make this assumption.

Howell (1985) describes in more detail the mechanisms which lead to singers (and speakers) not hearing the same signal as their listeners, and presents data which illustrates the effects involved.

With respect to 'bone-conducted' sound (the term conventionally refers to sound conducted by all body tissues) he points out that transmission of the vibrations of the air

in the vocal tract via body tissue will have a minimal auditory effect (except possibly indirectly by causing the cheeks to flap). Forced vibration associated with movement of the vocal folds is of much greater importance; this may have the same period as vibration of the airflow but is unlikely to have the same shape (and will not, of course, be filtered to reflect the resonances of the oral cavity). An additional contribution to bone-conducted vibration will come from the resonances of the skull and other structures involved (which are affected by the state of contraction of the muscles around them).

Howell (1985) describes further distortions likely to be introduced by the transmission path[113] and then presents data from one subject showing measurements made with an accelerometer secured on the mastoid bone. These show low frequencies over-represented in the bone-conducted signal compared to the airborne one, a reduction in amplitude of vowel formants[114], and the presence of what appears to be skull resonances at frequencies between 1 and 2 kHz which will affect the intelligibility of sounds through confusion with the formants.

Howell goes on to discuss the distortions that arise from the effects of middle-ear muscle activity and the external path taken by the airborne component of self-produced sound. He concludes that auditory feedback is not likely to be used in the way conventionally imagined for vocal control in singers.

Howell's experimental results are supported by Pörschmann (2000) who used masking with adult subjects to psychoacoustically evaluate the 'bone conduction' pathway of self-produced sound, comparing unvoiced /S/ and voiced /z/. He concluded that, "For frequencies between 700 Hz and 1200 Hz bone conduction dominates the perception of a person's own voice."

His more detailed findings were that,

> "When compared to air conduction, the perception of bone-conducted sound has the greatest influence at about 1 kHz. It can be observed that between 700 Hz and 1200 Hz

---

[113] It seems unlikely that signals from different vowels will be distorted in a similar way. Thus I think that Vallabha and Tuller's (2004) rejection of bone-conduction effects to explain their results was premature.
[114] Howell's graphs of frequency spectra (p.279) seem to me to portray a greater degree of distortion for the vowel formants than he describes.

> bone conduction is the dominant pathway, for the other frequencies air-conducted sound dominates in the perception process of a person's own voice." (p.1044)

Up to around 1200 Hz the voiced sound had a greater masking effect than the unvoiced one. However, his bone-conduction transfer function shows that, "in a frequency range between 1.2 kHz and 3 kHz ... the decrease in the amplitude response of the transfer function towards higher frequencies [for voiced phones] is steeper than for the unvoiced phones," with the effect of the unvoiced phones not, in fact, decreasing significantly. Pörschmann explained that this effect, "might be caused by the different excitation points and the resultant differing pathways of the voiced and unvoiced phones."

In addition,

> "Resonances of [the bone conduction transfer function were observed] at about 900 Hz, 1800 Hz, and 3600 Hz, which are all multiples of 900 Hz. The resonances could be interpreted as inherent to those of the human skull." (p.1042)

Finally, Shuster and her colleagues have based therapeutic work on an assumption of separate input and output phonological lexicons (which I will discuss later). In the course of probing their underlying assumptions, Shuster (1998) found that the self-perception of her subjects' speech was, indeed, different from their perception of others' speech, but not in a clear cut fashion. She identified a possible cause to be that the subjects had been presented back their recorded speech via headphones, without, therefore, the proprioceptive feedback they would normally receive together with any distortions that bone-conduction would create. Shuster and Durrant (2003) investigated the potential effects of this, but while their data are consistent with other research on bone-conducted sound, their novel approach did not generate a transfer function.

It is difficult to know how to relate all these results to infants. The first formants of infant vowels are higher than those of adult speakers, so are potentially more vulnerable to the distorting effects described. However, the characteristics of the infant body will certainly give it a different vibratory profile to that of the subjects investigated by Howell and Pörschmann. Potentially, though, the conclusion in Howell (1985) may be extendable to infants: they may not be able to use auditory feedback alone for vocal control, at least for some sound types, particularly voiced ones.

If this is the case to any significant degree, then there would be no need to consider the other arguments in this chapter for why at least vocoid speech sounds are not copied on the basis of signal similarity. It would not be possible.

On the other hand, the effect may only handicap such a mechanism, it might not be fatal to it.

Before moving on, let me raise one objection to my use of Howell's proposals: if the sound of the infant's own voice is heard by him as highly distorted in an important frequency band, then how can we explain pure echolalia, progressive idioms and other examples of mimicry?

An answer to this might make use of some of the following:

- Presumably bone-conduction affects perception of some of the cues for speech sound identification more than others (for vowels, for example, their steady state formants rather than the formant transitions between preceding and following consonants). So infants may be able to perceive features of their own output that enable them to judge their tracking of the overall shape of a phrase, but not those features that would enable them to judge the similarity of particular segments to those of adults.
- The match we perceive in progressive phonological idioms, for example, may not be as close as we imagine with respect to individual segments. Reports of infant mimicry are based on the outcome of infant speech, not their output. Adults may think they hear more than is actually there. Speakers generally 'sympathetically reconstruct' what they hear (Linell 1982; Nathan 1997), even to the extent of restoring phonemes that have been removed (Warren 1999), and top down expectations are a powerful factor in speech perception.
- Additionally, matching on one basis might unintentionally but fortuitously generate a match on others, because an infant shares his basic body morphology with adults.

Can a child hear himself adequately (2)? The effect of forward models of speech

Putting to one side the potential problem Howell has drawn attention to, we can also ask whether a child normally hears himself when he is talking. As adults we have an auditory experience of ourselves during speech, but phenomena such as inner speech

and the McGurk effect demonstrate that such experiences need not be generated by hearing the side tone.

It seems that this question cannot be answered for adults yet, much less for children. (Postma (2000) reviews experimental data, but his focus is on error detection rather than on the evaluation of sound qualities. Jones and Munhall (2000; 2003) and Houde and Jordan (2002) report experiments showing some effects of auditory feedback on production, but primarily with respect to pitch or loudness, or in ecologically unrealistic situations of whisper and listening through headphones.)

There are suggestions that in normal circumstances our primary percept of the sensory consequences of what we do is based on the output of forward models rather than intrinsic sensory input. Thus Blakemore et al. (2002) describe how the motor system might operate:

> "... the forward model predicts the sensory consequences of movement and compares this with the actual feedback – this comparison occurs after a movement is made. This prediction can be used to anticipate and compensate for the sensory effects of movement, attenuating the component that is due to self-movement from that due to changes in the outside world. The results of several studies suggest **that this prediction, which is based largely on the efference copy of the motor command, is available to awareness** ... The experiments ... also suggest **that the actual state of the motor system and the actual sensory consequences of a movement are normally unavailable to awareness.** Furthermore, we seem to be unaware of the results of the comparison between the predicted and intended outcome of motor commands, and the comparison between the predicted and actual sensory feedback, as long as the desired state is successfully achieved."
>
> "We propose that there is only limited awareness of the actual state of the motor system whenever it has been successfully predicted in advance. We suggest that **under normal circumstances we are aware only of the predicted consequences of movements.**"[115]

Wienen and Kolk (2005:298) support the application of these ideas to speech.

With respect to signed speech, a synopsis of Emmorey (2005) reports evidence that visual feedback from one's own signing appears only in peripheral vision. Signers, it seems, do not look at their hands; they monitor, instead, their internal representations of their signs.

---

[115] This conclusion was partially reached based on experiments undertaken by this group to explain why we cannot tickle ourselves.

Speakers may, then, be running an emulator/forward model which generates their auditory experience during self-produced speech. Levelt et al. (1999) suggest that the speech comprehension system parses the output of the speech planning process and this might be expected to generate an auditory experience which, being based ultimately on the phonological score (Levelt 1999:112), could differ from the side tone. If so, then this mechanism is presumably the same one that generates the neutral correctness of inner speech.

I feel I get a sense of a forward model in action from prolonging a vowel and then changing the mental basis of my actions from production of a phonetic entity to the maintenance of a given articulation. As the /i/-ness or /a/-ness of the sound fades what remains for me is a rather unfamiliar noise. This, however, is the noise that I would have to use for a comparison with a vowel that I was trying to learn to say.

Further support for a forward model mechanism come from Heinks-Maldonado et al. (2005) who report evidence for selective suppression of auditory cortex activity to self-produced speech.

If the suggestions above are broadly correct then a speaker is not typically having an experience while speaking that can be used to make a judgment of acoustic similarity. We are familiar with making similarity judgments when we can 'step back' from what we are doing, but this is only possible when the effects of our actions persist (e.g. with handwriting or a drawing). Speech only exists while it is being performed. Perhaps one reason why we need singing teachers, tennis coaches etc for skills whose effects are similarly ephemeral is that we cannot attend to our actions or their effects objectively (with AS) while we are simultaneously engaged during production with the intention underlying them.[116]

There must be (and are, of course) circumstances when we do hear something of our own output. It would seem that this has to be true when we are doing something new, since then, by definition, there is no forward model to supply a sensory expectation of the results. So when an existing inverse model is being driven, we may experience what

---

[116] In a language class, I can usually tell when a fellow student's pronunciation of a vowel, say, is incorrect, but I am often surprised by the judgment on my pronunciation when I think that I have done what is appropriate and have 'heard' it to be acceptable. Perhaps a child has a similar reaction when he is told he is saying 'fis' for 'fish'.

we expect; but when we create an articulation that is consciously new, no surrogate input is created and any suppression mechanism would be switched off. Thus while a forward model might disrupt speech sound learning to some extent under a copying account, it should not affect 'discovery' approaches (as described in the next chapter) where speakers consciously experiment with their own articulation.

In summary of this subsection, there may be two further reasons why the speech signal is opaque for the purpose of learning the pronunciation of speech sounds rather than being transparent. It may be masked by bone conducted vibration (presumably affecting the perception of vocoids more than that of voiced contoids, and the perception of voiceless contoids very little), and it may be replaced by the output of a forward model of speech that will reflect the speaker's intended rather than his acoustic output.

13.4.3 Comparing two auditory images*

(At this point, I am not in a position to do more than note some other issues for any similarity based equivalence account of speech sound development. Within these areas there may be further reasons for the speech signal being effectively opaque for learning to pronounce.)

In the third stage of a self-supervised learning process the child compares his attempt with the model, noticing similarity and differences. This raises issues of normalisation, how he learns criteria to judge similarity, what information the non-recognition of category membership furnishes, and the extent to which 'side-by-side' comparisons are possible between veridical sound images.

Normalisation

Kent (1981:167) describes the normalisation problem:

> "To learn speech, the child must ... establish equivalence classes between his or her acoustic patterns and those of the adult models. Equivalence classes have to be formed because the child's short vocal tract cannot produce exact acoustic replicas of adult speech sounds."

Most accounts assume that equivalence classes must be based on judgments of similarity made by the child. How are the criteria for such judgments developed, if the acoustic signals are in fact objectively and discriminably dissimilar (in at least some of their aspects)?

Johnson (2005) reviews pattern matching/translation theories, none of which seem satisfactory to him. On the other hand, Lieberman (1984:221) described an infant acting as if he, "were equipped with an innately determined neural device that enables him to effect the process of vocal tract normalization."[117]

Unless Lieberman is correct, normalisation remains mysterious, but its result is that a given speech sound produced in different circumstances is 'correctly' identified as being the same by a listener, despite the tokens being discriminably different.

Developmentally, normalising skills for perception seem to be developed early. The experiments of Kuhl, Hillenbrand and others in the late 1970's and 1980's demonstrated that infants of around 6 months could categorise signals that adults identify with speech sounds successfully across talkers of different ages, sex, and clarity of speech, for vowels that were sometimes quite similar, and for consonants varying in only one feature of place or manner. See Kuhl (1987, 1991) for reviews; and developments of her work by Bower and colleagues, e.g. Aldridge et al. (2001).

Under Kuhl's account, the infant's skill in judging perceptual constancy across others' speech is applied to his own, as it must be in all copying (similarity-based equivalence) accounts. This seems problematic to me. The processing of intrinsic input seems likely to be very different from the processing of extrinsic input. What pre-attentive processes are applied in each case? What is filtered or suppressed? What is the effect of expectation?

Similarity

Most accounts describe the child forming equivalence classes through judgments of similarity. I will argue below that a notion of similarity between his own and others' output might emerge later, out of the equivalence classes the child forms rather than as the means by which these are formed.

[117] Donegan (2002:6) points out that the evidence of perceptual constancy in infants assembled by Kuhl and her colleagues might support Lieberman's proposal.

Recognition and error signals

Judging that a token is not a member of a particular category does not, of itself, furnish information about why this is so. I can know that there is "something different" about another person without realising that it is because he or she has changed hairstyle, glasses or some other detail of their appearance. A page printed in two slightly different typefaces can be recognisably different without an understanding of the source of the changed appearance.

To make a comparison across a time interval that will furnish information about changes (and that will therefore inform attempts to correct imitation attempts) we must be able to evoke the original situation. To do this, we need to have marked the details that will be re-created, with the result captured either as an individual awareness or as part of a theory (a redescription of a group of awarenesses).

Holding two echoic images in memory at the same time

Returning to the example of learning to write letters and Chinese characters, it is clear that the normal copying process is supported by the persistence of the results. I could move my attention back and forth between my attempt and the model because both existed simultaneously. This allowed me to examine detail, to become aware of discrepancies which were not apparent at first glance.

Is such a 'side by side' comparison possible with sounds? It is, if one 'sound' has become a category (which can be represented by a symbol). The other can then be compared against the category characteristics. But if both sounds must exist as veridical images (possibly normalised in some way) then my personal experience is that producing a sound after hearing a model destroys any echoic image I have of the model. As an analogy, it is hard to imagine doing a 'spot the difference' exercise on two pictures that are displayed not contiguously but in succession; and for only a short period of time.

13.5 Two further arguments against copying

13.5.1 Pattern of learning seen in practice*

The pattern of learning which we actually see in children may tell against copying accounts. Menn et al. (1993:430) point out that:

> "Error correction patterns and output variation patterns [in child speech] correspond to those of reinforcement learning, not supervised learning [or 'copying']. **With supervised learning** (e.g. back propagation of sound error through a forward model) **we should see at least some tendency to correct erroneous forms**. The more often the model makes erroneous pronunciations, the more error correction should occur. This is not what happens. Erroneous pronunciation rules are usually very stable. Daniel's velar harmony rule lasted for nearly a year. Instead, **error correction occurs only in response to differential reinforcement**."

13.5.2 How complex motor skills are learnt, other than by copying*

Consider general motor skill learning in children (i.e. their non-linguistic motor behaviour). Play includes many activities where motor skills previously learnt are deployed in new ways (using a toy telephone, for example), providing us with examples of the learning of sequences of actions by imitation. This has also been the case in recent experiments which have investigated the goal-directiveness of children's copying behaviours (using a rake, turning a light switch on and off etc).

However, if we examine fundamental motor skills such as reaching, sitting, crawling and walking, then does anyone suggest that these are acquired by imitation? Superficially it seems possible that they could be: in the same way that a child is surrounded by adults modelling speech sounds he is surrounded by examples of at least three of these activities. Yet infants who have never seen anyone crawl manage to acquire this skill, and as far as I am aware there are no serious suggestions that the other three skills are not learnt by a similar 'discovery' process of the child exploring his own effectivities rather than observing and trying to copy how adults do these things. Perhaps adult behaviour inspires early attempts but modelling does not seem to play any significant role beyond this.

Discussing these types of actions, Gattegno (1973:50) says,

> "... the process is entirely an inner one. Only when a baby knows how to perform certain acts can he find that a similar act is being performed by others – similar, not the same, for the action he sees uses another soma [body] and is perceived from outside while he knows his own action from inside."

The question of whether fundamental motor skills are learnt by imitation is not central to this thesis, and I may well be ignorant of a body of literature that discusses this. It seems to me, though, that when an activity is very complex a child will not be able to

mark aspects of it through observation and it will then not be imitated. The way that locomotion operates seems to be a good example of something far too complex for a young child to divine from surface observation, but discoverable from within. Once discovered it can be recognised in others, and then walking styles imitated. But this is not a skill that, to begin with, can be copied by a child.

With respect to speech, it may have been imagined that babbling gives a child a rich understanding of the relationship between sound and articulation. However, MacNeilage and Davis have drawn a picture of babbling as highly constrained motor activity, without the segmental independence it appears on the surface to show (see the discussion of their work in section 10.4.3).

If this is right, then it may be that learning to speak is more similar to learning to walk and all other basic motor skills than we have imagined. Imitation (re-enactment) may not be possible until a great deal more has been discovered by other means than was gleaned through babbling.

13.6 Summary

In this chapter I had two broad goals. Firstly, to develop a theoretical framework for thinking about how speech sounds might be learnt if this is achieved by copying in any sense of the word. Secondly, to use that framework to describe some difficulties that may face the mainstream, 'common-sense' view of how children learn to pronounce sounds.

With respect to the first, I hope I have justified a sharp conceptual separation of mimicry and purposive copying (which, in the speech context, is represented by re-enactment). I based this upon them being the result of us attending to the output from two modes of perception of sound, AS and MP, so that mimicry is not just imitation without a purpose, but a more fundamentally different process.

Mimicry is based on resemblance between the forms of actions or their results. It is used by children, but it is too demanding of attentional resource to provide a basis for speech. Re-enactment is based on equivalent effect and attention to acts. Some confusion arises between them because of the supporting role that resemblance can play in learning

episodes for re-enactment, and because of the supporting role that recognition of actions as acts can play in structuring complex mimicry.

In line with this separation, my attempt at the second goal described potential difficulties with the learning of speech sounds from two perspectives: as if they are mimicked, and as if they are learnt in side episodes of learning but based on a judgment of similarity made by the child. I concluded that the first route was theoretically impossible: tracking an ephemeral target is incompatible with learning new motor skills. With respect to the second, I raised some issues that demand further attention. At present I cannot demonstrate that any are fatal to the conventional account although some have the potential to be[118].

Many accounts suppose that it is the child who makes a judgment of (normalised) sound similarity in some form to guide his learning. For this, the vocalisations of adults and children would have to be perceptually 'transparent' (a child would have to perceive inputs from listening to himself in a similar fashion as those he perceives from listening to others). A number of the arguments in this chapter suggest, with greater and lesser degrees of plausibility, that speech sounds produced by himself and others are actually opaque rather than being transparent to a young child. The arguments with greatest potential seem to me to be:

- that the need to recognise words in MP mode prevents him from retrieving a veridical sound image from AS processing;
- that bone-conduction makes the evaluation of the qualities of at least his own vocoid production impossible;
- that experiencing his voice as the output of a forward model rather than the sidetone (hearing his intention rather than his actual output) similarly precludes judging his own output;
- that babbling may not provide a developed inverse model for sound production, in which case the complexity of speech suggests that it might be broached by a process of discovery rather than imitation, as are other complex motor skills.

[118] Of course, there are other accounts that I described in the previous chapter. Another task for the future would be to determine which of the potential difficulties for the mainstream account that I have identified would also be difficulties for these.

If there is no innate imitative mechanism which can accommodate this, then how is the correspondence problem (learning to imitate) solved? As discussed in section 12.2, a generalist account of imitation (like Heyes' ASL) allows there to be a mechanism such as an environment which contains an imitative social partner – a mirror - to achieve this. The evidence for the working of such an environment for speech sounds follows in the next chapter.

I can now add a final diagram, showing what happens to learners if neither mimicry nor re-enactment of sounds enables them to solve the correspondence problem.

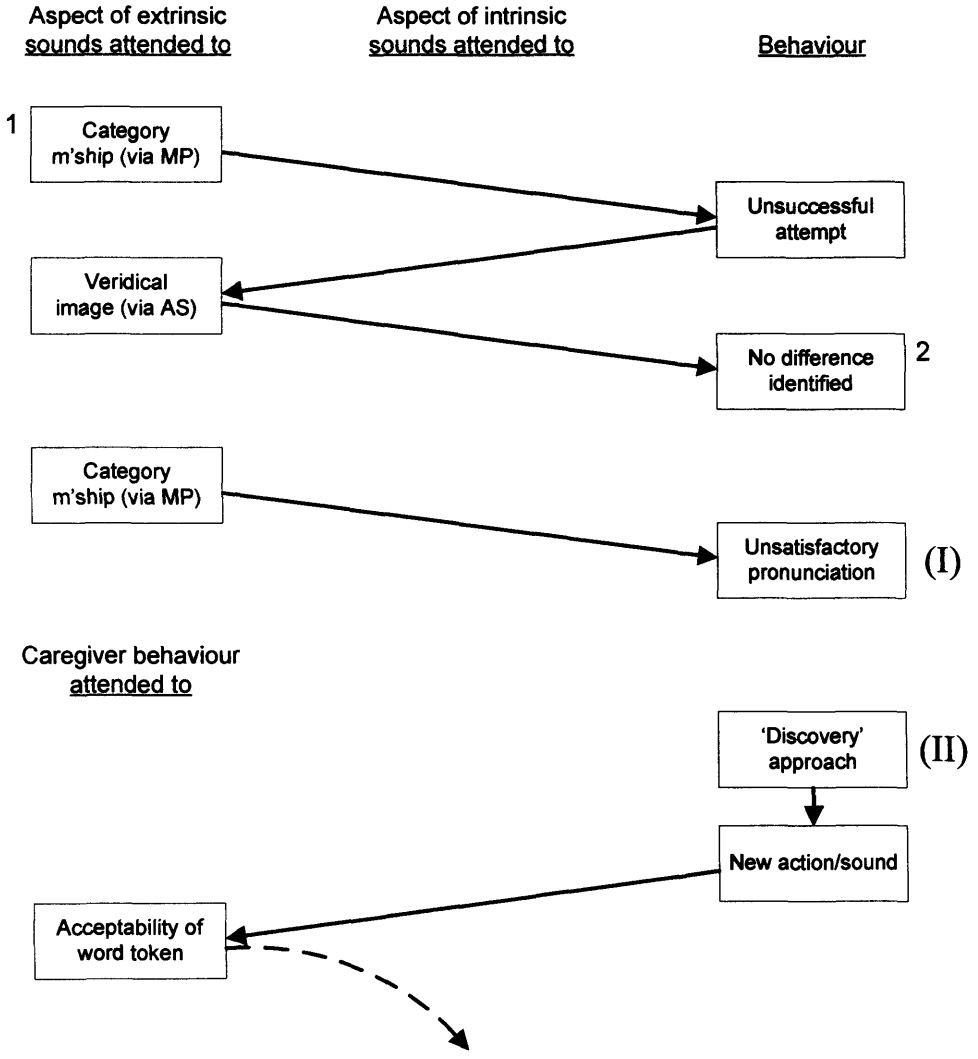

Figure 13-5. If the learner is unable to 'copy' a speech sound ... e.g., if 'No difference identified' in box 2 because of (i) categorising perception in an adult learner (effect of L1); and/or (ii) failure of identification during short presentation; and/or (iii) insensitivity to the characteristics needed for identification:

(I) he makes no further attempt to improve his pronunciation, persisting with a sound correspondence from his existing production repertoire and disregarding the evidence that it is unsatisfactory; or
(II) he adopts a 'discovery' approach (subverting any implicit instructions to mimic the model) in order to solve the correspondence problem.

At box 2, the difficulties identified in this chapter block progress. Many foreign language learners now take path (I): they stay with what they know how to do, aware, however, that it is unsatisfactory. (Some young children may, for a time, do something similar.) Path (II) would provide a way out of their predicament: to guide their production they switch from paying primary attention to what the environment produces, to paying attention to how the environment responds to their attempts at production.

For a child, in other words, what he first learns from adults is that saying words is useful. He can exploit this through mimicry of whole word shapes, but as the cognitive effort this entails becomes increasingly limiting he then learns not how to say the sounds of speech, but that production using the sounds of speech is useful. With the help he gets from his social environment he is able to construct the necessary movements to reproduce speech sounds for himself, using his success in using them (as determined by others) as his guide.

14 Learning to imitate: creating sound to movement correspondences

> "[A]t least one mother in our sample was aware, on reflection, that she did unconsciously imitate her child and that this was important. In answer to the question, 'Can you think of any example of when N imitates you?' she replied:
>
> > 'Well, it's a very complicated thing this, because I've found out that she ... I mean sounds, for instance, let's restrict ourselves to sounds – she suddenly discovers a sound and it rather fascinates her and I reinforce it; I make the sound as well. And that tends to make her want to do it more. And then perhaps on a different occasion when she's forgotten all about it, if I make that sound she will imitate it. But the sound seems to have to come from her in the first place.'
>
> Obviously one cannot rely on an insightful comment by one single mother to confirm a theoretical view about how imitation normally originates."
>
> S. Pawlby, *Imitative interaction* (1977:222)

> "Solution of the problem [of motor equivalence] is a pressing issue in general research on motor control. For speech ... we have an added twist: the arbiter of equivalence is not some effect on the external world - seizing prey, peeling fruit, closing a door - but a listener's judgment."
>
> M. Studdert-Kennedy, *The phoneme as a perceptuomotor structure* (1987:70)

How does a child acquire an inventory of speech sounds with qualities that match the pronunciation of the speakers around him? The widespread lay and scientific assumption is that he copies (or mimics) what he hears. This mechanism is superficially plausible but may be problematic for the reasons I discussed in the previous chapter.

Speech sounds are learnt as part of saying words. In this chapter I will propose that the spoken aspect of word adoption may start with the child copying or mimicking word forms using his auditory inverse model (IM), but that it becomes smooth and efficient when he has developed a specialised IM which links a speech sound he hears to the movements he needs to make to reproduce its equivalent in his voice. Unlike the auditory IM, the development of the speech sound IM (as I call it) depends upon social interaction: the mirroring of the child's speech behaviour by his mother and others in a variety of ways.

This process happens in parallel with production via the auditory IM, starting pre-linguistically with the 'imitation' (or reformulation) of a child's utterances by

caregivers. What the child learns from such interactions later enables him to discover that he has already developed some articulations that are equivalent to the speech sounds which are used communicatively by adults. His ability, therefore, to imitate at least some speech sounds (the correspondence problem solved) means that he can start to learn word forms by imitation. Reinforcement helps him to refine his output and to develop new speech sounds as required in a generalisation of this 'discovery' style of development.

At the end of this chapter I show that in analogous situations the mechanisms underlying this account enable vocal and other learning in humans and animals, young and old. I then review other accounts in the light of my proposals.

14.1 The entry into speech sounds

I have described some possible problems with the conventional and other accounts of a child learning sound qualities on the basis of his own judgments of similarity. But how else might an infant solve the correspondence problem for speech (i.e. create the 'vertical' links in figure 12-1)?

For a different conceptual approach, consider walking, a skill learnt during the early period of speech acquisition, and notice that, as with speech, there are models of this behaviour in the infant's environment that he could, in theory, imitate. Yet no one suggests that a child copies what he sees of walking (after some process of normalisation to his body size). No one imagines him thinking (non-verbally, but in a way that if verbalised would equate to), "Daddy lifts his knee, swings his lower leg forward, flexes his foot at the ankle ... now I'll do those things."

Instead the child discovers how to walk for himself, albeit with adult encouragement and assistance. At most he takes from the adult models only the idea or inspiration that walking is possible. Indeed, the facts that (i) blind babies learn to walk and (ii) crawling, an earlier form of locomotion, may well be learnt by a first-born in the absence of any model, together suggest that even this role for a model – providing a goal to be emulated – might be minimal.

After teaching himself to walk, a child may find that he can refine or supplement what he does by observation of other people: to mimic them in play, for example, or to copy

their actions in a dance class. By then, though, the basic correspondence problem for these skills will be largely resolved, partly because the child shares an overall body morphology with other humans. He will discover that much of what he has taught himself to do is already similar to what others do, because he is working with muscles, bones, etc that are configured more or less as they are in others. So, many aspects of learning to dance, the sequencing and integration of movements for example, may fairly be described as 'learning by imitation' (creating the horizontal links of figure 12-1). Refinement of existing movements, some new movements, improved control of balance, and so on will also be required, so there will also be correspondence problems still to solve; but by now some judgments of similarity can be made by the child, rather than him relying exclusively on external evaluation (which, when originally learning to walk, was provided by basic reinforcement from the success or failure the child encountered in staying upright and moving forward, and now comes from a teacher or devices such as mirrors).

I have suggested that the speech signal may not be perceptually transparent, at least for a young learner. So if a process of discovery is to be used to explain the appearance of early speech sounds, we need to explain how external evaluation rather than a self-made judgment of similarity can inform him that the correspondence problem has been solved. How can he become informed about his performance? Heyes and Ray (2000:224) discuss this general problem. They describe various mechanisms by which an observer can form vertical associations for perceptually opaque actions. One way is through evidence of behavioural synchrony, which may result *inter alia* from imitation of the observer by the 'model' (i.e. of B by A, with 'imitation' going in the opposite direction from that normally expected). The 'model' informs the observer of what the observer has just done by performing it for him.

14.1.1 Reformulations

We might expect that in the imitative episodes of early childhood, it will be the infant who imitates his mother. In fact the opposite is usually seen.

Pawlby (1977) studied eight mother-infant dyads while the children were between 17 and 43 weeks of age. With respect to speech sounds, she found that more than 90% of the speech-related matching observed was attributable to mothers imitating their children (p.215). She asked:

> "[W]here does imitation begin? Paradoxically our study suggests that the whole process by which the infant comes to imitate his mother in a clearly intentional way is rooted in the initial readiness of the mother to imitate her infant. In other words, almost from the time of birth there seems to be a marked tendency for mothers to reflect back to their infants certain gestures which occur spontaneously within the baby's natural repertoire of activities. She appears, however, to select actions which she can endow with communicative significance, especially vocalizations, and it is these acts which the infant may first perform inadvertently or unintentionally that are automatically reflected back as if the infant had deliberately initiated them for the purpose of social exchange."

Pawlby went on to describe "striking" strategies the mothers used to create simulations of deliberate acts of imitation on the infant's part, and she commented that infants, "pay special attention (in that they laugh and smile and appear to be pleased) when the mothers themselves imitate an action which the child has just performed." At the end of her report she made the observation that I have used as an epigram for this chapter.

Papoušek and Papoušek (1989:149) summarised similar results for vocal matching in infants from 2 to 5 months of age as follows:

> "[M]aternal propensities to modify their speech, to echo infant sounds, and to provide a modelling frame significantly contribute to the high incidence of vocal matching. The question of the extent of infant participation remains."

Snow (1977), Veneziano (1988) and Kokkinaki and Vasdekis (2003) are among other studies reporting similar results: that in the majority of 'imitative' interactions it is the infant who leads and the mother who follows[119]. This sequencing of events can be coded as 'infant-mother'.

Adults change their responses as they attribute greater powers of speech to their child (e.g. Snow 1977). Stoel-Gammon (2000) discusses this with respect to the pre-linguistic stage:

> "Adults tend to respond to [babies'] vocalizations in predictable ways, and pre-canonical utterances are often imitated, which, in turn, sometimes elicits an imitation from the baby. Canonical utterances, in contrast, tend to be interpreted as attempts at meaningful speech; a simple 'ba' from the baby will elicit the response of 'ball ' or 'bottle' from the mother."

---

[119] For discussion of why caregivers might behave in this way see Newson (1979:208), who describes the theoretical and experimental paradigm which Pawlby's study was conducted within. He claims that mothers interpret their infants' reactions by a process of 'adultomorphism', crediting them with fully human powers of social responsiveness. See also Snow (1977), Pawlby (1977:221), and Locke (2001:294).

Later, when recognizable words are being produced, expectations rise again (e.g. Velleman et al. 1989:170). Otomo (2001:29) reports mothers of children between 12 and 21 months now not responding to non-word-like utterances. However, they continue to infer meaning where they can and respond contingently, almost always in linguistically correct forms. Their range of responses now reflect features of the child's utterance beyond phonetic performance, but phonetic evaluation remains part of the process of reformulation of child productions by adults until at least 4 years of age (Chouinard and Clark 2003).

It seems that infant-mother interactions should be a source of useful information, but when this is discussed, the context is usually assumed to be that of a child attempting to form similarity-based equivalences for sounds. For example Menn and Stoel-Gammon (1995:354) say:

> "Adult imitation of the child's sounds intensifies the linkage between the adult-produced and child-produced sound patterns, and because the adult's imitation will be filtered through the adult's perception and production, it should produce some feedback about the difference between the child's output and the repertory of adult phones in the given language. Meltzoff (1990) indicates that children are highly attentive to adults whom they perceive as imitating them."

I am only aware of infant-mother interactions given any great significance in four descriptions of how children learn to imitate words and produce speech sounds, those of McCune (1992:331), Lacerda (2003:51), Clark (2003) and Yoshikawa et al. (2003).

In the subsection after next, I will propose that infant-mother interactions provide all the evidence that Heyes and Ray, in the previous sub-section, said would be needed for an observer to solve the correspondence problem for an opaque signal. I.e., that the child is informed by his mother of what he has just done by her performing it for him.

This will raise questions about the incidence of reformulations, the extent to which they may be needed and their generality (particularly in the light of reports of different cultural child-rearing practices across the world, e.g. Ochs and Schieffelin (1984)). These issues are discussed by Chouinard and Clark (2003:658-660), Messer (1994:229-235) and Zukow-Goldring (1996:208). Evidence that some young children do not get any chance to be informed about their production this way would challenge my account.

I will use 'reformulation' in a more restricted sense than in the literature, where it can cover the recasting of utterances for reasons of grammar, intonation, etc. I will use it to refer to the interpretation by a mother of just her child's vocal production and her contingent response with what she regards as the L1 equivalent of his output, either in normal or 'marked' forms ('motherese'). My use differentiates reformulation from vocal mimicry (matching of the surface form).

14.1.2 Mirroring

I will use the term 'mirroring' to describe any response by his mother from which the infant can obtain information about himself and his behaviours, in contrast to his obtaining such information from direct self-perception. As various authors have noted (e.g. Pines 1985; Stern 1985:144) 'mirroring' is not an ideal term for this: a 'response' allows for information to be conveyed both by different behaviour on the part of the mother compared to that of the child and by this behaviour being subsequent to the child's, while a mirror normally reflects an exact copy of the object in front of it and does so instantaneously[120]. Nevertheless, 'mirroring' is well established in the literature with the wider interpretation I have just described.

Winnicott (1971:112) provides a description of the process in its early stages:

> "... what the baby sees [in his mother's face] is himself. In other words, the mother is looking at the baby and what she looks like is related to what she sees there."

Pines (1984:32) explores the dynamics of a mirrored interaction further:

> "It is mother who selects only certain patterns of activity to respond to in her child, thus presenting him with an image of himself through her mirroring behaviour ... The child can begin to learn who he is through attending to his mother's response to those aspects of his behaviour which make sense to her. Mother inserts meaning and intentionality into her baby's behaviour and so in this way he begins to recognize himself."

Pines (1984) and Gergely and Watson (1996:1187-1203) discuss various ways in which mirroring has been hypothesised to play developmental and therapeutic roles. Fonagy et

---

[120] In fact looking in a real mirror is usually a more complex psychological process than that of observing identity. Romanyshyn (1983:7-20) describes how the reflection we see is actually a figure in a tale rather than the person on the right side of the mirror, and how the observer, too, is transformed into a character in the tale. Understanding the experience of looking in a real mirror in this more sophisticated way supports the wider, metaphorical usage of the term 'mirroring'.

al. (2003:423) list psychoanalytic theorists and developmental psychologists who regard, "the caregiver's mirroring of the infant's subjective experience ... as a key phase in the development of the child's self," particularly with respect to affect regulation. Rochat (2001:201) describes such emotional mirroring as providing,

> "a perceptual scaffolding for the objectification of [infants'] own affects ... [by which] infants are exposed to an explicit, analyzable form of what they feel privately at an implicit level."[121]

As Rochat (2001) makes clear, in this developmental paradigm it is the origins of specific knowledge about an infant's own affective disposition that the infant is believed to develop. In time, direct self-perception of both his inner states and their outward expression will also come to be a source of affective self-knowledge.

Stern (1985:142) places imitation/mimicry at one end of a spectrum of mirroring behaviour and so-called 'affect-attunement' at the other. Here, a mother reflects back to the child her understanding of his internal state rather than his overt behaviour.

> "Affect attunement ... is the performance of behaviors that express the quality of feeling of a shared affect state without imitating the exact behavioural expression of the inner state."
>
> "The reason attunement behaviors are so important as separate phenomena is that true imitation does not permit the partners to refer to the internal state. It maintains the focus of attention upon the forms of the external behaviors. Attunement behaviors, on the other hand, recast the event and shift the focus of attention to what is behind the behavior, to the quality of feeling that is being shared. It is for the same reasons that imitation is the predominant way to teach external forms and attunement the predominant way to commune with or indicate sharing of internal states. Imitation renders form; attunement renders feeling. In actuality, however, there does not appear to be a true dichotomy between attunement and imitation; rather, they seem to occupy two ends of a spectrum." (1985:142)

Stern (1985:140) originally suggested that the mother begins affect attunement when the infant is around 9 months of age, but Jonsson et al. (2001) found that episodes of affect attunement were seen at just 2 months, were already more common than those of imitation (i.e. mimicry) by 6 months, and increased further in relative importance from then to 12 months at which point their study finished.

---

[121] See also Trevarthen and Aitken (2001:5, 12) on parent-child interaction from 2 months onwards, and Markova and Legerstee (2006) for an investigation into the relative importance of different forms of mirroring behaviours for affect regulation. Gergely and Watson (1996:1191) explain how contingency detection need not require perfect regularity in the structure of contingent response-stimulus events.

If this is all correct, then at the start of learning to pronounce speech sounds an infant already has extensive experience of picking up information about both his behaviour and his inner states from the mirroring behaviour of others. Within this experience, interactions where either his affective behaviour or vocal output is mimicked by his mother have been well documented. With respect to the other end of Stern's spectrum, we can see a parallel between affect attunement and vocal reformulation as follows.

In the case of affect attunement, the 'output' of the infant is interpreted by his mother to be the expression of a particular inner state (affective disposition). In the case of reformulation, the 'output' of the infant is interpreted by his mother as if it is the expression of a particular linguistic intention (an L1 utterance). In both cases the output thus interpreted is then reflected back to the infant.

From vocal reformulation, the infant is potentially assisted in objectifying any element of his sound making activity that he can isolate, i.e. a VMS. He can also associate this with what his mother believes to be its equivalents in her output. Note that I am not describing the infant matching his acoustic output to his mother's, but, rather, associating the inner representation of his VMS (his 'inner state') to the demonstration by her of what she has imputed to the acoustic expression of this.

So the infant is not attending to the outputs of each party and regarding them as similar. The outputs may be dissimilar, as where a child seems to have matched his movement for a given speech sound with an inappropriate adult output. Thus Locke and Kutz (1975) found that their 5½ year old subjects would substitute /w/ for /r/ in production even though they could distinguish the tokens both in others' speech and in their own speech played back to them. (Locke and Kutz's own conclusion was that the children were judging their own intent, and I would suggest, as explained in section 13.4.2, that this was in fact the auditory experience they would have been having.)

Note, then, with respect to this parallel:

1. I would be proposing that reformulation becomes an important interaction for speech later than affect attunement does for affect regulation.

2. As mentioned earlier, both affect attunement and vocal reformulation are only proposed to be mechanisms that help to explain the origins of self-knowledge in their particular spheres. Increasingly the child will be able to perceive aspects of his affective state and speech utterances for himself.
3. Not all reports of linguistic mirroring behaviour have distinguished between 'imitation' (i.e. mimicry) and reformulation, sometimes classing all such episodes as imitation.
4. The behaviours may not form a dichotomy; it may be correct to describe even a single interaction as containing elements of both. With respect to intonation, for example, while the exact frequencies of his mother's voice may be unattainable, the infant may well be able to mimic the shape of her pitch contour, which may then serve as a linguistically equivalent signal. At the same time, the infant may be informed about the qualities of sounds involved by reformulation.

If attunement is a functional mechanism for affect regulation, then it is plausible that the infant will pick up information about the linguistic value of his speech behaviour through the parallel interaction of vocal reformulation. If so, then 'speech sound mirroring' is a reasonable theoretical construct which would include a number of mechanisms by which an infant can inform himself about the meaning of his behaviours from the responses of others. It would encompass at least vocal mimicry, reformulation and reinforcement.

I have, perhaps, laboured the introduction of this idea, but the development of speech sound pronunciation has often been seen (or at least documented) as a continuation of how Hsu and Fogel (2001:88) describe most accounts of vocalisation as a precursor to language; where the emphasis has been on how infants, "creatively explore the sound-making abilities of their anatomical system with little need for adult motivation or shaping." This, of course, contrasts with the research tradition into vocalisation as a component of early parent-infant communication which emphasizes the intersubjectivity of the process.

I can see no reason for us to imagine that learning pronunciation is an activity undertaken by the infant using only his own resources directed towards his own and

others' speech[122]. My contention is that intersubjective mechanisms of learning are (1) available and (2) familiar to the infant. So while the mechanism of imitation for learning the pronunciation of speech sounds may not be available to a young learner (for the reasons given in chapter 13), the social mechanism of mirroring provides the alternative.

14.1.3 Mirrored equivalence (ME)

As I discussed in section 10.4.4, one feature of the pre-L1 stage of reduplicated babbling and protowords is that infants develop some Vocal Motor Schemes (VMS's): sound-producing articulations that are regularised and repeatable.

As we have just seen, at this stage caregivers reformulate some of the child's utterances, saying them back to him as if what he said had been well-formed (phonetically) in what will be, but is not yet, his first language. For example, his αα (where the Greek letter indicates a movement[123], in this case a VMS) can become his mother's *baba*.

By now, the infant is familiar with the various characteristics that mark out imitative games in general, including speech games, and he will realise that his mother is engaged in one now. He therefore knows that she is doing what she considers him to be doing. He can deduce that she thinks that whatever he does to produce his α is equivalent to what he perceives her to do, i.e. to produce the sound [ba][124]. He can thus establish an equivalence mapping (or correspondence) between his movement and what he perceives of her speech sound. Equivalence, rather than similarity of output, is what Kent (1981) and Studdert-Kennedy (1987) emphasised to be required for a speech sound IM (the relevant quotations from them are reproduced in section 13.4.3 and as the second epigram for this chapter, respectively).

This is illustrated in figure 14-1 for a general VMS α. The movement creates a sound of an indeterminate nature, but one which his mother takes to be equivalent to the speech

---

[122] And of course I am not unique in this, although I am giving more importance to it than any account I am aware of in the mainstream speech development literature. I discuss support from outside of the mainstream in section 14.2., but here note that Zukow-Goldring (1996) criticises theoretical approaches in language development which concentrate on processes internal to the child, and presents evidence in favour of a greater significance for dyadic interaction. In this she follows Lock (1980), whose notion of 'guided reinvention' would be a good description for the process of learning speech sounds that I am proposing.

[123] In MacNeilage and Davis's description of 'frame dominance', discussed in section 10.4.3, they present evidence that babbling does not have the segmental independence it appears to show. For this reason, my example has just two movements, provoking a caregiver response containing four 'segments'.

[124] My use of 'deduce' reflects my prejudices about the experience of being an infant. I do not think it affects any aspect of my argument if this learning is in some way implicit.

sound /x/ (the letter being used as a variable rather than designating a velar fricative). She produces an example of this, which the infant uses in the development of a perceptual category /X/ (which will eventually be the same as his mother's /x/, but may differ during development). The 'vertical' link between /X/ and α which solves the correspondence problem is created either by 'deduction' (the infant knows that his mother's response was, in her judgment, equivalent to what he has just done, because he knows from the behavioural context that she was imitating him) or by simple association based on contiguity.

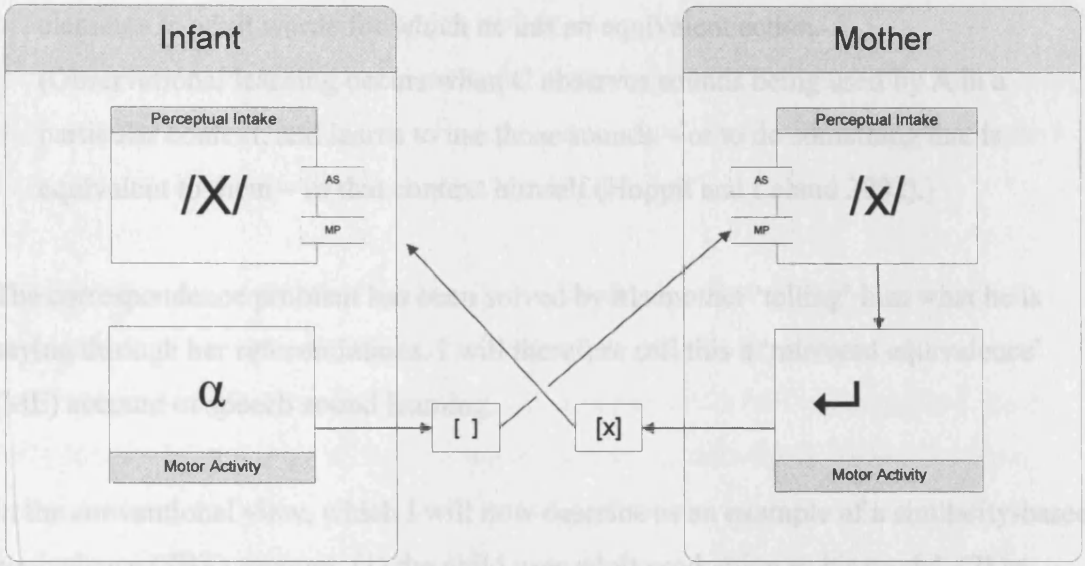

Figure 14-1. The infant's VMS α creates a speech sound whose exact quality is not important. The mother interprets this to be equivalent to her /x/ and produces a speech sound [x] in response. The child categorises this as /X/, and, knowing that his mother was imitating him, infers a correspondence between α and /X/.

Note that for the infant there is no distraction of word meaning at this stage; the interaction is about movements, auditory sensations, categorising and equivalence. His cognitive task is therefore relatively straightforward.

Correspondences between sounds and actions created this way provide a first entry into particulate speech, giving the infant equivalence pairs that he can transfer to mother-infant interactions. When he later begins to adopt adult words, any instances of a /ba/ that he hears will now be particularly salient, as will other sounds and syllables to which this process can apply. When he finds a /ba/ – or something similar – he can now do

something he knows that his mother will take to be equivalent: he will execute the movement α.

If this proposal is correct, a child's entry into a particulate sound system does not start in mimicry, but in a two stage process:

1. A 'training' period, during which he finds that actions he can already perform (the VMS's that he taught himself in babbling), are treated by his caregivers as equivalent to sounds that they produce.
2. Observational learning of words from his caregivers' speech: the child discovers elements in adult words for which he has an equivalent action.
   (Observational learning occurs when C observes sounds being used by A in a particular context, and learns to use those sounds – or to do something that is equivalent to them – in that context himself (Hoppit and Laland 2002).)

The correspondence problem has been solved by his mother 'telling' him what he is saying through her reformulations. I will therefore call this a 'mirrored equivalence' (ME) account of speech sound learning.

In the conventional view, which I will now describe as an example of a similarity-based equivalence (SBE) account, (1) the child uses adult production as his model, (2) he copies the acoustic output, and (3) he himself judges how adequate his attempt has been.

The ME account differs from a conventional one on each of these three points. Thus:

(1) The judgment of equivalence (probably, but not necessarily, based on similarity) is made by the adult, rather than a judgment of similarity being made by the child[125].
(2) The child discovers that, in certain circumstances, adults take what he already does to be equivalent to the acoustic results of what they do.
(3) "What he does" is considered by him to be his movements, not the sound output he or an adult hears.

[125] Ian Howard pointed out to me the benefit to the infant of relying on the perceptual system of an experienced listener rather than on his own underdeveloped one.

The first point means that speech sound acquisition is more straightforward for the child than we have imagined. His mother's behaviour transforms the challenge. He has less cognitive work to do, his perceptual system need not be as acute, and he does not need to solve the normalisation problem, whether via an innate brain mechanism or in any other way.

The third point may, at first sight, seem surprising. We are more familiar with the notion that the conceptual unit of speech for a young child is acoustic. However, this might only be the result of the widespread assumption that copying is the mechanism for speech sound development; an assumption which demands, of course, that the child conceives of his productions in acoustic terms. I am not aware of any independent evidence for this. (Or, as discussed in section 13.4.1, for the related notion that a young child has an inner voice, capable of evoking sounds. In its absence, I would struggle to imagine how acoustic representation would be possible.)

So while the adult will perceive the imitative interaction as a mapping from sound to sound, there is no reason to believe that the infant does. In fact Locke (1986:245), Kent (1992:84) and Davis, MacNeilage and Matyear (2002:102) have all suggested some early form of motor or sensorimotor representation for first words, rather than an acoustic one. A ME account allows for this to persist into later word learning.

For an example, imagine a game where the child makes an association between peeking round a doorway and his mother saying *boo*. Clearly it is an action here that is associated with a sound. Now, instead of the peek, substitute the child clapping, to which the mother still responds *boo*. A clap may have an acoustic component, which the child will be aware of and can use to help inform himself of what he is doing, but that doesn't mean that he need regard the sound of the clap he makes as similar or equivalent to the adult utterance. It can still be his movement – the act – that he associates with the mother's sound. Finally, substitute a vocal act for the clapping act: the structure of the interaction and its results need not change.

In infant-mother imitative interactions, the vocal act that will cause a mother to respond *boo* is, of course, for the child to blow his lips apart while voicing, maintaining that voicing for some time while his tongue is in a position where the maximum oral constriction is approximately velar.

For a detailed discussion of these ideas, see Bates (1979:52). She describes how, for children using gestures for naming, "vocal gestures and manual gestures are things that one simply does in association with certain referent objects and events." It is only, "from the comfort of our adult perspective," that we know that the vocal gesture made by the child, "has a solid functional basis"; in my example, that its acoustic results are 'similar' to the referent event (the adult sound).

If we look at the general development of motor skills, we should expect a motor stage of top level control to precede any structural displacement to a sensory one. After all, an inverse model has to be built before it can be used. Control of clapping (the action) comes before control through the intention to produce an approving sound.

In principle, then, the sound an infant makes might even be dissimilar to his mother's. In practice, she will usually use her notion of similarity to match her utterance to his. However, she might not, if, for example, the child had a speech defect that meant he could not articulate certain sounds. Then her judgment of equivalence – but not similarity – would be sufficient for her purposes, and from the child's point of view the action-sound mapping he inferred would also suffice, at least whenever his mother was his interlocutor.

14.1.4 Word adoption: parsing for production and word assembly

Children learn pronunciation through learning to say words, not learning to say sounds *per se*. In the previous sub-section I described word adoption as happening by 'observational learning'. Here I will expand upon what that might entail.

Clearly some comprehension of a word precedes its production for, "how else can speakers know which words to use to convey a particular meaning?" (E.Clark 1993:246). Clark and Hecht (1983) discuss and explain apparent contradictions to this.

So a word has been listened to and is now recognised. Let us use conventional terminology and say that a representation for it has been formed. If the correspondence problem has been solved for some of the elements that make it up, can this representation – or an abstraction from it – form the basis for an attempt at production?

The conventional assumption – the one underlying copying theories – is that it can. The result is that speech is believed to require only a single phonological lexicon[126]. This belief underlies the 'listen and repeat' and 'ear-training' pedagogies of language teachers and phoneticians (respectively), and is widely held among phonologists and speech scientists. The results of Eerola et al. (2003), however, fail to demonstrate the similarities between perception and production one might therefore expect. (The nature of the representation itself has been controversial, of course, with proposals and evidence put forward for auditory, motor and gestural coding.)

On the other hand, there are reasons to believe that speech is supported by two phonological lexicons, one for input and one for output. First, some psycholinguists point to the very different nature of perception (recognition of a word-form and recall of a concept) and production (recognition of concepts and recall of word-forms). They argue that this implies two very different forms of phonological knowledge (e.g. Huttenlocher 1974:335; E.Clark and Hecht 1983:336; H.Clark and Malt 1984:200). Analogously, perhaps, speech engineers create different lexicons for automatic speech recognition and text-to-speech systems because the tasks involved in each case are so dissimilar (but they make no claim to be modelling human faculties when doing this).

Second, neuroscientists have found evidence for dissociable routes of phonological processing from studying patients with disorders, and from neuroimaging healthy subjects (e.g. Romani 1992; Martin et al. 1999; Howard and Nickels 2005; but, *contra*, Coleman 1998; Shelton and Caramazza 1999). Most recently, Jacquemot et al. (forthcoming) have added to the evidence for separate phonological codes.

---

[126] Speech is produced with one apparatus and perceived with another. So there must be initial differences in the way that the brain processes the signals involved in each activity. But at the point where these signals interact with a semantic lexicon (or lexicons), there has been disagreement about whether any part of the phonological stage of processing is shared between hearing and speaking, or not. Depending on how the brain's activity is modelled, this issue has been framed as being between accounts of one or two phonological lexicons, one or two entry paths to the semantic lexicon, or one or two phonological networks. I have been using the 'phonological lexicon' terminology, but I have no particular commitment to this conceptualisation rather than any other.
The possibility that there are separate semantic lexicons for comprehension and production (e.g. Straight 1986; 1992) is not important for the arguments in this thesis. In fact, Straight's view, which he calls processualism, is more radical even than this:
"In place of the traditional view of a language as a set of abstract pairings of sounds and meanings, specified by structural rules, language is seen as two arrays of perceptual and motor processes, one for converting an input sound into an inferred meaning and another for converting an intended meaning into articulate speech."
H.Clark and Haviland (1974) make a general case for a process- rather than a structure-based approach to speech and language.

Third, some data from development, developmental disorders and therapeutic intervention argues in favour of two phonological lexicons. See Shuster (1998) for discussion and references.

I am impressed by the arguments in favour of two phonological lexicons, but some accounts leave open the question of how separate auditory and articulatory representational systems might be related. For example, Straight (1980:64) acknowledges this issue but does not attempt to resolve it. Before elaborating on my own proposal, I would like to discuss E.Clark's (1993:245-251) solution, for contrastive purposes.

Clark posits the existence of separate representations for comprehension (C-representations) and production (P-representations). She explains that children start by setting up a C-representation for a word they hear in input, and that they map meaning to this.

Importantly, she claims that, "the C-representation must contain information about the articulatory form of the word: the sound segments and their ordering" (p.246). This is not uncontroversial. Many researchers argue that the word representations used for recognition are holistic until at least some way into the word learning process. (Fowler 1991; Walley 1993:292, 2005:453/457)

Children can now start trying to produce the word. Clark is not explicit about the starting point for this, but says that they will need to set up a P-representation, containing articulatory information for it. The inclusion of detailed information about words in the C-representation allows Clark to propose that the child can improve his production (align his representations) through self-monitoring. He will be able to compare the output from his P-representation to the corresponding C-representation and to correct the P-representation accordingly[127].

[127] This will not necessarily be possible. As I pointed out in section 13.4.3, the type of comparison Clark presumably has in mind (one that will generate an error vector) requires the child to be able to evoke the image of a correct token. A simple recognition that a token fails as a category member does not, of itself, furnish the reason for this judgment. In this situation, the child has no indication about what to do to improve matters.

Clark claims (p. 251) that, "this view is incompatible with all accounts that simply take for granted that there is a single set of representations in memory, neutral between comprehension and production." However, it seems to me that the mechanism she describes cannot justify this statement. It is certainly true that comprehension and production require different forms of information, but these can be associated with a single representation. The segments of her C-representation – or a representation derived from them – would seem to be capable of playing this role.[128]

Two forms of listening

I would like to propose a different solution, though one which still generates two representations for a word. My starting point is the observation, from non-speech examples in daily life, that a P-representation is informed by a different type of perceptual activity than that creating a C-representation.

For the latter, only ordinary-noticing (as defined in section 10.1) is required. As a result, any feature of a new instance may be picked up and subsequently utilised for recognition. At the start, perhaps, gross features such as the overall shape of what is observed, but potentially finer and finer detail thereafter. Obviously there will be a functional aspect to this: where similar word forms contrast in only one portion of the signal then this must be attended to for discrimination to be possible. But functional concerns do not limit what will be noticed and then made use of. None of this, however, is organised or selected with any regard to reproduction. Indeed, if it is only the result of ordinary-noticing it is not amenable to evocation.

To create P-representations, on the other hand, we attend to a scene on the basis of what we know we can do, or what we think we might be able to do. For example, I can recognise a Ford Mondeo even from seeing very little of one (if it is parked among other cars, say). However, I cannot reproduce this car in a drawing from memory, because I have never examined a Mondeo with that idea in mind. I cannot evoke what

[128] Clark's proposal has the same structure as Kuhl's account of speech sound development. An auditory representation is acquired first, which serves as the standard against which copying attempts are judged. This may well function to help the child with the serial order of segments – learning by imitation – but may not help with improvement to sound qualities – learning to imitate. Like Kuhl, for the latter Clark would have to assume that there are no significant normalisation problems, that the child can hear himself veridically, that the auditory experience he retrieves from self-monitoring is of hearing the sidetone rather than the output of a forward model driven by his intention, and so on.

makes a Mondeo different from my general understanding of how cars look. The information in my C-representation is not available to me for that purpose.

If I now had a Mondeo in view, the way I would examine it, given this task, would be informed by knowledge of my own limitations as an artist. I would mark only the simplest aspects of its form in 2-dimensions, because I feel I would have a chance to produce this. I would ignore the 3-dimensional variation (the form of the front lights or the profiling of body panels, for example) because I know I do not have the ability to draw these things. A more skilled artist would examine the car in a very different way.

In this example, I have ordinary-noticed many things about Ford Mondeos in the past, which enable me to recognise one with ease. However, for reproduction in the form of a drawing I need to mark aspects of the car's form, which I will do with regard to my effectivities (possibilities for production). What I have marked is what I will later be able to evoke and to remark upon; or in this case literally re-mark, in the form of a drawing.

I would contend that the same is true of word forms. Production requires a separate examination of a word for that purpose, which is done with the speaker's speaking capabilities in mind. To do this we adjust our attentional set: we set up a filter on how we attend to what we hear. In the child's case, he will be looking for parts of words for which he knows he has equivalent VMS's. Of course, for many words these will only be a subset of everything that is needed to reproduce the word fully. So he will have to 'project' (Locke 1983:83) the best segments he can into gaps he identifies, and these will sometimes be judged to be incorrect by adults.

In summary, the production process will include a "listening" stage that is very different from the way words are listened to when they are acquired for comprehension. I will use the term 'parsing' for this special type of listening, performed with one's own effectivities in mind. Parsing retrieves a string of soneme-VMS units (and some gaps, perhaps), which form the P-representation.

Both listening for recognition (i.e. not recognition itself, but listening in the way that will make recognition possible) and parsing for production, are performed in MP mode. As a mature speaker I find it hard to disentangle these two auditory ways of attending to

words in myself. Parsing has become an effortless, automatic process so my impression is that even attending to new words is a unitary activity. However, I can still find examples in my own behaviour where the two processes are dissociated. For example, if I hear a complex, unfamiliar name on a news bulletin (a government spokesman for a Central Asian republic, pronounced by the newsreader, though, to conform to English phonology) or the exotic name of biochemical compound, I may resist parsing it for later reproduction, knowing that I am unlikely to ever need to say it. But I retain the overall sound shape, and recognise that I have heard the name before if I hear it again. Of course, the dissociation is clearer when one is learning a foreign language.

The auditory-motor correspondence described earlier, learnt through mirrored equivalence rather than by sounds being copied, enables a child to parse adult words for production purposes. His output is not dependent on the parallel processes that recognise words (which can be specialised for this task), but instead is supported by a separate lexicon specified in forms derived directly from his own effectivities.

I would claim that the two-lexicon model I have described will resolve the present conceptual disagreement across speech related disciplines. To demonstrate this by reviewing all the evidence goes beyond the scope of this thesis, but such a review would include the work of Straight (1980; 1982; 1986; 1992; 1993) as well as that reviewed in Jacquemot et al. (forthcoming).

Changing sounds

How is the pronunciation of words produced in this way improved? One piece of evidence that Clark advances in favour of the idea that children align representations by self-monitoring, is that they make spontaneous repairs to their own production, which typically move this closer to the adult form. To do this, she argues, they must be able to detect mismatches for themselves.

When these spontaneous repairs concern the sequencing of sounds or any other 'horizontal' aspect of a word, then a monitoring process of the kind she describes may well be in operation. However, if they are improvements in the qualities of speech sounds, I would contend that these are likely to be the result of the child's dissatisfaction with the reliability of his motor control (as when a golfer realises his swing needs attention). If so, they would involve him attending to auditory and other

MAP feedback (a different form of monitoring). They do not necessarily imply an auditory self-monitoring process where output is judged against an externally derived standard. (Although, as I will explain shortly, I can imagine such a process gradually becoming possible, and to the extent that the child is able to make judgments of SBE, he will, of course, make use of this.)

So, if not by auditory self-monitoring, how do the qualities of speech sounds in words improve? In the next subsection, I will argue that the child makes primary use of the reinforcement he receives from the environment, which may cause him to improve his motor control through practice or may prompt him to reparse the target with his attention drawn, perhaps, to where greater sensitivity to its content is required. This will be a view of an infant attending closely not so much to others' speech as to how others respond to his speech.

It is a view supported by some evidence I will present in section 14.2. It is consistent with Zukow-Goldring's (1996) vision of caregivers who act in ways that are sensitive to the learning needs of their charges, and Chouinard and Clark's (2003) evidence of pervasive reformulations of a child's output by his caregivers.

14.1.5 Development of sound qualities: reinforcement learning

As Hewlett et al. (1998:163) point out, a psycholinguistic model of speech must provide a means for the (progressive) revision of one pronunciation of a word in favour of another. Amongst other things, this could involve either differentiation of the sounds involved or their refinement. Differentiation will be needed because not all the speech sounds that a child will need to recognise and produce will start as naturally discovered VMS's that his mother responds to while they are still part of babble. Refinement will be needed because while parents are forgiving of pronunciation accuracy to begin with, their expectations rise.

In theory, there is no reason why the need for differentiation or refinement could not become apparent to a child from reinforcement only. This can take a multiplicity of forms. The child may already have a starting point for both types of development in an awareness of the approximate sound quality change required. However, for the reasons given in chapter 13, the child may not be able to perceive either a target or the results of

his own articulation with sufficient fidelity to make copying a possible mechanism to take things forward from here.

In this case, with the experience of having developed speech sound equivalences in the past, the child can embark on a more conscious use of a discovery mechanism. He can enhance his speech sound inverse model – give himself a new behaviour – in a way similar to that described by Young (1995) for foreign language learners (quoted in section 13.3). This process will be interwoven with testing of the new sound for its acceptability to listeners. So the infant would be creating a new or improved articulation (1) by doing something different and by feeling and hearing something different as a result, and (2) by then judging the new articulation by its efficacy (how listeners respond to it).

The judgment made by his interlocutors on the results of the new articulation will be based on the qualities of the sound it produces (as embedded in a word, of course). They can provide feedback on this in a number of ways (intentionally or not), including by reformulation of the child's output which, as I noted above, they do until the child is at least age four. So the child's search could be terminated on the basis of external evidence of acceptability rather than his judgment of similarity to adult models. The correspondence problem would again be solved in terms of an extrinsic sound to articulator movement equivalence.

This said, as time goes by, the child will increasingly be able to make his own judgments of similarity (the general view is that we do end up with a certain facility at hearing ourselves and copying sound qualities). To the extent that this ability exists at any time, the child will, of course, make use of it and he will have more than one source of information to draw on to judge the adequacy of his production. So the process will not be as strictly non-imitative as, for example, the original entry into speech sound production was.

The process of producing a new speech sound in this way avoids any perceptual problem of the type that might affect a copying account, where the child would be

required to perceive something new to begin a development[129]. Instead, learning a new articulation is the key to the system moving forward. With the mechanism described, action on the part of the child will simultaneously develop the articulatory skill needed for production and make his perception more sensitive to the acoustic features of the new speech sound (in particular, how it is distinct from what he produced for it in the past).[130]

As I noted at the beginning of chapter 12, there is less in the literature dealing with how children learn speech sounds than one might expect, partly, of course, because vowels are mainly learnt at an age where it is difficult to assess the basis of a child's behaviour. However, Jakobson (1972) argued that children principally seek to develop contrast among speech sounds, an idea that is consistent with Schvachkin's (1948/1973) belief that children will only learn contrasts among speech sounds when these make a difference to meaning, and with Clark (2003:121) who says that, "... vowel qualities in children's attempts at words appear to be selected to be discriminable for others."

As empirical support for this, Otomo and Stoel-Gammon (1992) observed that one of their subjects (who was significantly older than the age when vowels are normally acquired) did not learn to produce a particular vowel through copying the environmental input. Instead,

> "It seems that she actively explored ways to effectively contrast /a/ and /æ/, reorganizing her vowel system to create the necessary phonemic distinction."

The age of this subject may have provided the chance for an insight into what is normally an opaque process. She may have been doing what others do at an earlier stage.

Reinforcement learning

So far, I have only referred to negative reinforcement as a motive for change. (Where reformulations give the child evidence that his output does not correspond to what he

---

[129] As many philosophers have noted (and I discussed in section 13.4.1) we perceive what we are prepared to perceive, and what we are prepared to perceive is what we have perceived in the past. So how do we perceive something new?

[130] Hence the results from experiments where second language learners are given either perceptual training or production training, and the former only improves perception while the latter is found to improve both production and perception (e.g. Tsushima and Hamada 2005).
Shvachkin's (1948/1973) account of the interplay of production and perception also makes better sense in this light.

expected it would, I am taking this to be a reinforcing signal, rather than as the provision of something he will use as a model.) Reinforcement learning is sometimes looked at slightly askance. For example, Menn and Stoel Gammon (1995:354) write,

> "Perceived communicative success or failure in obtaining the desired response from the adult gives feedback, but if there is no adult imitation of the target word, this type of feedback – e.g. getting a smile, a different word, or a cookie – can only reinforce globally. The poverty of information in such responses – the fact that they can only tell the child that she/he did or did not get close to the target word – reminds us that **the primitive notion of reinforcement by obtaining a desired real-word objective is a rather blunt instrument for teaching**. Nevertheless, its contribution may be non-negligible, as communicative failure may lead the child to monitor the quality of his/her output more closely."

In modelling applications, by contrast, the star of reinforcement learning as a mechanism for problem solving seems to be in the ascendant (Sutton 1999). According to Sutton and Barto (1998), reinforcement learning is of great value in "uncharted territory" (situations where examples of desired behaviour that are both correct and representative are impractical to obtain) where an agent must be able to learn from its own experience. I have suggested that speech may be more like this than we have supposed. In the next sub-section I will give some examples of vocal learning by a form of reinforcement ('social shaping') that demonstrate its efficacy with infants even before the transition to words.

As Chouinard and Clark (2003) demonstrate, reformulations of all types, including phonological reformulations, are plentiful until at least age four. Other forms of reinforcement will also be a regular feature of a child's linguistic experiences.

In Part 1 of this thesis I questioned the assumption that imitation is the means by which temporal and some other speech phenomena are replicated. This assumption may have contributed to the role of reinforcement learning being downplayed in speech acquisition. To explain the appearance of pre-fortis clipping before 2 years of age, for example, it has seemed that young children must be attentive phoneticians, motivated to perform speech <u>well</u> (sometimes almost for its own sake). If children are highly motivated and expert imitators then it has not been necessary to pay much attention to the alternatives to copying[131].

---

[131] Also, as I noted in section 12.3.1, an objection to reinforcement learning as a general mechanism has been made based on covert contrasts in VOT times. The argument was that parents couldn't reinforce a

If, instead, this phonetic detail does not appear as a result of being copied but for other reasons, then we might change our picture of the young child as a junior phonetician for one where he is much more practical about his speech. His main motivation for revising the quality of a speech sound may be when he feels that he has executed what he intended correctly, but that he is still not understood.

Note that I am regarding reinforcement as a form of mirroring, so include this within the mechanisms that make up a ME account.

Criteria for similarity

It is generally believed that we can judge the similarity between speech sounds we produce and speech sounds we hear. To the extent that this is the case, how does this ability develop?[132]

One possibility is that it is just an extension of whatever process leads us to be able make such judgments about the sounds that other people make. Kuhl (1991) demonstrates this form of perceptual constancy in very young listeners, despite the objective differences in the acoustic signal produced by the wide range of speakers that were used (including children). If this ability can be applied to self-produced speech then it will be available at the time that words are learned.

However, if any of the reasons I have given for doubting such a transfer are valid, then the criteria for similarity between self- and other-produced signals could arise in another way, emerging later out of the equivalence classes for sounds created by children, rather than being the means by which such classes are formed.

Bates (1979) describes how means-ends relations, "can be 'acquired' through perception of contiguity, and 'reproduced' through imitation." I have described the equivalence relations which solve the correspondence problem for speech sounds being

---

distinction they could not hear. This too would be invalid if the proposals in Part 1 about the relationship of respiratory system activity and VOT are correct.

[132] I am trying to be careful in the way I describe this because I doubt my ability to make such comparisons myself and a part of me doubts it in others. But I am also poor at visual imaging yet have to accept that other people really can evoke vivid red dresses with yellow polka dots on them. I am certainly able to assess some voice qualities in my own speech, but I struggle with others.

created in this way. But she then asks how these individual relations come to be organised:

> "How do we ever get beyond memorizing long lists of relations by brute force? Surely we eventually come to truly understand how things work. Earlier on I mentioned that, after a cultural means is acquired, we can analyze the means-end relationship at our leisure. What do I mean here by "analysis"?
>
> There is ample evidence in all of cognitive and learning psychology to suggest that learning by association makes greater demands on memory than learning by similarity, and/or by extraction of general rules to account for classes of instances. There is, then, a constant psychological pressure to get rid of knowledge by contiguity alone. In the Peircian terms used here, by "analysis for understanding" we mean that subjectively arbitrary vehicle-referent relationships are broken down, examined, and replaced by subjective icons and indices whenever possible." (p.62)

After describing how this process might operate she asks about one of its products, the notion of similarity:

> "[T]he discovery of deeper iconicities or similarities seems to be in some way cognitively appealing, aesthetically preferable to a list of associations." (p.63)

A complementary perspective on this is given by Karmiloff-Smith (1992), who describes a general developmental process of 'representational redescription' which would lead to the same result (a theory of relations, or, in this case, the notion of similarity). Again, this is based on a "stock" of behaviours the child has mastered (implicit knowledge). In her description (p.18) this is, "the process by which implicit knowledge *in* the mind subsequently becomes explicit knowledge *to* the mind."

Thus criteria to judge acoustic similarity between self- and other-produced sounds could arise from a developing speech sound IM.

(See also Neisser (1987), Sloman and Rips (1998) and Mompeán-González (2004) on how similarity may not be the basis of categorisation but instead be a useful heuristic for membership identification that emerges from an initial theory-based classification.)

14.2 Support for a ME account

In section 13.4 I described some potential problems with SBE accounts of speech sound learning. One advantage of the ME account is that it is not vulnerable to any of these. In section 13.5, I gave two further reasons to doubt that speech sound qualities are copied: the pattern of development actually seen in children, and how children generally learn

complex motor skills by discovery rather than by copying. These two behaviours are consistent with a ME process.

I would now like to start to present the positive support for a ME account, beginning with its plausibility as a learning mechanism.

I have only recently come across the work of Yoshikawa et al. (2003), who demonstrate the potential of a constructivist/ME approach to infant vowel acquisition by means of a robot model. They make some of the points that I, too, have made, including the following:

- The normalisation problem is not addressed in most modelling studies.
- Caregivers often reformulate, rather than mimic, their infants' vowel cooing.
- Bone conduction would make a SBE mechanism of vowel acquisition difficult.

Their robot infant begins by making random vocalisations. A human 'caregiver' interacts with it through the two mirroring processes of reformulation and reinforcement. The result is that the robot solves the correspondence problem satisfactorily: it matches its articulations with the caregiver's vowels. Clearly their work directly parallels (and has preceded) mine.

Lacerda (2003:51) also makes the central point that I have made, that mirrored equivalence classes can be formed through adult feedback, independent of sound qualities:

> "Adult feedback in response to the infant's vocalizations is, in terms of the emergent perspective presented here, an important component of the language acquisition process. Although, as discussed above, the vocal output produced by the infant does not necessarily involve adult-like articulatory-acoustic correspondences, adult listeners often tend to interpret the infant's vocalizations in terms of speech sounds used in their ambient language. This adult interpretation can therefore be seen as a systematic bias (a "phonological filter," e.g., Sundberg 1998) that effectively structures the infant's phonetic variations (Routh 1967). **In other words, by providing feedback to the infant's spontaneous utterances, adults may help the infant to establish equivalence classes between babbled utterances and adult speech sound categories.**"

I will now note some 'circumstantial' evidence that ME would operate effectively in practice.

West and her colleagues have documented what they describe as 'social shaping' mechanisms operating in both songbirds (V.Smith et al. 2000; King et al. 2005) and human infants (Goldstein et al. 2003). In the first reports, the song development of young male cowbirds was affected by contact with non-singing older females. Vocal imitation was not a possible mechanism; instead, the females mirrored the males through reinforcement, by use of rapid wing actions ('wing strokes') and the degree to which they opened their beaks ('gaping'). In the human experiment, some features of the babbling of 8-month-old children (e.g. rate of syllable production, proportion of voicing) were shown to be affected by the non-modelling behaviour patterns of their mothers (e.g. smiling, touching, movement towards or away from the child).

These results extend older studies reported by Locke (1993:163), where mirroring affected infants' vocal output as a result of both positive and negative reinforcing behaviours. Altogether, this body of work demonstrates the effectiveness of reinforcement on its own for the development of sound qualities. An infant can and does attend to how his caregivers respond to his speech, and will modify his pronunciation on this basis.

Pedagogical practice in four speech disciplines demonstrates that an auditory model is not needed to develop the production of sounds, provided the learner receives appropriate feedback on his attempts. (In some cases articulatory guidance is given, but this is not practical, for example, for vowels. In this case, the learner may make rudimentary attempts based on auditory copying, and follow this by detailed investigation of the articulation required through trial and error.)

1 In speech therapy, successful results have been reported for various visual biofeedback devices (e.g. Shuster et al. 1995; Hardcastle and Gibbon 2005).

2 In second language teaching, the Silent Way (see section 16.3 for references) demonstrates that older learners can master the pronunciation of a new language through mirroring, with no vocal models apart from hearing other students' attempts at improvement.

3 I understand that articulatory training and reinforcement play a significant role in the instruction of deaf children for vocal speech. While reports of the results are

mixed, the principal cause when results are disappointing may not be the fact that the children cannot hear a model so much as their inability to hear themselves. Another reason may be how our limited understanding of some aspects of speech production have translated into pedagogical practice: for example with respect to the 'durational' phenomena discussed in Part 1 of this thesis. Deaf children who learn to speak well demonstrate that both these handicaps can be overcome, and that vocal models are not required by them to learn speech.

4 Although 'ear training' is conventionally central to courses in practical phonetics, there are phoneticians who would downplay its importance. To my knowledge, the merits of articulatory versus auditory approaches do not seem to have been the subject of published experiments with the exception of Catford and Pisoni (1970). They found that:

> "Performance on production and discrimination tests indicated a striking superiority for the subjects who received systematic training in the production of exotic sounds as opposed to those subjects who received only discrimination training in listening to these sounds. The results of this study suggest that what is effective in the teaching of sound production and discrimination is the systematic development by small steps from known articulatory postures and movements to new and unknown ones." (p.477)
>
> "For nearly every exotic sound [the group given purely articulatory training] did not *hear* the sound at all until they themselves produced it." (p.479)

Fischer-Jørgensen (1984:266) reports Sweet taking a similar view:

> "He criticizes the German phoneticians for basing their vowel systems on auditory similarity instead of production, and he recommends whispering the vowels in order to better feel the muscular sensations and says that training of the vocal organs is a better way of learning sounds than doing it by ear."

14.3 Comparison of ME with other accounts

In my discussion so far I have mainly contrasted the mirrored equivalence (ME) account with a conventional copying one. The way I am characterising these and some of the accounts described in chapter 12 is shown in figure 14-2. Here I will expand the discussion to other copying and 'discovery' accounts (both SBE and non-SBE), which will help to establish the potential scope of a ME one.

	Copying	Discovery
Similarity-based equivalence (SBE)	Conventional Kuhl ...	Vihman 'echo re-enactment' Davis and MacNeilage
Gestural equivalence	Goldstein and Fowler Studdert-Kennedy ...	
Mirrored equivalence (ME)		Yoshikawa et al. Messum

Figure 14-2. Characterisation of accounts by (i) basis for equivalence and (2) form of matching process.

I am not in a position to prove that a ME account is correct, or that any other account is incorrect. However, there are the potential problems for all the SBE accounts described in chapter 13, which do not seem to have been widely appreciated. The ME account is not vulnerable to these problems.

The positive support for ME that I summarised in the previous sub-section establishes, at a minimum, that the mechanism is successful in other contexts: robot, animal and human, and with young and old learners. In the next chapter I will claim a different form of support, in its ability to explain data that has hitherto been perplexing.

14.3.1 Comparison with copying accounts

It is perhaps worth saying again that there is no experimental evidence that children learn the qualities of speech sounds by copying them. This has only ever been an assumption, made in the apparent absence of alternatives.

The potential problems of chapter 13 would tell against the SBE copying accounts, perhaps less so with respect to Kuhl's than the conventional one. I also described copying accounts which solve the correspondence problem through innate mechanisms for cross-modal signal matching, with mirror neurones being put forward by some to

play a central role in this[133]. As someone whose understanding of learning is thoroughly imbued with Gattegno's theoretical ideas (e.g. Gattegno 1975, 1987) I am suspicious of appeals to innateness, and I would argue that a ME mechanism makes such an appeal unnecessary.

Goldstein and Fowler (2003) propose that children copy gestures directly. This is part of a research programme which encompasses Articulatory Phonology (Browman and Goldstein 1992) and the body of evidence that Fowler and her colleagues have assembled in favour of the 'direct' perception of speech. The proposal must remain as one possibility pending the resolution of the wider debates on the nature of perception and the basis of speech production.

14.3.2 Comparison with discovery accounts

I have argued that learning speech sounds may not be an exception to a general pattern whereby children learn complex motor skills by discovery rather than by copying. If the discovery relies on a match made by the child based on acoustic similarity, then in section 12.3.3 I described a process of 'echo re-enactment' as taking place.

There are points of similarity between this and the ME account. Both describe an infant creating something first (a sound or a movement) and only then discovering that its counterpart is used by others linguistically. Both potentially allow the learning of speech sounds comprising consonants and vowels, together and separately. And both mechanisms potentially allow for pronunciation development from the early stages of word adoption to the final completion of the speech sound inventory.

There are also differences:

1 Most obviously, the matching mechanism in the ME account is not based on the child's judgment of similarity. Instead, he establishes a two-way correspondence between a VMS and those sounds produced by his mother in response, based on his judgment that she is taking them to be equivalent. This primes his listening, leading to recognition of members of this category of sound (i.e. this speech sound) in words spoken by his mother, which he is then in a position to reproduce using his VMS.

[133] There is no necessary connection between mirror neurones and accounts which invoke innate matching, of course. Mirror neurones linking perception and production might be the result of learning in an emergentist account, as Vihman (2002) describes and as I think would also be true for my account.

2 The scope of each account is different. Based on the potential problems with SBE accounts I described in chapter 13, the need for a ME mechanism rather than a SBE one is likely to be greater with respect to vowels than consonants:

- Firstly, because the psychoacoustic difference between consonants produced by children and those produced by adults is probably less than the difference between child vowels and adult ones. If little or no normalisation is required, self-made judgments of similarity by the child on his and adult output would be easier for consonants.
- Secondly, because bone conduction is likely to affect the discrimination of voiceless sounds much less than voiced ones (and voiced contoids less than voiced vocoids).

So, at the limit, it seems possible that learning to produce an instance of [s] in a word by echo re-enactment would be little affected by these problems. A weaker version of my proposals, therefore, would have speech sounds being learnt by a mixture of discovery processes; by echo re-enactment for voiceless contoids, and ME for all voiced sounds or for all vocoids.

3 The profile of cognitive activity would be different for the two accounts. While an echo re-enactment mechanism is simpler in some ways than the ME one (not depending on interaction with another person, for example), the latter would be less demanding on the young child during the period when words are adopted by being parsed and reproduced rather than by being re-created holistically. The significant cognitive work in the ME scenario is done before the child encounters the speech sound in a particular word in the stream of speech. At an earlier point, he creates a VMS and then a perceptual category for the sounds that his mother makes which are equivalent to his VMS. His later task at the stage of word adoption does not then require either normalisation or judgment of distance from a prototype. He just has to perceive this latest token as falling within the category that determines the speech sound.

I will now compare the ME account to the 'discovery' accounts I described in section 12.3.3.

The developmental focus of Vihman's articulatory filter is different from the ME account. Vihman describes the articulatory filter operating as part of the process of early

word selection and word production (and does not discuss later differentiation of speech sounds *per se* or refinement of pronunciation under lexical or parental demands). The ME account is concerned with the production of speech sounds.

If the earliest words adopted from caregivers are produced not with speech sounds but as *gestalts,* and therefore re-created using the auditory IM rather than a speech sound one, then it is possible that the articulatory filter as Vihman describes operates at this first stage. Later, when the basis for word production changes to being particulate, the rationale for ME being the mechanism which develops speech sounds for a speech sound IM would apply.

Note, though, that the working of the ME mechanism does allow for the possibility that early adopted words are not completely mimicked, though. Their basis may be the first equivalence pairs created through reformulations during babbling. There may even be a hybrid form of production from the start (or later), with some parts of words reproduced, the rest re-created. On an analogy with learning to write this does not seem too fanciful. I may be able to write a <d> and an <o> at a stage where I still have to draw a <g>.

That said, many authors have described qualitative changes between production of the earliest adopted words and production of later ones which would certainly be consistent with a change from re-creation of a word shape to reproduction of a word. For example, Snow (1988:348):

> "There is, clearly, development during the one-word stage in the kinds of pragmatic forces words convey and in the complexity of the phonological system needed to deal with the various distinctions introduced with successive words. However, it seems that the major developments occur after the thirtieth and fortieth word is attained. In fact, it is striking (and related to the discussion of continuities versus discontinuities in development) **how fragile are the early 10-15 words, and with what difficulty they are acquired**. Only after a lexicon of 40-50 words is attained do children really become efficient word learners. Before that, each item represents a long and difficult process. **Thus, it seems that all the first 15-25 words are within one system, a system in some ways closely related to babbling**, and that only somewhat later does the process of lexical acquisition really change, with some systematization of the sound representation and production and the introduction of stable referentiality."

The account given by Davis and MacNeilage also supports this, so from now on I will make the working assumption that the first adopted words are mimicked, but that over

time there is a change in the basis of production which leads to the use of speech sounds – learnt by ME.

I make use of the same reformulation mechanism described by McCune (1992:331), but her account of this came within a conventional paradigm of imitation of sounds by the infant based on their similarity (p.330). Also, her focus is on elucidating how VMS's function within situations linked with meaning. It seems important to me that we can apply the idea earlier than this. I have asserted that when meaning is present, word recognition mechanisms in MP mode will dominate, limiting the availability of the acoustic sensory data for any possible matching process.

In the passages I quoted in s.12.3.4, Locke describes a mechanism by which there may be, "a motor basis to speech perception – and internal representations – from the very start." This might operate prior to a more systematic period of word adoption when a ME mechanism takes over. I.e., the two may be complementary.

14.4 Summary

The main question I have considered in this chapter is how the correspondence problem for speech sounds is solved by a child. How, in other words, he learns to re-enact (or imitate) speech sounds so that he is able to adopt words (to learn them by imitation).

Kent (1981:167) notes that,

> "To learn speech, the child must ... establish equivalence classes between his or her acoustic patterns and those of the adult models. Equivalence classes have to be formed because the child's short vocal tract cannot produce exact acoustic replicas of adult speech sounds."

The word 'establish' here can be read in different ways. Most naturally it seems to imply that the child must himself match his acoustic output to the adult acoustic input, implying a judgment on his part which he is only in a position to make on the basis of similarity. This is how most accounts describe speech sound acquisition: the child as a largely autonomous learner working to make his output resemble the input, with the precise nature of the required normalisation process ill-defined at present. No necessary role seems to be played by adult responses, although presumably negative reinforcement like misunderstandings may be helpful in drawing his attention to the need for continuing work on his articulation.

I would like to rephrase Kent's formulation to emphasize that while equivalence classes must be formed, the child might do this between his articulatory movements and adult acoustic patterns, i.e. to solve the correspondence problem in its classical formulation.

My account is of the child not being in a position to copy speech sound qualities at the start for a number of possible reasons. Most of these imply that the signals he would have to compare perceptually are opaque for this purpose; more than one reason may cause this. But his mother provides him with a bootstrap into particulate (as opposed to mimicked) word reproduction with her reformulations of his output during late babbling and beyond. She 'tells him what he has just said', and he can make use of the equivalences he deduces in the reverse direction when he starts to adopt words by the ultimately preferred route of parsing, representing and reproducing them using a speech sound IM. Continuing reformulation and reinforcement of his output by his caregivers guides the subsequent development of his pronunciation.

Adult responses, then, are not marginal but central to the process. I used the term 'mirroring' to encompass at least the mimicking, reformulation and reinforcement behaviours on their part that this could include.

Of course, as the child develops his production it will educate his perception. To the extent that extrinsic and intrinsic input thereby become perceptually less opaque, he can start to modify speech sound qualities based on the copying of others' output. But the criteria for similarity that would support this will emerge from the evidence contained in his functionally defined equivalence classes, rather than criteria of similarity forming these classes in the first place.

Word adoption may start with whole shapes being mimicked using the auditory IM developed through early sound play. But the mature form of word adoption for particulate speech has a stage where the child parses the target with his effectivities (the speech sounds he knows he can produce) in mind. This perceptual process is different from the 'ordinary-noticing' that he has used to build the phonological input lexicon by which he recognises words for comprehension. Hence production has a phonological lexicon that is distinct from the perceptual one.

15 Some implications of a mirrored equivalence (ME) account

I have described SBE accounts of speech sound development and their potential problems, and proposed an alternative account that is not vulnerable to the same concerns and which is plausible for a number of other reasons.

In this chapter I evaluate these different accounts from a psycholinguistic perspective. I will consider how representations are developed in speech sound learning, the implications of this for underlying structures in the brain, and evidence from shadowing experiments and neurology.

Then I show how the ME mechanism generates a new and integrated model of speech production and perception, which also resolves some longstanding problems in speech.

15.1 Comparison of ME and SBE: categorisation behaviour

My focus in this section is on the categorisation behaviour of the child and its psycholinguistic consequences. I will compare the mirrored equivalence (ME) account with three similarity-based equivalence (SBE) ones: what I have called the conventional account, Kuhl's variant on this and echo re-enactment. I have described the first two as 'copying' accounts and the third as a 'discovery' one.

All the SBE accounts assume that the child can and does hear his own output veridically. (I will call his own output the 'intrinsic input'.) They also assume that he can normalise it, at least to some extent. A number of the potential problems I discussed in chapter 13 would undermine these assumptions, but let us take them to be possible for the present.

Figure 15-1 is reproduced from chapter 12, as a general diagram of the processing underlying SBE accounts of speech sound development. (Chapter 12 also has derived figures that show Kuhl's proposed mechanism and echo re-enactment separately.) The accounts differ in the sequencing and timing of events. Thus in the conventional account the extrinsic input is held in memory and is followed by the intrinsic input. In Kuhl's account the extrinsic input is that used many months previously during perceptual learning, creating categories of sounds now co-opted to guide production. In echo re-

enactment the intrinsic input comes first, creating a prototype against which the extrinsic input is matched.

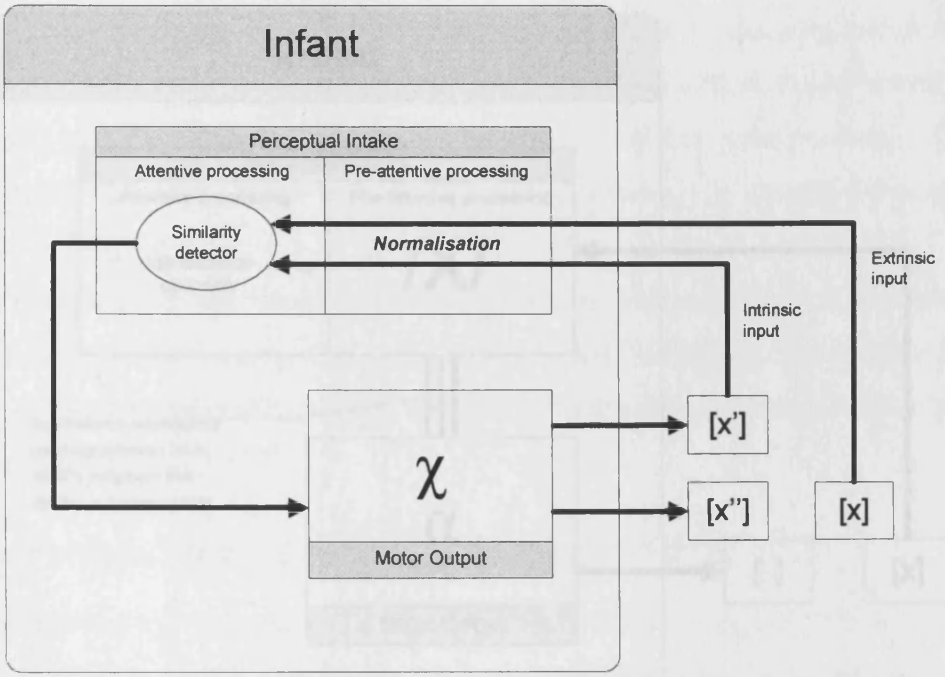

Figure 15-1. Generic mechanism if judgments of similarity are made by the infant.

However, the post-normalisation judgment of similarity between inputs means that in all the SBE accounts there is an auditory representation mediating between the extrinsic input and the movement which will reproduce it. The extrinsic input must be recognised "as a …", even when this process is automatic and the listener is not consciously aware of it. The result of the comparison is the trigger for the action. Thus the extrinsic input indicates the most appropriate articulation for the child to use to match the speech sound in a word that is to be produced[134].

Figure 15-2 illustrates how the ME account would differ from the accounts above, particularly in the neural structure it creates. From the pre-L1 training period onwards, the infant's developing VMS's are reformulated by his mother. The perceptual category /X/ is formed from examples of her (and others') speech in response to a particular VMS α. Category formation will be accomplished prior to the system being used for

[134] Rather than specifying it. The distinction between the two is discussed by Studdert-Kennedy (1987:69).

adoption of words. The category will not be based on a prototype, but on an integration of the various examples presented to the child in response to any single VMS.

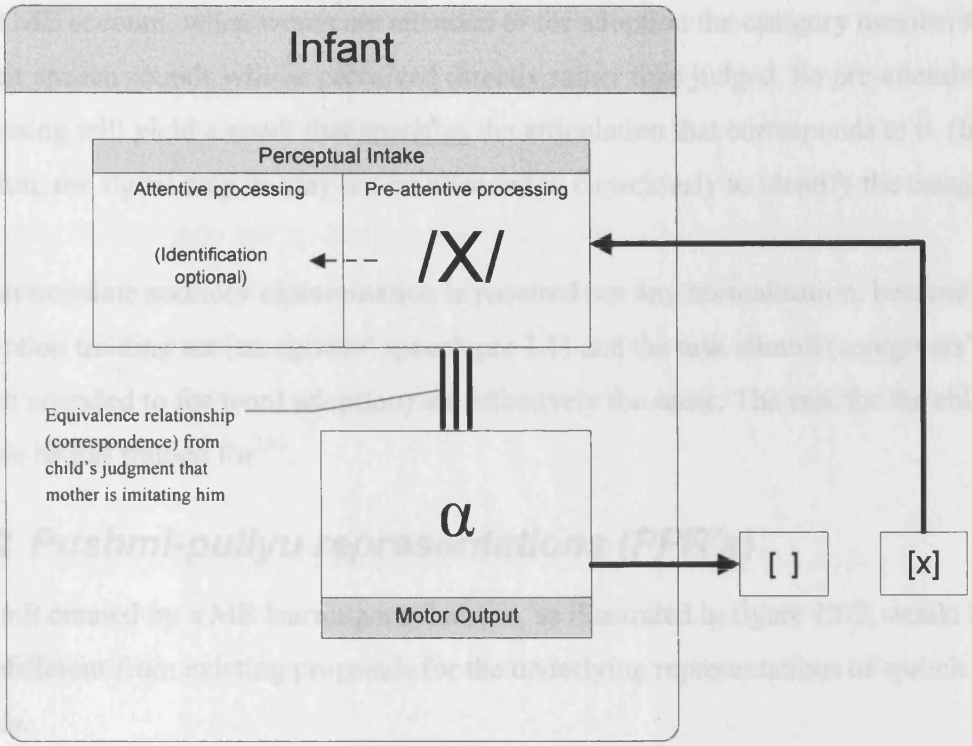

Figure 15-2. Expanded view of left hand side of figure 14-1 (mother's activity to produce [x] in response to infant's α is omitted)

Notice that the perceptual category is built around the occurrence of the VMS in the context of an imitative interaction. The VMS category is the perceptual category's identity or 'label', and the reverse is also true; the label for the VMS category is the perceptual one.

Bruner et al. (1956) drew a distinction between perceived and conceptual (judged) categories. They observed that recognition of 'perceived' category membership appears to be immediate, while 'conceptual' judgments of membership take time and some mental effort[135]. This now has greater theoretical and experimental support, including neuroimaging data which shows that the processing associated with categories formed

[135] They also pointed out that conceptual categories can be altered into categories that can be utilised with more immediate perceptual cues, giving the example of how untrained and more experienced doctors might come to a diagnosis. But in the early stages of speech acquisition, I am assuming that a child using echo re-enactment, for example, would not have attained the faculty of immediate perception with respect to speech sounds uttered by others in relation to his own prototypical output from a VMS.

by integration of information from multiple sources and those formed from prototypes differ in significant ways. See Maddox et al. (2002), Keri (2003) and Ashby and Maddox (2005) for further information.

In the ME account, when words are attended to for adoption the category membership of their speech sounds will be perceived directly rather than judged. So pre-attentive processing will yield a result that specifies the articulation that corresponds to it. (In addition, the signal may or may not be attended to consciously to identify the category.)

No intermediate auditory representation is required nor any normalisation, because the perception training set (caregivers' speech pre-L1) and the task stimuli (caregivers' speech attended to for word adoption) are effectively the same. The task for the child is the one he has trained for[136].

15.2 Pushmi-pullyu representations (PPR's)

The unit created by a ME learning mechanism, as illustrated in figure 15-2, would be very different from existing proposals for the underlying representations of speech sounds.

15.2.1 Description

The unit seems to be a natural development of the perceptuomotor structure that Studdert-Kennedy (1987) and MacKain (1988) proposed. From what seems to be a largely shared starting point, they developed accounts of how speech must be represented in the brain. These treated different aspects of the issue but appear largely complementary, and most of their arguments seem to have continuing force.

However, both accounts incorporate the assumption that children replicate sounds by copying. It seems to me that a unit created as a result of a ME mechanism solves the difficulties that they themselves identify without it being necessary to appeal to innate mechanisms of matching in any form. It would be compatible with the other general requirements they describe.

The unit is distinctive in a number of ways. Firstly, the usual assumption made by speech modellers is that articulation is moulded by the speaker to meet his own acoustic

[136] Rachel Smith pointed this out to me.

goals. This is not the case here: the unit's motor component is not derived from its auditory one.

Secondly, the reverse is also not true; its auditory component is not derived from the output of its motor one. Such a derivation would occur, for example, as a result of echo re-enactment, where the acoustic results of the child's VMS's form the prototypes for categories by which extrinsic input is recognised.

Instead, both the motor and auditory components of the PM unit are developed independently. The former by discovery (trial and error), guided by the child's sense of what is new in terms of articulation, and new in terms of the auditory and other proprioceptive information he receives as feedback. The latter by categorisation of the acoustic input.

Thirdly, the PM unit is distinctive because it does not require the link between perception and production of speech sounds to be made through an intermediate representation, whether amodal, acoustic or motor. Direct associations are made instead.

I am aware of four recent precedents for a representation of the type I am proposing and how it might be created.

Millikan (1996/2005, 2001) calls the first a "pushmi-pullyu" representation (PPR), for the following reason:

> "I have argued, however, that there are two opposing directions in which mental representations may operate. They may be designed to represent facts they are not responsible for creating but which need to be taken into account to guide (physical or mental) activity. Or they may be designed, like blueprints, to help create the affairs that they represent. All of the most primitive representations, however, face both these directions at once. I call them "pushmi-pullyu" representations (Millikan 1996) after Hugh Lofting's charming creatures having that name. My favorite example is the bee dance, which tells the bees where the nectar is (facts) but tells them where to go as well (directions)." (2001:895)

In other words, "... the same complete representation token can have two functions at once, being both a fact-presenter and an action-director." Her favourite example from ordinary language is, "No, Johnny, we don't eat peas with our fingers."

Jeannerod and Jacob (2005:313) comment on the similarity between this idea and a visuomotor structure that they argue is distinctive from that underlying normal visual percepts (which represent the world) and is the result of dorsal stream processing that supports intentions for action.

The third recent precedent would be the direct links connecting sensory and motor representations posited under Heyes' Associative Sequence Learning model of imitation (e.g. Heyes 2001:258; Bird and Heyes 2005). She suggests that when a movement is perceptually transparent these links are formed by co-activation arising from correlated experience of observing and executing the same movement unit. This contrasts with theories of imitation that require there to be intermediate representation.

I am suggesting that a mirroring interaction could lead to similarly direct links if a movement was perceptually opaque.

Finally, of course, a fourth possible precedent would be a representation instantiated in some way by the much-discussed mirror neuron system (Rizzolatti et al. 1996, 2006), since under some circumstances mirror neurons respond both when an event is observed and when it is performed. This is how the representation I am proposing would be expected to behave, so the mirror neuron results are supportive inasmuch as they show that such a structure can be implemented in as small a structure as a single neuron[137].

Millikan's terminology is so attractive that, with apologies if I am misapplying her concept, from now on I shall describe the neural structure created by ME as a pushmi-pullyu representation (PPR).

---

[137] Speculatively, it may be that to create mirror neurons requires mirroring interactions of the type I have described.
Where B imitates A through a judgment of similarity he compares an intrinsic and extrinsic input that are both sensory. If attention and resonant states are needed for learning, there is no reason to expect this to lead to a neural connection between the sensory input from her action and the movement he makes in response, as his attention will have been on the two sensory representations. (The one he generates will be linked to his movement, of course, but that leaves the motor scheme at one remove.)
On the other hand, a deduction of equivalence between A's output and his (already in-repertoire) movement, based on B's recognition that she is now imitating him, gives B a reason to associate his movement to the sensory result of her action. We experience such connection phenomenologically, of course: as an "aha" moment or a *prise de conscience* (a 'taking' of awareness, in the sense that a blancmange or cement 'takes', or sets). It does not seem unreasonable to imagine that such vivid experiences (and less vivid ones) lead to a neural instantiation of the 'connection' made.
In this scenario, B must recognise he is being imitated. The cognitive sophistication required for this recognition would explain why mirror neurons have only been found in higher primates.

15.2.2 Implications

It is obvious that the motor/sensory aspects of speech production and perception are specific to each process. At higher levels of brain function, however, almost every possible organisation of processing for speech has had evidence advanced in its favour.

One axis of disagreement has been whether there are one or two phonological lexicons (or 'access paths') to one, or two, semantic lexicons. I discussed this earlier in section 14.1.4.

Another is whether the underlying representation for speech sounds is fundamentally articulatory or acoustic. These alternatives have been extended to the possible existence of both types of representation simultaneously, of an underlying representation that is neither one nor the other, and to the possibility of all three being present (Nolan 1990; Coleman 1998). I have just explained how a PPR is a departure from all these proposals. Further complexity is introduced by the different requirements of the groups who develop models of speech organisation (phonologists, phoneticians, developmentalists, psycholinguists, neurologists, etc) and the different evidence they use to judge them by.

Ultimately, we must develop a unified account of how the mind, the brain and the body deal with speech. I think that a different view of how children learn speech sounds is the comb that could straighten out some of the tangles that now exist. I now address one of these.

Shadowing and repetition impairment

As just mentioned, there has been a longstanding debate between those who regard speech as being best characterised as gestures made audible (a view held by Stetson, Motor Theorists and Direct Realists) and those who would view it is as a primarily acoustic code (a view held by Sapir[138], many others before and since, and probably the majority of contemporary speech scientists).

---

[138] "Language is primarily an auditory system of symbols. In so far as it is articulated it is also a motor system, but the motor aspect is clearly secondary to the auditory. ... The motor processes and the accompanying motor feelings are ... merely a means and a control leading to auditory perception in both speaker and hearer." (Sapir 1921: 17-18).
See the papers collected by McGowan and Faber (1996) for a concise introduction to many of these issues.

Supporters of the 'gestures made audible' position can point to evidence from shadowing, which shows that choice reaction times with speech material are not significantly longer than simple reaction times (e.g. Porter and Lubker 1980). On the face of it, this is incompatible with a traditional model of how speech is processed; it appears to show that in some way, 'speech is special'.

Fowler et al. (2003:398) summarise some of the early results:

> "Porter and Castellanos (1980) ... compared latencies to make vocal responses in both a simple and a choice reaction time test. In both tests, a model speaker produced an extended vowel /a/ and then after an unpredictable period of time shifted to /o/, /æ/ or /u/ (Porter and Lubker) or to a consonant-vowel (CV) syllable (Porter and Castellanos). In the simple task, participants shadowed the extended /a/, but as soon as they detected the change to a new vowel or CV, they were to shift to /o/ (Porter and Lubker) or /ba/ (Porter and Castellanos). In this test, shifting from /a/ merely signalled that a change had been detected. In the choice task, in contrast, the subject's task was again to shadow /a/, but then to shift to whatever vowel or CV the model shifted to. Accordingly, listeners had to do more than detect a change from /a/; they had to determine what was being said and to say that themselves.
>
> Luce (1986, p. 208) reports that, generally, simple response times are faster than choice response times by 100-150 ms when the two tasks are made comparable. This latency difference is understandable in that the simple task only involves detection of a stimulus (or stimulus change) whereas the choice task requires that a choice among responses be made depending on the identity of the stimulus.
>
> In contrast to this typical finding, Porter and Lubker found just a 15ms difference between the choice and simple response times; Porter and Castellanos found a 50 ms difference. Moreover, response times were very fast, so the small difference between conditions was not due to simple responses having been slow. Why, under their experimental conditions, is the choice/simple response time difference so small? Porter and Castellanos suggest that "subjects are able to directly and accurately realize the results of perception in articulatory terms" (p. 1354)."

Fowler et al. (2003) have extended the earlier results. Their subjects shadowed VCV utterances (e.g. /apa/) where the initial V was of varying length, and were required to change to either a predetermined or a stimulus-dependent CV as soon as the initial V changed to the second syllable. The results confirmed the approximate equivalence of simple/choice reaction times.

Fowler et al. favour either a motor theory or direct realist account of their data. They propose that a subject perceives the vocal tract gestures of the model and is thus informed of the articulation required in the choice reaction time test as part of his process of perception, without the need for an additional stage of processing after

recognition of an auditory category. However, they acknowledged that an "augmented" acoustic theory might also explain their initial data, if it allowed for,

> "... articulatory properties as well as acoustic ones to be associated with phonological categories. From this perspective, in the choice task, listeners perceive the disyllables acoustically, but the consequence of mapping the cues onto a phonological category is that articulatory properties are made available." (p.407)

In other words, stored with the perceptual category information is information for producing a token of the category. The former specifies the latter, rather than indicating it. Discussing this possibility, Shockley et al. (2004:422) point out that they are unaware of any theory in which this is actually proposed[139].

However, it would seem to be exactly the effect that having a PPR as the structure for the underlying representation of speech sounds would produce. As can be seen in figure 15-2, the most basic correspondence established between perception and production would be between a phonological category and the action which the child learns will be taken to be equivalent to it.

Other work has examined the shadowing of longer passages of speech material. Marslen-Wilson (1985) summarises a series of experiments which identified so-called 'close shadowers', women able to reduce the delay between hearing and production to 250 msec or less[140]. The difference between close shadowers and 'distant shadowers' (latencies averaging over 500 ms) was that the former were able to use the products of on-line speech analysis to drive their articulatory apparatus before they were fully aware of what these products were. Word specification as PPR's would explain this very naturally.

McLeod and Posner (1984) proposed that shadowing was an example of a special class of input/output transformations that they called "privileged loops". These would be separate from general information processing and performed with relatively little interference from other cognitive activities. They pointed to patients (echolalics) who

---

[139] In Fowler et al.'s fourth experiment, data was collected which would appear to tell against any such theory. However, this interpretation was based upon the premise that VOT's are a timing phenomenon. If, as I argued in section 6.3, the timings of VOT are epiphenomenal, then the plausibility of a ME account is not undermined.

[140] It is hard to imagine that there are no men at all able to do this, but the 8 subjects who did were all female, out of 40 men and 25 women tested.

can repeat speech without showing any awareness of its meaning in support of this, and described neuropsychological evidence for the reality of a separate system for repetition.

McCarthy and Warrington (1984) provided further evidence for this, from repetition impairments in patients with two types of aphasia. They described "a two route model of speech production" (see figure 15-3). Using their terminology, the auditory/ phonological transcoding process bypasses verbal semantic systems. However, they pointed out that, "the biological necessity and *modus operandi* of dual processing routes in speech production remain obscure." (For recent developments with this basic model, see Hanley et al. 2002, 2004.)

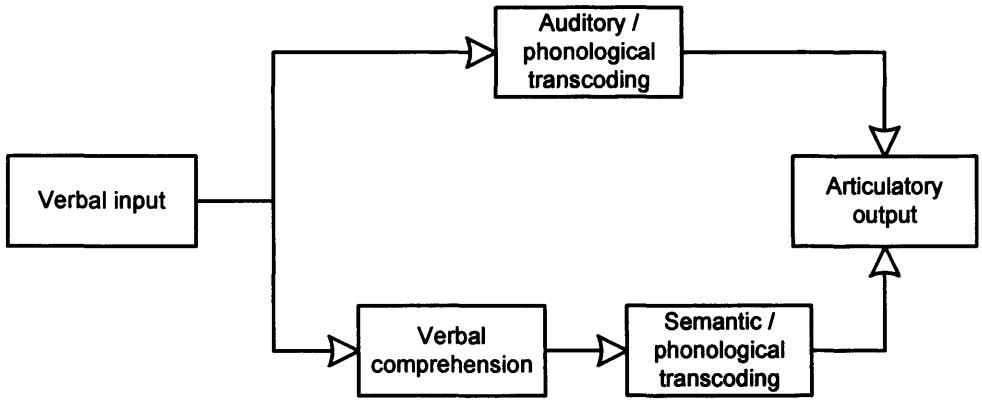

Figure 15-3. Redrawn from McCarthy and Warrington (1984) "A Two-Route Model of Speech Production." Fig 4 (p.481) "A functional model of the speech production process"

My contention would be that two routes to speech production are a natural result of speech sounds being learnt through mirrored equivalence, and the construction of the PM units that this entails. In fact, if we count mimicked phrases then there may even be three routes to speech production. I discuss this further below, when describing a model of speech and hearing development, but the idea is illustrated in figure 15-4 (organised to be easily compared with the model of speech production described in Hewlett et al. (1998)). My 'parsing for production' would equate with McCarthy and Warrington's non-semantic route and McLeod and Posner's privileged loop, and would be the route taken by close shadowers.

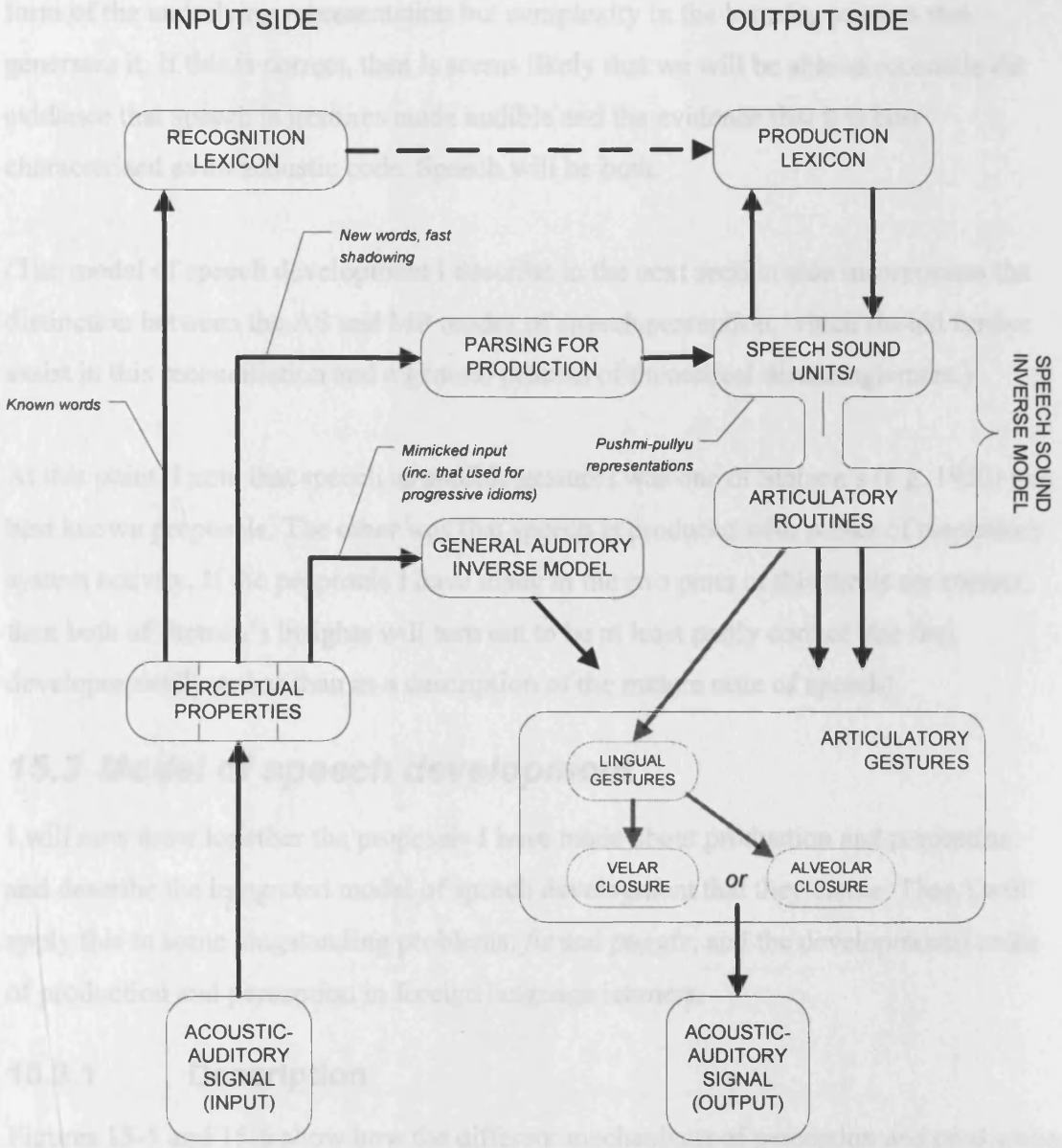

Figure 15-4. A model of the components involved in the processes of speech perception and speech production, structured for direct comparison with Hewlett et al. (1998) Fig 4. (The original illustrated the case of a target form of 'cow', which the child perceived correctly but sometimes pronounced with an initial alveolar sound. Hence the particular articulatory gestures shown lower right in grey boxes.)

Motor and auditory representations of speech

I have proposed that the learning of a speech sound by mirrored equivalence creates a PPR. Its perceptual and motor aspects can be used by the child to represent and reproduce speech sounds. A PPR is, perhaps, the most natural of possible forms of underlying representation for a speech sound. The forms that have been proposed previously fail to account for all the data because they place complexity in the brain and

simplicity in the learning process. Instead, I am proposing that there is simplicity in the form of the underlying representation but complexity in the learning process that generates it. If this is correct, then is seems likely that we will be able to reconcile the evidence that speech is gestures made audible and the evidence that it is best characterised as an acoustic code. Speech will be both.

(The model of speech development I describe in the next section also incorporates the distinction between the AS and MP modes of speech perception, which should further assist in this reconciliation and a general process of theoretical disentanglement.)

At this point, I note that speech as audible gestures was one of Stetson's (e.g. 1950) two best known proposals. The other was that speech is produced with pulses of respiratory system activity. If the proposals I have made in the two parts of this thesis are correct, then both of Stetson's insights will turn out to be at least partly correct (the first developmentally rather than as a description of the mature state of speech).

15.3 Model of speech development

I will now draw together the proposals I have made about production and perception and describe the integrated model of speech development that they create. Then I will apply this to some longstanding problems: *fis* and *puggle*, and the developmental order of production and perception in foreign language learners.

15.3.1 Description

Figures 15-5 and 15-6 show how the different mechanisms of perception and production lead to a range of adult speech processes. Perceptual mechanisms are separated according to their use of awareness of sensation (AS) or meaningful perception (MP), to the left or right. Production mechanisms are located centrally, but this area is itself divided into three streams: for babbling, protowords (L0) and the language of the environment (L1). (This diagram is a simplified version of one which explores in detail the distinctions this layout allows us to draw.) Developmental time moves from top to bottom, where the results of different activities are shown in dashed boxes.

Figure 15-5. The processes leading to two of the speech phenomena of adult production and perception (in dashed boxes, below). Areas for the two modes of perception (labels top left and top right) flank a production area, which is subdivided into babbling, protoword (L0) and native language (L1) streams. Developmental time moves from the top to the bottom of the diagram.

As many people have noted, an infant alone in his crib can experiment with the articulators that will become his speech production apparatus and he can simultaneously attend to the consequences of his actions. For all children these include changes in pressures, flows, temperature etc, and for a hearing child there will be auditory feedback as well. He will repeat articulatory gestures to the point where he feels he can rely on his ability (i) to create the same movement and (ii) receive the same sensory feedback from it (as he does when learning any other motor activity).

For this, we assume that he has a version of the echoic memory we are familiar with as adults, meaning that he can track the auditory results of actions and compare them with the auditory results of actions performed immediately before. Hence repetition can become controlled by sensory as well as by motor criteria: control can pass 'to the ear' (i.e. there can be a so-called 'structural displacement' (Ivry 1996:286) in what he attends to in order to control an articulation). I have asserted that these activities will be undertaken with AS, so any notions he develops about sounds will be based on their surface similarities (or, more correctly, on the effects the signals have upon his auditory apparatus).

Longer term memories of auditory feedback will also develop. The result will be an inverse and a forward model linking articulations and the auditory feedback he receives from them. This is noted at the top left of the main area of diagram.

He can also listen to utterances produced by others in his AS mode. The echoic memory of such utterances (with their pitch movements probably particularly salient) can form a target for him to apply his inverse model to. So my label "Recreating others' output" would be more carefully described as capturing their output in AS mode, and matching the auditory results of his actions against this. Sometimes this process will produce an output that adults recognise as words or a phrase, a so-called 'progressive phonological idiom', when the child's 'pronunciation' of an utterance he has copied appears to be in advance of his general pronunciation as judged by the qualities of speech sounds in other words (e.g. Jones 1967; Ferguson and Farwell 1975:432).

These activities lead to an increasing ability for the mimicry of sound patterns. The key characteristic of this is that resemblance is judged by AS. It is meaningless surface characteristics that are re-created, although the activity as a whole may not lack

purpose: a mimic may have the intention of getting his audience to think of a specific action or event.

The AS mode of listening and the re-creation of a sound pattern from matching its results endures as a possible mechanism of vocal communication. Adults do learn to 'run off' some words spoken to them by L2 language informants. But, as with a child's progressive phonological idioms, the limitations of this approach to speech soon become apparent. A basic communication system can be constructed this way, but the re-creation of long sound sequences controlled by their effect on our auditory apparatus is extravagant of attentional resources. Mimicry exists, but it is not the basis for speech.

On the right hand side of the diagram, the perceptual process by which the sound patterns that a child hears become associated with meaning is described, initially, as "Word recognition". In chapter 11, I argued that there is no reason for us to imagine that this is an active process to begin with. Until there is some benefit to the child in this activity we should view it as an example of learning by acquaintance, and this characterisation helps to explain some features of infant performance (see the next section of this chapter).

The basis for early word recognition is likely to be 'holistic' word-shapes (Hallé and de Boysson-Bardies 1996; Vihman et al. 2004). In time, a listener learns multiple cues for words, including cues based on what he learns about the structure of words once he starts producing them. These and his top-down expectations mean that recognition can sometimes be achieved at a very early point in a word's reception; and in other cases at some time after reception. There is no reason to believe that recognition relies upon the knowledge of any phonological organisation of words that is developed for production. Hence an independent phonological lexicon for words is created, with a wide range of information to enable word recognition (bottom right of diagram).

In figure 15-6 the main path to word production has been added. Babbling creates VMS's. The first PPR's are formed from interactions with caregivers that are not linguistically meaningful, so the child's bootstrap to learning speech sounds does not have the distractions that communication would bring (nor an understanding of how it will benefit him).

As I mentioned in section 14.3.2, the first words selected for adoption by a young child (perhaps by Vihman's articulatory filter) may be re-created as whole-word shapes, using the auditory inverse model developed during pre-linguistic vocal play. But a time comes when the child perceives the perceptual 'head' of one of his early PPR's in a word spoken by his mother. (The head is a phonetic category, so identification occurs within the MP mode of listening.)

If performance of the articulatory head now reproduces the word successfully enough for it to be recognised by his mother, then he has discovered a new and more economical route to word production. Now, in addition to recognising words in MP mode, he can attend to them with his PPR-based effectivities in mind – or mark speech sounds that 'pop out' if this occurs - to enable his reproduction of the words. I have labelled this process of observational learning 'Parsing for production'. This leads to 'Word assembly', whose major challenge is the serial articulation of the elements identified. No judgments of sound similarity need to be made within this process.

Figure 15-6. The main route for the development of speech production has been added to figure 15-5. The processes leading to all the principal speech phenomena of adult production and perception (in dashed boxes) are now illustrated.

That there should be a second way of attending to words within MP in addition to 'Word recognition' is what we should expect from our knowledge of how other skills are learnt. Prior to adopting a behaviour there is usually a need to parse it for its effectivities (the aspects of it that the observer can reproduce) rather than just for identification. I may have considerable experience in watching horse racing, even to the point where I can identify riders by their style, but if I learn to ride myself I will start attending to a wholly different set of attributes in my instructor and other riders: ones that have been present but hitherto unnoticed or not understood[141] by me in the absence of a need to reproduce them. (I gave a similar example to this, relating to drawing a Ford Mondeo, in section 14.1.4.)

If words were stored in a form equally suited for perception and production there would be no need for a separate parsing for production once they had been represented for recognition. However, the fact that even adults can recognise words that they cannot yet say[142] suggests that the representations for recognition and production are separate. This is clearer in children than adults, because as mature speakers we do not often hear words that are unfamiliar, and when we do, our 'parsing for production' automatism is so smooth and practised that it can be hard to discern it at work. (I gave an example earlier, though, of my resisting the parsing of the name of a foreign government spokesman which I judged I would never have reason to either say or to evoke in private thoughts.)

Not portrayed in the diagram is the later development of speech sounds from information not originally gleaned in pre-linguistic infant-mother interactions. For the differentiation and refinement of sounds when the child is already a speaker, he can employ a more informed and increasingly deliberate use of the interplay between production, perception and mirroring.

---

[141] By 'not understood' I mean aspects of other people's actions for which I have no practical or experiential understanding. I may, however, have noticed the surface characteristics of these actions. This means that I can mimic some aspects of being a horse rider or mimic being a tight rope walker without having any hope of staying on a horse or a rope if I actually attempted to do so. Similarly, mimicry of word shapes by a child does not imply any ability on his part to perform the generative acts that syllabic speech articulation requires. For this, an inverse model must be built for each speech sound individually.

[142] This is not the same as the issue of one's passive vocabulary being larger than one's expressive vocabulary. That is a function of recognition and recall mechanisms. In an adult, even a word which is not normally available for expression will usually have received a production parsing.

One effect of 'Parsing for production' will be that the child develops a sensitivity for the internal structure in words and thus for auditory cues which identify particular speech sounds. This will inform the ways that words can be recognised (as will his increasingly sophisticated top-down expectations). The dotted arrow pointing to "Listening for meaning" has been added to indicate this. He now has the chance to develop a number of new heuristics for word recognition in addition to the whole-word approach that he may have started with.

"Listening for meaning" has replaced "Word recognition" to indicate the change in how perceptual word forms are attended to. Prior to the possibility of self-expression the child acquired these largely by acquaintance. Now, the power of words for self-expression and his success in doing something that adults do[143] give him an incentive to more actively acquire sound-meaning correspondences for both perception and production.

At the bottom of the diagram, the two dashed boxes "Processes for word reproduction" and "Processes for recognition" generate the two phonological lexicons or phonological entry/exit paths to whatever form semantic information is stored in.

The 2 (or 3) routes to word production are (i) the typical route during expression, out of a semantic lexicon via the phonological reproduction lexicon; (ii) a non-semantic route involving parsing an input and producing an output 'on the fly', used for unfamiliar words, by 'close shadowers', etc; and, occasionally, (iii) mimicry of foreign words, names etc.

Figure 15-6 makes clear how it can be the case that perception generally precedes production (inasmuch as words are recognised before they are spoken and speech sounds appear to be contrastive before they are used contrastively), but that this is of less significance than has been imagined in learning to pronounce a language. The different type of perception that production requires is closely bound to a child's effectivities, and, indeed, to the effectivities of an L2 learner.

---

[143] Marilyn Vihman pointed out this second motivation to me.

Shvachkin (1948/1973) gives a careful account of the influence of articulation on the phonemic development of speech and the interaction of hearing and articulation. In many ways what I have proposed seems in accord with what he describes.

Compatibility with neuroimaging data

Is the model I have been describing consistent with recent neuroimaging of the speech process?

My 'parsing for production' seems compatible with the flow of information between brain structures supporting perception and production described by Scott and Wise (2004:26). Further, my model has a separate mechanism which would correspond to their description of a flow of information through the planum temporale acting as a sensori-motor interface for mimicry (p.27).

Similarly, a mode of 'parsing for production' would naturally support the working of a phonological loop for verbal working memory. This would use motor systems (articulatory rehearsal) or virtual motor systems (evocation) to keep sensory-based representations (the phonological store) alive, along the lines Baddeley has described (Hickok and Poeppel 2004:87).

Hickok and Poeppel (2004:91) argue that the evidence available supports partial but not complete overlap in the neural systems supporting speech perception and speech production, as described in my account. Further, that, "the mapping of sensory representations of speech onto motor representations may not be an automatic consequence of speech perception, and indeed [is] not necessary for auditory comprehension."

Scott and Johnsrude (2003:100) suggest that,

> "the robustness of speech perception might, in part, result from multiple, complementary representations of the input, which operate in both acoustic-phonetic feature-based and articulatory-gestural domains. ... These parallel, hierarchical processing 'streams', both within and across hemispheres, might operate on distinguishable, complementary types of representations and subserve complementary types of processing."

If so, then, "speech is processed both as a sound and as an action," (p.105) as I have described.

On the other hand, Hickok and Poeppel (2004:77) describe the explicit segmentation of a word into phonemes as being a dorsal stream function. My model would locate such processing within an MP mode. Hickok and Poeppel, however, seem to have been strongly influenced by the notion that children learn speech sounds via self-supervised learning, in which case a, "sensory-motor integration network" (operating in an AS mode of listening) would indeed be necessary (p.86). However, I believe that their premise is mistaken.

Finally, Pulvermüller et al. (2006) report that, "during speech perception, specific motor circuits are recruited that reflect phonetic distinctive features of the speech sounds encountered, thus providing direct neuroimaging support for specific links between the phonological mechanisms for speech perception and production." I have only seen an abstract of their paper, but this also seems compatible with what I have proposed.

15.3.2 Implications

Production deviation from the adult norm

Since the *fis* issue receives an uncomplicated solution within the model I have described I will not review previous arguments around it. There is detailed discussion in Clark and Clark (1977), Locke (1979) and Priestly (1980).

I would contend that a child recognises the adult *fish* correctly but creates a *fis* outcome from a faulty 'parsing for production'. Let us represent PPR's within curly brackets, showing the sound perceived from adult intake followed by the articulatory movement that becomes associated with it. Then if the outcome of a movement sigma is /s/ (i.e. this is what adults 'hear'), the child who says *fis* for *fish* may be articulating,

{/f/:phi} {/I/:iota} {/S/:**sigma**}

because he has not yet developed a distinctive movement for the palato-alveolar fricative which should finish the word. Alternatively, of course a distinctive movement may be being made, but the outcome is still /s/ because it does not cue successful adult recognition of a /S/. (He is producing a covert contrast.)

However, the child is aware of what he intended to say or, perhaps equivalently, he is monitoring the output of a forward model driven from the phonological score. His primary auditory experience will not therefore be the result of hearing the sidetone, unless he allows "sensation to trump expectation" (Grush 2004:383), in which case he will be aware of his error without necessarily being able to do anything immediate about it (see the discussion of auditory experience in section 13.4.2). From his perspective he has done what he can to produce the word correctly.

The real question is why he is not willing to act on the evidence from reinforcement that his pronunciation is unsatisfactory. He has, after all, been willing to be guided by this in the past and hence has entered into cycles of differentiation to create new sounds. However, anyone who has struggled to learn to pronounce a second language might find it easy to imagine several reasons for his stasis. If he has tried and failed to make a distinction in the past, and if he usually 'gets by' with what he does now, then he has a disincentive to change without any real incentive to do so. Menn and Stoel-Gammon (1995) describe how this situation normally resolves itself.

The *puddle* /pʌgəl/, *puzzle* /pʌdəl/ problem (N.Smith 1973, 2003; Macken 1980) receives a similar explanation. The surprise here is that despite being able to articulate /pʌdəl/ – as demonstrated by his pronunciation of *puzzle* – a child doesn't use this articulation when saying the word *puddle*. Here his PM units are {/d/:gamma} and {/z/:delta}. So he does not have an auditory /pʌgəl/-like representation anywhere: this is just the outcome of the movements he makes. And there is no reason for him to respond to *puddle* with a {/z/:delta} that is apparently inappropriate.

Menn's (1983:38) description of Daniel repeating *duck* correctly and then reverting to saying /gʌk/ when he retrieved the object would be explained, as she does, by assuming that initially he did not comprehend the word. His repetition was mimicked. When he recognised the word he produced it using a PPR which equated /d/ to a velar articulation rather than a dental one.

Production and perception of forms in L2

The model I have described also suggests a natural explanation for what has appeared to be puzzling data on the temporal relationship between production and perception in L2

learners (reviewed by Llisterri (1995) with more recent experimental contributions by Bradlow et al. (1997) and Tsushima and Hamada (2005)).

The well-known results of Goto (1971) and Sheldon and Strange (1982), that some aspects of Japanese learners' production were in advance of their perception, are disconcerting to a view that sound qualities are learnt by copying. However, if my model is correct then the results should come as no particular surprise. Mirrored equivalence does not require any relation of auditory similarity between the perceptual head and the motor head of a PPR, so children's pronunciation can be in advance of their perception for a particular speech sound. It is, of course, unlikely to be judged this way because of the non-adult sound quality they produce with a reduced size vocal tract.

Furthermore, there is no requirement that the perceptual head of a PPR be related to a speech sound from the environment. A perceptuomotor association can be made with any arbitrary symbol; for example the spelling of a speech sound or a coloured rectangle on a chart (as in the Silent Way). These can then trigger the production of a sound.

If learners have worked on their articulation and then base their production on how they know a word is spelled, then there is always a possibility that this will be superior to their ability to perceive the sound involved. In fact, since production will educate a (hearing) learner's perception but perception only weakly improves production, it is efficient to teach learners through production, as Catford and Pisoni (1970), Sweet, Gattegno and others have asserted (see section 14.2).

*15.4 Early speech perception**

I gave a brief account of some issues in speech perception in chapter 3. My main purpose was to argue that word recognition in infants is unmotivated and happens by acquaintance. Only ordinary-noticing would be needed to support this. However, ordinary-noticing would not generate representations adapted for use in word production.

In this chapter and the previous one, I have described how a different perceptual process, 'parsing for production', creates the separate representations that support word production. Faced with a word they wish to adopt, young children and adults will mark

and then reproduce those elements that they believe they have the ability to say themselves.

Testing these assertions will have to be a project for the future. Here, I will elaborate a little on what I have described so far.

- There seems to be no need to imagine an infant recovering strings of phonemes (or sonemes) as part of recognition, as Warren (e.g. 1999:169-173), Suomi (1993) and others have also argued. (See also Scott and Wise 2004:19.)
- At the same time, he may ordinary-notice aspects of the signal at any scale, from 'holistic' properties to fine detail. Recognition would be multi modal, as described, for example, in the Polysp model (Hawkins 2003).
  (What is ordinary-noticed may, of course, include portions and parts of the acoustic signal which happen to map onto phonetic categories and features, but the infant would only perceive these by their 'appearance' rather than by their linguistic function, i.e. as rectangles and textures rather than bricks and properties of bricks.)
- Both parsing for production and production experience will help to educate perception, enabling the creation of new heuristics for word recognition.

One further issue is the level of detail encoded by a young child which is made use of for recognition purposes (Werker et al. 2002; Swingley and Aslin 2002; Bailey and Plunkett 2002; Swingley 2005). On the one hand, I would not imagine this being limited to what is functionally necessary to distinguish words. As I have just mentioned, as part of a child's deepening acquaintance with words, he may ordinary-notice many aspects of a word's form. On the other hand, the detail encoded need not support production; that is the responsibility of the separate 'parsing for production' process.

It seems to me that these considerations will support a new perspective on results previously reported. However, I will not attempt to describe this here.

15.5 Summary

This chapter examined some implications of a mirrored equivalence (ME) mechanism for speech sound development, as opposed to a similarity based equivalence (SBE) one:

- Such a mechanism leads to the creation of a new form of underlying representation for speech sounds. I adopted Millikan's (1996/2005) terminology for this, calling it a pushmi-pullyu representation (PPR). It is formed from one 'head' that is a perceptual category of speech sound (a soneme), and a second that is a regularised movement of the child's articulators (a VMS). Between them, there is an associative link, created by the child on the basis of the evidence he has that the two are equivalent.

  Unlike the representation of a speech sound described by SBE accounts, neither head is derived from the other. Nor is there any need for a further, abstract representation.

- This structure explains the otherwise puzzling data from shadowing experiments. It may also resolve the longstanding controversy about whether speech is best characterised as gestures made audible, or as an acoustic code.

- I was then able to describe a model of speech development which integrates speech perception and speech production (see figure 15-6). This incorporates my proposals from chapters 10, 11 and 14: perception is split into two modes of AS and MP, initial PPR's are created prior to words, 'listening for meaning' is separate from 'parsing for production', and so on.

  This then provides straightforward explanations for the *fis* and *puzzle/puddle* phenomena.

16 Conclusion

In this thesis I have argued against the idea that children learn either the replication of certain 'timing' phenomena by modelling or the reproduction of speech sounds by imitation, and I have argued in favour of alternative mechanisms that would lead to the outcome we observe. I will summarise my proposals shortly, after noting a few additional reasons why I find them attractive.

Firstly, they make learning to pronounce a more normal activity:

> "... a 'production-based' approach has the important advantage of bringing together language learning and other kinds of learning that occur in childhood. For instance, no one would seriously defend the idea that a child learns how to build with blocks primarily by analyzing the block constructions produced by others. Rather, one would assume that the child learns from his or her own constructive operations and their outcome .... Yet theories of language acquisition, of whatever signature, mainly acknowledge the role of input in the learning process, not that of children's constructive production." Elbers and Wijnen (1992:341)

Secondly, it is satisfying that an apparently arbitrary and disparate set of 'timing' phenomena in English can be given principled explanations which bind them into a now more coherent whole.

Thirdly, I have been able to present an integrated model of speech perception and production in development (in section 15.3, and illustrated by figure 15-6). Many people have drawn attention to the fact that such a model is needed. This is the first I am aware of that takes account of our different modes of listening to a sound signal (as now supported by the neurological evidence of dissociated streams of perceptual processing).

Fourthly, the neural structure that would be created by a mirrored equivalence mechanism for learning speech sounds promises to resolve the longstanding debate about the underlying nature of speech sound representation. Speech would not be either gestural or an acoustic code; it would be both, at the same time.

Finally, since the inspiration for this thesis came from Gattegno's insights into speech, I am happy that my proposals are consistent with his general view of learners and learning.

Much theory, research and practice in speech disciplines relies on the assumption that the aspects of pronunciation I have considered are learnt by imitation. There is no evidence for this. It has simply been assumed from the fact that children do come to produce speech that resembles that of the speakers around them. Thus the decision about which account we should prefer is not one that pits a challenger against a champion who has earned his position. The incumbent has never demonstrated any claim to the title. In fact, theory, data and some circumstantial evidence have allowed me to argue in favour of the challenger instead.

16.1 Summary of Part 1

In Part 1, I considered how children replicate various 'timing' phenomena. I dealt in detail with one 'language universal', pre-fortis clipping, and three phenomena that characterise West Germanic languages: foot level shortening, vowel characteristics that create tense and lax classes, and long lag/short lag voice onset times.

It is generally believed that children learn these phenomena by imitation (or modelling). A child is supposed to identify the underlying temporal structure in the speech of others, and to adapt these models or 'rules' to his own production apparatus for planning his output.

I described an alternative: that it is the constraints arising from speech being embodied that lead to these phenomena appearing. Their timing characteristics are then epiphenomenal.

The constraints I invoked relate to speech breathing and speech aerodynamics. I pointed out that because speech breathing is a motor skill that is learned both (1) during the period of speech acquisition and (2) at a time when a child's respiratory system is so much smaller than an adult's, we cannot expect it to have mature characteristics from the start. Instead, I argued that a child's speech breathing will be pulsatile, and that each pulse will be produced with the same level of effort whatever the 'segments' it is used to produce happen to be. Thus the activity of the respiratory system and that of the upper articulators are in a frame/content relationship. Speech breathing is dominant, and articulation is adapted to what it allows.

With respect to aerodynamics, I pointed out various asymmetries in the scaling relationship between child and adult speech production systems. These impose further constraints on child speech that are absent from adult speech. I used the phrase 'breath stream dynamics' (BSD) to encompass the SB and aerodynamic factors that I considered.

Within this framework, pre-fortis clipping emerges naturally as a result of the need to distribute a limited aerodynamic resource among parts of a syllable that have different aerodynamic requirements.

I then extended my model of speech breathing to take into account how a child speaking a West Germanic language will implement stress-accent. I argued that he will continue to use a pulsatile style, but with greater effort expended on foot-initial syllables.

By reanalysing the weak syllables of a foot in terms of their aerodynamic characteristics – the load they present to the respiratory system – I was able to claim that foot level shortening is essentially the same process as pre-fortis clipping. The need to allocate limited aerodynamic resource also explains other 'compression' effects, which do not, therefore, have a basis in rhythmicity.

I also explained the characteristics of tense and lax vowels by reference to the implementation of stress-accent. The greater respiratory system effort it demands poses a number of threats to a child-sized speech production system. To deal with potential loss of the pressure head, unwanted turbulence, etc, the child has to modify his production strategies for vowels. The side effects of this include the lengthening of tense vowels in some contexts, and the phonotactic constraint against lax vowels appearing unchecked (e.g. word finally).

With respect to VOT, I argued that it will be natural for children learning English and German to discover a distinctive set of long-lag plosives as a result of making greater respiratory system effort at the start of feet. The primary feature for a child, however, is aspiration, rather than the relative timing of the laryngeal and oral gestures.

In my account, then, stress-accent precipitates these and some other phenomena. The imitative account of their replication is rather implausible for a variety of reasons. Chief

among these is the complexity of the modelling task. On the other hand, a BSD account is more consistent with the developmental data, requires no substantial and ongoing cognitive engagement by the child, and explains some otherwise puzzling aspects of English phonetics.

Although the BSD constraints on a child need not directly affect adult production, I argued that speech timing is probably planned in such a way that it can be implemented with a range of styles of respiratory system control. In this way, the mechanisms discussed continue to affect speech into maturity.

16.2 Summary of Part 2

In the introduction to Part 2, I described a Ruritanian family in which both child and mother have unimpaired seeing and hearing, but where the child can sign, but not speak, and the mother speak, but not sign. Thus the infant perceives speech in one modality (sound) but can only produce it in another (gesture). Clearly he will not be able to learn to sign by imitation – either by mimicking whole-word shapes or by mimicking the cheremes of the language - because he has no model to copy. Nevertheless, I asserted that he will learn to both sign and hear if his mother provides appropriate assistance (without her even having to know that this is what she is doing).

In chapter 13, I gave a number of reasons why we might think of a real infant as being in essentially the same predicament as the Ruritanian one with respect to speech sounds. If any one of these is true, then during the process of learning to say words using speech sounds, the perception and production of a child are effectively operating in different modalities, contrary to the general assumption that the acoustic signal provides common ground between them.

However, as in my Ruritanian family, a child is able to learn to pronounce words using speech sounds because of the way his mother acts as a mirror: first bootstrapping his entry into speech sound to movement correspondences during pre-linguistic imitative play, when the 'signing' he does with his articulators is 'observed' aurally by his mother and reflected back to him as speech sounds, and then responding to his attempts at speech. These mirroring interactions enable him to develop speech sounds with the acoustic qualities that the environment expects and demands.

In more detail, I started to develop my argument[144] in chapter 10, with three preliminaries:

1. There are several phenomena that we call 'noticing'. 'Ordinary-noticing' creates a trace which can help with subsequent recognition, but which does not support evocation. For that, an awareness has to be marked.
2. There are two modes of perception. In one mode, we attend to a signal by being aware of the effect it has on our perceptual apparatus. I called this 'awareness of sensation' (AS). In the other, we bring previous experience to bear to find what a signal reveals about the world. I called this 'meaningful perception' (MP). I asserted that mimicry is an AS function, based on the resemblance of the sensory effects of the signals produced by the model and the mimic.
3. When a child's production routine for a sound or sound pattern becomes regular, we can call it a vocal motor scheme (VMS). For some sounds, VMS's develop before his output has any linguistic significance.

In chapter 11, I argued that a child has little motivation to learn the meaning of words before he is able to use words to express himself. Nevertheless, word recognition develops by acquaintance and through ordinary-noticing. As a result, the way that words are represented will not support their reproduction through speech sounds. These representations may, of course, support whole-word shape re-creation, and may contain fine acoustic detail.

I then pointed out that the equivalence between speech sounds produced by a mother and a child might be established on the basis of function rather than by criteria based on similarity.

In chapter 12, I distinguished between learning <u>to</u> imitate and learning <u>by</u> imitation. The latter normally refers to learning sequences of acts, when for each individual act the observer has an equivalent to what the demonstrator produces. Learning word forms becomes a challenge of this type when speech sound reproduction replaces whole-word re-creation.

---

[144] I should acknowledge again the precedence of Yoshikawa et al. (2003). However, I came across their work after independently developing the ideas that we share. I was motivated by Gattegno's (e.g. 1973:4) argument that children do not learn speech sound qualities by imitation and inspired by his practical demonstration of an alternative in the Silent Way.

However, prior to this, the equivalences have to be established: the observer has to learn to imitate (to solve the so-called 'correspondence problem'). If the sounds produced by mother and child are perceptually transparent to the child this may be straightforward: a 'matching-to-target' process of increasingly good approximation should be possible. If the signals are perceptually opaque this is not possible, but an alternative is for the child to be given a 'mirror' to inform him of the results of his actions.

In chapter 13, I looked in more detail at mimicry and imitation (or 're-enactment', which may be a more suitable term for the imitation of sounds). I elaborated on an earlier assertion that they are very distinct processes, and asked how suitable either could be as the basis for learning speech sounds.

In mimicry, a sound image is used to drive the auditory inverse model that the speaker started to develop in his earliest vocal play. It is, of course, possible to re-create sounds and sound patterns this way, but the process requires the speaker's attention – the sound image must be evoked or kept 'in mind' to some extent - and it is therefore not a suitable basis for speech.

In re-enactment, a new IM is created in which each speech sound is represented as a lightweight token which mediates between what is heard and the movements needed to reproduce it. (The nature of the token is controversial, of course.) In leading accounts, this equivalence is determined by the child making a judgment of similarity between his output and that produced by others.

However, there are a range of possible problems with this mechanism, which would make any signal conceived as speech sounds effectively opaque to a child. These include the following:

1. While he can perceive the sensory patterning of sounds and words through AS mode and thus attempt to re-create them, he may not be able to perceive his mother's speech sounds in this way. To understand speech and to recognise a speech sound within it, he must attend to the signal as informing him of meaningful events, in MP mode. But as speech is ephemeral, the opportunity to then attend to it as a sensory experience is gone.

2. He may not hear himself adequately because of interference from bone-conduction of sound.
3. Or alternatively, because he 'hears' what he intends to produce, not the sidetone.
4. He may not be able to compare the sound images captured.

If any of these are valid then such a judgment of similarity would be impossible. (These problems might affect vocoids more acutely than contoids.)

In chapter 14, I presented an alternative mechanism, by which the correspondence problem is not solved by the child's judgment of acoustic similarity between (normalised) speech sounds. Instead, his mother makes the judgment of equivalence on his behalf, initially during the pre-linguistic imitative exchanges where she reformulates the output of his VMS's into her L1 speech sounds. Because he knows that she is taking what he does to be equivalent to the speech sounds she produces in response, he can associate the two directly.

I borrowed the terminology of a pushmi-pullyu representation (PPR) from Millikan (1996/2005) to describe this movement-to-sound mapping. It is established without the need for a mediating token because the child has only to realise that his mother has taken the two to be equivalent. (This equivalence is, in fact, the only requirement for a speech sound. Similarity, *per se*, is not needed.) There is no need for a normalisation problem to be solved.

Word adoption may have begun with the re-creation of whole- or part-word acoustic images. But as the disadvantage of this becomes apparent, the child discovers that he can use his PPR's in reverse on words, as he did, of course, during pre-linguistic imitative play. This gives him a way of reproducing words that can become fully automatic.

Note that this is based on the child marking those speech sounds in his mother's words that he believes he can reproduce. This is a different type of listening from the ordinary-noticing which has supported (and will continue to support) word recognition. I called the marking operation 'parsing for production'. It supports a phonological word production lexicon that is separate from the phonological word recognition lexicon.

Over the course of speech development, new speech sounds are developed for the speech sound IM, and existing ones refined. These both occur through the child first experimenting with his production apparatus, and the motivation for both can come from simple reinforcement as can the child's judgment of his success (i.e. through meeting his listeners requirements). In due course he may also end any differentiation/refinement process by a judgment of similarity made by him, but depending on the sound type, only to the extent that this may gradually become possible.

In chapter 15, I explored the implications of a PPR that directly associates (1) the category of speech sound produced by others and (2) the movement produced by the child as equivalent to it. The structure of this unit, at once both auditory and motor, explains otherwise puzzling data on shadowing latencies. It may also enable a resolution of the longstanding issue of whether speech is best characterised as gestures make audible or as an acoustic code. The underlying representation of a speech sound would, in fact, be both; a simple brain structure having been created by a complex learning process, in contrast to previous theoretical proposals where the simple learning process of imitation has failed to generate representations that can account for all the complexity of the data.

Finally, I described an integrated developmental model of perception and production, incorporating the ideas above. Simple explanations emerged from this for various problems, including the *fis* phenomenon and the apparent precedence of production over perception in some L2 learners.

The table below summarise the distinctions I am making between the two possible routes to word production. The drawing and handwriting of alphabetic letters is a parallel in some of their aspects.

	Auditory inverse model	Speech sound inverse model
Creation	From earliest vocal play onwards, by self-made judgments of signal resemblance	From evidence of equivalence, e.g. during reformulation and other mirrored interactions
Input	Signal is attended to in AS mode (as a sensory pattern)	Signal is attended to in MP mode, and speech sounds identified
Output	Conceived as a sound (re-created via a movement)	Conceived as a movement (a VMS which creates a sound)
Inner representation?	No. The model is driven by a sound image which has to be "in mind" (evoked)	A pushmi-pullyu representation: a direct speech sound to movement association

Figure 16-1. Routes to word production. The speech sound inverse model allows for greater automaticity, so is preferred.

16.3 Afterword: teaching pronunciation

I came to consider how children learn to speak from my experience learning French and Japanese and my attempts to teach English as a foreign language.

If the proposals in this thesis are broadly correct, then I hope that one practical result will be changes in the way that pronunciation is taught in language classrooms. The copying mechanism has been and continues to be thoroughly tested; and clearly does not work[145]. Current practices survive, though, largely because people believe that young children learn to pronounce this way, and hope – in denial of the evidence in front of them – that older learners can, too[146].

---

[145] In Messum (2002) I argued that those students who do succeed in acquiring good pronunciation have probably been wise enough to subvert the process.

[146] In fairness, the other reason current practices survive is because of an assumption that their ineffectiveness is due to L1 perceptual interference. But the usual attempts to deal with such interference – more listening – do not help, and the net result is still failure and frustration.

In the classroom, copying doesn't work for either temporal phenomena or for sound qualities. The reason is expressed concisely by Mason (1994:178):

> "The more explicitly the teacher indicates the behaviour which would arise from understanding, the more likely students are to be able to produce that behaviour without generating it from understanding."

To generate pronunciation behaviour 'from understanding' in an English language class, we need to involve students more in the use of their respiratory system for stress-accent (which may then result in apparent 'stress-timed' rhythm, in tense/lax vowel length variation, in aspirated plosives when appropriate, and so on), and more than we do now in their articulation (for developing speech sounds).

My description of how an infant learns a speech sound inverse model through mirroring interactions is similar to the process Gattegno designed for learning pronunciation in L2 Silent Way classes.

Gattegno's insight that children don't learn sounds by copying was just one of many he had about speech, language and how they are learnt. (And these, in turn, form just part of his pedagogical contribution. He is equally or better known, in fact, for his work in the teaching of mathematics and of reading and writing.) Another result that I hope for from this thesis is a more widespread consideration of Gattegno's ideas about language teaching and about learning in general.

He published three books specifically for language teachers (Gattegno 1972, 1976, 1985) and a treatise on the theoretical aspects of education (Gattegno 1987). Young has written several short introductions to the Silent Way (e.g. Young 1984, 2000) and a full treatment of Gattegno's pedagogical theory (Young 1990). Much of Stevick (1980) is concerned with the Silent Way, although Allard and Young (1990) question his understanding of the approach at that time. Stevick (1990) is a more reliable guide.

Miller and Dollard's (1941:128-133) account of a non-singer learning to sing in tune may provide an instructive parallel. Here the student's initial failure to benefit from attempts at acoustic matching cannot be ascribed to any perceptual interference, and progress only began when the teacher started mirroring him.

I have written two articles about teaching pronunciation that were inspired by Gattegno's ideas (Messum 2002, 2004). Underhill (1994) acknowledges a similar debt, and contains many practical ideas of his own.

Appendices

Appendix A

Kneil (1972) "Subglottal pressures in relation to chest wall movement during selected samples of speech"

Kneil's PhD thesis was one of a series of studies undertaken at the University of Iowa into aerodynamics and the action of the respiratory system during speech. Other investigators included Cooker, Hardy, Kent and Netsell.

Kneil starts by pointing out (p.1) that despite criticism of Stetson's work (e.g. Stetson 1951), there was no alternative theory, "to account for the **generation and control** of subglottic pressures." (Studies had measured airflows and pressures, but these are emergent properties of the system, not directly controlled as such.) Only Cooker (1963) had tried to replicate Stetson's fundamental observations concerning respiratory activity (looking at chest wall (CW) activity in relation to pressures developed), but his study was limited by the recording of intraoral pressure rather than P_{sg}, and by instrumentation that gave limited data at high syllable rates. (p.14)

Two principal theories had been advanced to account for the control of respiratory pressures for speech. Bell's:

> "The prime requisite for speech is a store of compressed air which can be let out little by little, as wanted. It is obvious that the air would escape with a gush unless restrained. The trap doors [vocal folds] constitute the chief means by which a too rapid escape of air is prevented." (Bell 1916:4)

And Stetson's (as described by one of his collaborators):

> "...the air within the chest at the beginning of the phrase is not under pressure. The pressure rises with each syllable pulse of the chest muscles and subsides between the syllables. Thus the air column is moved upward through the trachea not in a continuous stream, but rather in a series of pulsations ... the breathing muscles themselves regulate and control the air flow during speech." (Hudgins 1937:347)

Kneil's own summary of these positions is as follows:

> "[O]ne theory is that a supply of air under pressure is 'valved' to create airflow as needed while the other is that the respiratory bellows is itself a pressure-pulse generating system, functioning in response to syllable demands."

Investigators prior to Kneil had noted that rate of articulation may have played a role in the formulation of these positions:

> "Ladefoged (1960) stated that for Stetson to obtain one pulse per syllable, his subjects must have been 'talking more loudly, slowly and distinctly than is customary.' This inference may, in fact, be partially supported by a careful study of the rates of utterance in Stetson's records. In this regard it is worthwhile to note that most of Stetson's data were generated from trains of syllables or short phrases. In his data both chest wall movements and tracheal pressure pulses are clearly evident at low rates. His records also show that, as syllable rate increased, the pulses in both the chest wall movement and tracheal pressure traces tended to diminish in amplitude and nearly lose their identity as discrete pulses at rates of 3 to 4 per second and above. It may be hypothesized that at the conversational rates of utterance used by Ladefoged's subjects (4-5 syllables/sec; Ladefoged, 1963, Fig. 2) little or no evidence of discrete pulses would remain." (p.8)

Indeed, Stetson's 1951 records showed that at 3 syll/sec or greater the tracheal air pressure did not return to baseline between syllables. Instead, an elevated background pressure was maintained.

Further, Cooker (1963) had found clear evidence of CW activity preceding syllables at low rates, but concluded that above 2 syll/sec the observed motion of the CW was best accounted for as the result of back pressures. (Kneil's results in fact suggest otherwise, p.188.)

Kneil set out to replicate Stetson's work using similar observational techniques, but using improved instruments and a focus on rate variation to discover, *inter alia*, if there are two modes of respiratory system activity, at low and high rates. Following Stetson he measured CW activity by a cup placed on the epigastric region, high on the abdomen in the midline just below the level of the xiphoid process. Within this cup, pressure changed as the "underlying musculature firms". He obtained P_{sg} readings by tracheal puncture.

He presents data from 4 subjects. Most of this was obtained from trains of 4 different syllable types, chosen to reflect various aspects of speech production:

/ʌ/ With only one closure of the vocal tract, like Stetson's "OVO" syllables.

/pʌ/ Stetson had noted that higher rates are possible with an initial C than without.

/hʌ/ Cooker (1963) had found differences between syllables starting with /h/ and those starting with other C's.

/s/ Stetson viewed /s/ as a syllable when produced in isolation; also, the vocal folds should be abducted during its production, and since there is no complete closure of the vocal tract it may have provided a clear view of pulses, without confusion from back pressure effects.

The subjects were asked to produce syllable trains at fixed rates of 1/s, 2/s and 4/s, and at rates increasing during a single utterance from 1/s to the speaker's maximum rate and rates decreasing from their maximum rate to 1/s.

Kneil reports subglottal pressures as ΔP_s, the rise from the intersyllabic low to the syllabic high, and $P_{s(b)}$, the background pressure developed (the baseline for measurement of ΔP_s). The sum of the two gives the actual P_{sg}.

The data was too variable and there were too few subjects for extensive statistics to be derived. Instead, Kneil presents graphical displays of his results by type of syllable and by subject. See, for example, my figure 3-2, where the reversal of the positions of the open and filled symbols indicates the changeover from one mode of production to another. Kneil describes these modes (or 'styles' of SB) as 'pulsatile' and 'elevated/developed background pressure' (EBP).

In general, any changeover from one mode to another in the syllable trains depended upon rate, speaker, type of syllable and the starting mode (a fast or slow rate).

Other things being equal:

The higher the rate the more likely it was for a speaker to change to the EBP mode. This happened most often at rates somewhere between 2 and 4 syllables per second.

Speakers were far from uniform in their styles of production. For example, in contrast to the data I have reproduced above, speaker DN's data for variable rates of articulation showed no use at all of background pressure, even at rates approaching 6/sec, except for the /pʌ/ syllable type.

The rate at which a change of mode occurred depended on syllable type. For /s/ it happened at lower rates than for /pʌ/ and /ʌ/, which in turn changed at lower rates than

/hʌ/. Indeed, for /hʌ/ there were only preliminary signs of a change from pulsatile to EBP modes at the highest rate of 4 syll/sec and for just 2 of the 4 speakers whose data was presented. /hʌ/ is a high flow segment, and Kneil hypothesises that this is the reason for the difference:

> "The demands of airflow during /h/ seem to preclude or restrict the development of a background Ps at least in trains of repeated utterances of /hʌ/. The relatively high airflow would appear to place more of a demand on the respiratory system as a driving force than in the case of /s/. The [chest wall] trace generally shows greater deflection during /hʌ/ than during /s/ which may reflect that demand." (p. 123)

The mode of production that speakers started with affected the point at which a change was seen in the increasing and decreasing rate conditions. As Kneil says,

> "... it appears that it is preferable to operate the speech-respiratory system with a developed background pressure and that a change toward that mode of operation typically occurs as soon as rate (mean resistance over time) allows. However, when the system is operating in that mode and rate decreases, the tendency is to delay shifting to a 'pulse' type of operation." (p. 161)

Kneil elaborates on the reasons for the findings in point 3, above, explaining the articulatory significance of the segments in question:

> "Utterance of a series of /s/ syllables requires the formation of an oral constriction which remains essentially constant as the syllables are spoken. Though phonation is not involved, the vocal folds appear to operate in a valving fashion in order to control the air stream. The data in Figure 3 at the rate of 4/sec demonstrate pressure **in the oral cavity** *[intraoral pressure was being measured separately - PM]* that rises for each utterance and falls in the intersyllabic interval. Since P_s is relatively constant, the glottis must valve the airstream to produce the P_o pattern. As the glottis is not otherwise engaged, as in phonation or the generation of turbulent flow, it is free to perform such valving. The glottograph trace tends to confirm this observation. Since it appeared that a rapid, relatively inaudible glottal valving of the airstream was occurring during the series of /s/ productions, two subjects were asked to produce a train of /s/ utterances at a high rate followed by a train of /ps/ utterances at a similar rate. The glottograph traces for these trials indicated that glottal valving was prominent during the series of /s/ production but was not observed when /ps/ was uttered. This suggests that the subglottal pressure, which was comparable for both syllable trains, was controlled glottally for the /s/ since valving by any other articulator would have introduced unwanted sounds or modifications of the /s/ sound. In the case of /ps/ on the other hand, valving at the lips was required to articulate the sound combination and the additional control at the glottis was not required. This appears to be **evidence of a trading relationship wherein the effective resistances (to maintain $P_{s(b)}$) trade between the bilabial closure and the lingua-alveolar constriction, maintaining a mean resistance to air flow over time.**" (p.120)

> "The generation of /hʌ/, with two phonetic elements, has different requirements. For the /h/ there is relatively high airflow through a partially constricted glottis. Although the

glottis is sufficiently constricted to generate turbulent airflow, its resistance is less than the resistance of the oral constriction for /s/ and the airflow is correspondingly greater. The glottis must alternate between this adjustment and one that is appropriate for phonation for the vowel. The vowel adjustment requires greater adduction, increased resistance and reduced airflow. Since there is no oral constriction, the total resistance to airflow is a time-varying one at the glottis. ...

The remaining two syllables, /ʌ/ and /pʌ/ seem to fall somewhere between /hʌ/ and /s/, at least insofar as Ps(b) is concerned." (p.123)

For /ʌ/, some subjects were able to exercise precise and balanced control of the vocal folds, so that discrete utterances at up to 5/sec were achieved. Two subjects seemed to achieve discrete syllables by adduction of the vocal folds, two using the respiratory system as the controlling device. /pʌ/ allowed an easier maintenance of a constant background pressure.

On the relationship between chest wall movement and subglottal pressure, Kneil reports:

"It is evident that for discrete syllables at low rates of utterance, discrete actions of the chest wall and the respiratory system will occur if the system is permitted to relax between syllables. Under such circumstances one would expect a discrete movement of the CW to accompany a positive pulse in P_s. On the other hand, if an elevated $P_{s(b)}$ is maintained at high syllable rates, and variations of airflow are controlled completely by glottal and articulatory valving, little activity would be expected at the chest wall. This suggests that a fairly close relationship between the amplitude of CWM and the amplitude of ΔP_s should exist.

The data appear to confirm the relationship posited above. ... in general the amplitude of the chest wall deflection covaries with the degree of change in P_s. As with the data reported earlier, it is apparent that differences exist between syllables and, more clearly, between subjects." (p.161)

Kneil makes the following points among his conclusions:

"(1) An elevated subglottal pressure level develops and is maintained as the rate of syllable repetition increases. This elevated pressure level forms a base or background from which increases associated with syllables may be made. **The way in which the pressure is developed and the degree of its development varies among subjects and syllable types.**
(2) As a background subglottal pressure is developed, the amplitude of the pulse-like changes in the subglottal pressure diminish. At high syllable rates the record of subglottal pressure is relatively smooth.
(3) As a background subglottal pressure is developed, the movements of the chest wall that are associated with syllable production diminish in amplitude.
(4) With the development of an elevated background subglottal pressure, the egressive airstream is valved by the articulators as the means of control of the flow and the pressure. Normally, such valving occurs at all conversational rates of utterance.
(5) The glottis plays an important role as an articulator in valving the airstream under

certain circumstances. Such valving is relatively inaudible during conversational speech. ...

It appears that **neither Bell nor Stetson were wholly correct in their views of the operation of the respiratory system for speech; neither were they wholly wrong.** The system does engage in discrete muscular contractions leading to subglottal pressure increases for isolated or very low rate syllables wherein the system relaxes between utterances. At high syllable rates and during speech at normal conversational rates, the subglottal air is under pressure and the egressive airflow is controlled by valving. Such valving is the result of all constrictions of the vocal tract as they occur during speech.

In some complex fashion individual chest wall movements may be superimposed on a background of activities so that, when needed, pulse-like pressure increases can take place as for stress." (p.197)

Appendix B

Caleb Gattegno (1911-1988)

In chapter 12, I mentioned some of Gattegno's proposals on how infants and children learn to speak and in Appendix C I reproduce one of his own summaries. However, his ideas have been fundamental to my work in other ways, two of which I would like to acknowledge here.

Firstly, he alerted me to the idea that proper production of the phonetic differences between languages may require more than just new uses of the upper articulators. He described how he realised this in Gattegno (1998:39):

> "En 1947, Marcault[147] m'a fait faire incidemment une découverte. Il était venu à un de mes séminaires, à Marly - c'était un rassemblement international; il y avait même des gens des universités anglaises. Il y avait environ soixantes personnes qui l'écoutaient, qui ne savaient pas toutes le français. J'étais assis et lui debout (il préférait parler debout). Comme il était sourd, je lui glissais alternativement un papier écrit en français puis en anglais; il parlait donc alternativement en français puis en anglais. Comme j'étais à côté de lui, j'ai remarqué une chose extraordinaire: quand il changeait de langue, il se produisait un grand bruit (qu'il n'entendait pas) venant de lui...
>
> [C'est la première fois que j'ai eu l'intuition que parler différentes langues consistait, pour la volonté, à commander un agencement sur le plan somatique; et je l'ai entendu: il y avait un agencement dans l'espace et le temps avec un bruit considérable - sans cela c'était une théorie. A ce moment-là, j'ai su qu'il y avait une voie de recherche, un très gros problème qui valait la peine d'être étudié.]
>
> Chaque fois qu'il changeait de langue, il s'arrêtait et repartait avec un grand bruit, comme s'il y avait eu un réagencement des vertèbres pour donner un nouveau tuyau d'orgue. Si je ne l'avais pas vécu, je ne l'aurais pas cru. Mais cela se passait à cinquante centimètres de moi. Il était clair que, parce qu'il était sourd, il ne contrôlait plus... C'est sans doute la présence de l'ouïe qui, par un procédé encore mystérieux, agit sur le soma et fait que celui-ci fasse ce travail en silence, avec, disons, un lubrifiant. Chez Marcault, il n'y avait plus de lubrifiant, et l'on entendait le bruit de charnières."

In "Teaching Foreign Languages in Schools: the Silent Way" (1963/1972), Gattegno applies this idea to learning how to speak a language:

> "It seems to me that the practice of breathing in a certain way, and the use and practice of what I want to call the functional vocabulary will, to a certain extent, provide classroom equivalents of what is picked up naturally in an environment where the

[147] Jean-Emile Marcault was a French philosopher whose ideas and book, "L'Education de Demain" (written with Thérèse Brosse and first published in 1939), greatly influenced Gattegno.

language is normally spoken.

> Though it is obvious to all of us that we speak our own tongue differently when we are relaxed or tired, out of breath or in control of our flow of words, it rarely occurs to linguists to consider whether some languages present special features because of the demands they normally make on the pneumatic systems of their users. A language like English, in which so many short words are commonly used, gives much more often than German the occasion for short pauses, thus making the English more inclined to speak slowly and mutedly, and the Germans more inclined to embark with vigor upon what sounds like a speech on the most ordinary matter. This is inevitable because of their way of describing things by linking a number of words into one. Because Germans place a key word, their verb, at the end of a sentence, no matter how many clauses are inserted in it, German speakers tend to race to the end of every statement to convey their meaning. The breathing requirements of their language are thus different from those of English.
>
> So the spirit of a language can be reached still better if learners are made aware of the breathing requirements of different languages. This meets the melodic component somewhere, since the line of a melody, as distinct from its notes, is concerned with time factors, which in turn are connected with the qualities of breathing if the voice is the instrument or part of it.
>
> This has been a very brief and sketchy consideration of breathing as a way to the spirit of languages. To discuss it in greater detail would take us too far in a book of this kind, though I believe the problem to be of fundamental importance."

Reading this passage at the time that I was learning about tense and lax vowels and other curiosities of English phonetics, lead me to the idea that the WGmPh might be connected with a particular style of speech breathing used by speakers of these languages.

A second essential support for me has been Gattegno's theory and descriptions of learning, as developed in many articles and books over his lifetime[148]. He insisted that learning takes place through awareness and could be described in terms of awarenesses[149]. This gave me a tool to investigate both imitative accounts of learning to talk and the alternatives I considered.

He invited us to consider all aspects of life and living as energy transformations in time, and provided numerous example of human activities so described. He generated many insights in this way, including many connected with language.

---

[148] A bibliography of Gattegno's published work is available at http://www.cuisenaire.co.uk/gattegno/bibliog.htm (last checked June 2006).

[149] Unfortunately English lacks an equivalent to the French *prise de conscience*. The use of 'awareness' as a countable noun seems the best way to remedy this omission. Awarenesses produce internal energy changes that range from the barely perceptible to the almost overwhelming (for example, the 'eureka' moment). See also Mason's analysis of noticing, reported in section 10.1.

Finally, Gattegno made the following remarks in what turned out to be his final address to the Association of Teachers of Mathematics, which he had helped to found in the early 1950's. His final question is one he repeatedly asked himself, with fruitful results:

> "Have you ever noticed that children learn to speak their mother tongue by themselves? And that you are evading questions in saying, 'They do it by imitation.' 'By imitation,' indeed. The greatest nonsense I ever heard, and everybody repeats it. It's absolutely wrong. No-one can learn to speak the mother tongue by imitation. So, you have to ask the question: how did we - because we were babies - how did we learn our mother tongue? What sort of powers of the mind did we have to sort these things out by ourselves?" (Gattegno 1989)

Appendix C

Gattegno (1985:6-21), extract from "The learning and teaching of foreign languages"

Talking

Let us agree that this word will be used in this writing to cover what any baby does in the field of sound production as well as *what he can do entirely by himself* and for which he does not need anyone else of the environment. All that for which he must accept full responsibility to himself.

Our idea that babies are helpless should not extend to areas where they alone have entry and can take initiative knowingly.

The common idea that words are the stuff of verbal expression should be replaced by one in which the reality contemplated is truly described by us; and the study of distribution of energy over time takes much better care of that. In fact, everybody knows that *in the beginning there are no words* and that that beginning lasts a number of months for all of us, generally more than 9 months.

Our idea that babies hear words first and try to reproduce them is utterly non-factual. Hearing needs to be educated before it can function on words and each baby does this work of education, no one else does it. For no one else would know how to do it.

A very important observation needs to be made from the start: *our phonation system is voluntary and our hearing is not.*

Another important observation is that we have been told - by our ancestors who selected some words in all (?) languages to indicate their awareness of themselves - that "looking" and "listening" are functions of the self in contrast to "seeing" and "hearing" which conveys that what was without has been allowed to reach us and be held within. The first two are deliberate, the last two relative and non-compulsory, as is immediately and easily ascertained by referring oneself to seeing beauty and hearing harmony. We may be taught to see or hear but not to look or listen. This we must do through the movement of our own will.

Each of us, in the intimate non-verbal language of awareness, has reached this conclusion about him or herself and the world, as soon as our sensory nerves have been allowed to be myelinized, three or four weeks after birth.

Before leaving these generalities and concentrating on "talking," let us ask our readers to connect with their own critical intelligence and put it to work at every point of this pinpointed study of how babies work using their mental and other powers, and which we all take for granted because they are universally available and, smoothly and effortlessly, take people to a place from which they display commonly visible and noticeable behaviors.

In my crib I was left alone so long as I did not cry for help.

No one had any reason to ask what I was doing with my time when I was calm and satisfied, dry and comfortable.

Still I had my time to my self. I could open my eyes and move them. I could watch my breathing and play with the flow of air my voluntary chest muscles could allow me to vary. I could be present in any one of my voluntary muscles and study the variations of the muscle tone which holds them ready and permit myself to concentrate on obtaining what I can perceive as being different or the same. For instance, that I can shut or open my eyes at will; that I can hold my head up; that I can turn my head, etc. etc.

What no one can see is that I note every one of my new awarenesses and that increases my experience of myself by letting the new be known and letting it integrate the old. That there is so much to do that I need much time for it by exercising my learning powers *all that time*. My waking time was needed to relate to the non-self while my sleep took me back to my self to sort out my retentions of the day and integrate what these were with what was already there. That in sleep I grew, and in wakefulness, experimented.

Thus I learned how to handle inner energy shifts and impacts of energy, knowing which was which.

For our purpose of understanding "talking" we do not need to be in contact with what all babies do with themselves during those many days and hours of early childhood.[150] It may suffice that we be alerted to the reality of learning in both sleep and wakefulness, at the preverbal stages.

The acquisition of L1, is based on learnings which have their roots not in the environment but in the self at work. Until this is clear all efforts will only yield little understanding and even be a waste of time. Investigators of the acquisition of L1, must, themselves, learn to work with the energies within and their various dynamics. Before we can enter into "speaking" we must master the processes which go with the awarenesses of the distribution of energy over time just as we did when we were babies.

Energy is the ultimate reality in the cosmos, which includes us.

Variations of energy generate the possibility of *the awareness of time*. This in turn places at our disposal a receptacle in which we can make sense of temporal hierarchies of energy uses which represent our experiments with energy. Thus if I can become aware almost from the start of all the possibilities of the muscle tone of my lips by affecting it deliberately over a certain duration and a succession of durations, I must wait till my teeth break through my gums, to work on how those affect the flow of air I put off. If I can affect my vocal cords and note how their state (a result of ordering their muscle tone to be thus or thus) affects the flow of air, I will need this awareness as an instrument before I can combine that knowledge with a study of the conjunction of my acting simultaneously on my lips. My self elaborates instruments of study and then studies how to bring these instruments together to achieve more or something else, in analogy with the cosmos first making hydrogen and oxygen and with them later generating water. The latter is a possibility of the existence of the first but does not take their place; all can co-exist in spite of the temporal hierarchy establishing which came first to make the last exist.

---

[150] cf. our monograph "The Universe of Babies" Educational Solutions Inc. New York City, 1973

Babies do all this work knowingly and deliberately, thus making us understand why babyhood is needed, why we are small and at work on what only benefits ourselves. All our elaboration is within and absorbs all our time.

Unless we use the lighting of awareness we cannot come to grips with the challenges raised here. As we said earlier, learning is equivalent to living, *to consuming our time (which is our wealth) to give ourself experience* which stays with us either in the form of objectifications or of knowhows which are dynamic. All this clearly is or concerns energy. As energy systems we have no trouble accounting for our growth which is energy added and taken from the store found in the universe around.

* * *

In the temporal hierarchies we must place first the acquaintance of the self with its sound production system. This, because:

a. it is at hand,

b. is permeable to the will which can give commands in the form of altered muscle tones, and then

c. we can become aware of the consequences.

Breathing takes air through the larynx. Crying is one of its uses. Every baby may learn to modulate its crying, to prolong it and to stop it. Thus from the start we dwell in our throat and control it as well as the flow of air. We quickly learn how to make it loud or less so, finding out the exact mechanisms of that and therefore own it fully as well for any other purpose of ours which is not crying.

The important point here is to grant to babies their ownership of such a presence, conscious presence, in the various organs composing the phonation system needed to dedicate each part to the formation of specific and determined wholes integrating those parts. For instance, acting on one's tongue and one's larynx at the same time and studying their respective contributions to final products which encompass each of them separately and all the intermediate mixings of both. When lips are added, when the walls made of the cheeks, the palate and later the teeth, are called in, it becomes obvious

that a whole spectrum of complex sound productions are available to every child who can then play variations on them. These are gratuitous combinations produced by the self mainly for its acquaintance with a given somatic system generated *in utero* and whose possibilities can only be known and assessed *ex-utero* when air can flow through the organs under variable conditions proposed by the self for that study and ending in a thorough acquaintance.

Endless hours are spent by the young child not yet 10 weeks old, say, to make sure that learning has taken place, i.e. that a mastery has been achieved which might make possible other conquests than sound production. Vowels are produced first. But for each a baby has to find exactly how it is made in terms of quantities of energy poured (or simply added) in the relevant muscles in order to remake them exactly, thus leading to an awareness of *sameness*. Once this is attained a baby can act on the duration of the utterance and how the somatic meaning of that sound shortened or prolonged, reproduced continuously or staccato. An alternative, open to all, is to produce a different sound and recognize it as such by the amount of energy affecting the various muscle tones of the muscles involved. As soon as two sounds are known for what they are from within and hence how they differ, a possible exercise is the production of sequences of the two, intermingled in various ways. The "algebra" present is acknowledged and leads to the awareness that say *ai* differs from *ia*. that *aai* and *iaa* are not the same, but that *iai* is unique and remains the same by "reversal."

As many vowels as wanted can be produced and such combinations and permutations explored without effort or tuition. Rather every child can teach himself all this without consultation with anyone.

Of course, a special awareness is needed for a child to hear his utterances and know them as his. Because it is easy it is done quite early, mainly because the energy of an utterance can affect one's eardrums. But in its nature it is quite different from the awareness of the sound production: only a fraction of the energy of an utterance reaches the ears. Attributes of each utterance need be kept in mind to ensure recognition. This a baby does by reproducing the same sound a number of times and affecting the productions with distinguishable properties such as durations and permutations and concluding with certainty that what he hears is what he himself does.

Educating one's hearing requires the presence of the self in both one's throat and one's ear and molding hearing upon uttering until such time as the mere evocation of an utterance triggers the evocation of what is being heard. From then on the self can use hearing as a monitoring system of his utterances and give his hearing all the knowhow acquired through the effects of the will on the voluntary system in his mouth and consider this a transfer of awareness.

Of course, this education will continue and even be available all one's life (in particular, when acquiring an L2). It still remains that knowing directly through awareness of muscle tone variations is not replaceable by dwelling in one's ear, which is a transferred knowing.

Beside the ears, the whole bony structure in the head is also affected by the energy of the utterances emitted and a baby is aware of that, so that to reach pure hearing a movement of abstraction is needed. Most grown-up babies are unable to recognize their taped voice when this is audiotaped simply because the harmonies added by the echo chambers in one's head are missing on the tape. While we can all recognize other people's voices all the time we do not recognize ours until we work on this new demand and complement *mentally* what we actually hear. Our awareness through our throats knows a different voice from that gotten by our ears, at least in the single case of ourselves, thus stressing the profound difference of knowing language by ear and by uttering it.

The fact that musical melodies are also energy distributions over time but will reach first our ears before we attempt to voice them, while suggesting opposite connections, will help us understand the connections between uttering and hearing and hearing and uttering. When uttering, the self does two things: it is related to its intention, its project, *and* to the ordering of the relevant muscles to produce the equivalent of the project. In hearing, on the other hand it must surrender to the incoming flow of energies and their variations, letting them reach the ear and produce awarenesses which can be retained as objectivations or as time endowed with energy. A melody listened to, i.e. with a presence in one's ears to let the energy in, molds duration and thus can be retained as it affects the substance of our energy system of which our brain is a part. It is energy affecting energy. There is nothing to remember but all to be retained. No request to test its being able to be recalled, but a penetration of our substance which recognizes the

impact as such and can recall that addition of energy. If there is something to be remembered it would be the circumstances of the addition. Otherwise it is known as one's own but not as one's creation, as received and held, as integrated, i.e. making one with oneself, but also capable of being separated and contemplated by the self. Thus awareness can work on itself and know its working and workings. In particular, note the nature of utterances and that of sounds heard and attribute the first to oneself or others and the second to others or to oneself. The details as long as they are energy distributed over time can be evoked, i.e. brought back to awareness as a delineation of the actual energy received and thus can be re-lived either actually in real time or virtually in telescoped time. A whole melody can be triggered by two or three of its notes strung together in a time sequence even though it is not actualized in real time. The connection of the self to sounds received is of the same kind whether the sounds have been uttered by oneself as by others. No effort to retain is needed or implied.

As the uttering baby knows which are the additional energies contributed to sounds by stress or intonation, he will know how to blend or to separate these components on a packet received by his ears from himself or from others. Exercises to ensure these awarenesses so that there is no doubt that some energies are acknowledged as either sounds or stresses or silences or intonations, will provide babies with the necessary experience to relate intimately and immediately with flows of sounds produced first by themselves and later by others.

An educated ear is necessary for a serious engagement in the study of the attributes of what one hears and babies work on such an education by themselves in close contact with the challenges, since no one else can help. Hence, they give themselves not only the instruments needed but also the vulnerability which makes them specially sensitive to the minute and subtle energy changes over duration, which can be short or long.

All this of course does not show but we can accept that it is objectively there if we ask:

a. how could one do a number of these things otherwise?

b. who could help a baby do it? (and find no one)

c. how could we account otherwise for the fact that when things become tangible and observable, that they are there?

The mere fact that in the beginning there are no words and that genes cannot account for the delay of months before babies discover speech in the environment - and take a number of months to learn to speak - force us to look for deeper mechanisms in the mind and for certain kinds of experiences akin to the final results: a willed flow of words which are carried by one's voice which has a certain number of attributes that are individual and idiosyncratic and others which are collective and communal.

The first are taken care of by all the work done by babies under what we called learning to talk to which we must add the melodies of the spoken languages around the babies which, like those of music, are energy distributions over time, accessible to every child individually but resulting in a group carrying it.

As to the second, we find in the reality of the specific spoken language in a given environment, attributes directly reachable because they are energy - and sounds, stresses and melody are three of them - and other attributes which are arbitrary and cannot be known directly. These will require special functionings of one's imagination and intelligence, as we shall see when we consider "speaking," revealing new competences of babies overlooked for too long by students of early childhood and in particular, by those concerned with first language acquisition.

Often, perhaps usually or even always (a fact hard to establish) a child who has completed the learnings under "talking," invents his own language which displays all the energy attributes of a language found in his environment except that *none of the units* uttered is a word of the environmental language. The job of speaking clearly cannot be done by the child on his own. All he could invent and produce, he has, but the arbitrariness he used for his personal language does not cover the historic choices for words which his ancestors had reasons to select and hold to, over many generations. Although such children (and they probably are very numerous) could enlighten us on the origin and evolution of spoken languages, they personally have to give up their artistic verbal monument and use all their abilities to acquire the language of their environment. It is still unknown whether they are helped or hindered by such an

abandonment of a fully constituted first language. Clearly for such children, L1 should be called L2 ....

Summing up: "talking" is that part of the acquisition of L1 during which every child specializes a number of learnings which provide him with a vast number of inner criteria which ensure that he hears what he can utter and give him an intimate and immediate acquaintance with the various energy contents of what he hears, this because the learnings were deliberate, willed, acted upon a voluntary system located in the throat and mouth, reachable from within and guided by a vigilant awareness working on energy variations which go on also to form a system. This is normally adequate. Deaf people do not have access to it and do not form it although they are able to do many things (paralleling "talking") when talking does not require hearing.

From "talking" to "speaking"

Since babies are surrounded by an environment, generally a speaking one, and since they end up speaking like people in it, it has generally been accepted that children imitate speakers and learn in that way.

We shall not spend much time on this not very useful approach to the challenge. Since our understanding of the challenge has been presented in the section on "talking", readers will relate to our challenge differently.

What is new in "speaking" is that children are confronted with the arbitrariness of the lexicon of the language around them. Being wise and expert they know that the reality of spoken words is only in the energy of the sounds in them and in the stresses, phrasings and melody which appear when words are polysyllabic, run together and used in sets. Words have *no* meaning of their own - as everyone will make sure by listening to someone speaking an unknown language - babies look for what makes sense to them and find that intonation conveys information about the speaker and his or her emotional states. Intonation being a human component of languages and not a linguistic one, brings to the listener direct human information which later will be organized intellectually under categories such as: sadness, irritation, joy, satisfaction, doubt, bewilderment, etc.

There is need for a bridge between "talking" and "speaking" and to find it we need to return to babies in their cribs entertaining their production of sounds. In that area, they are consciously engaged in working out how to generate syllables which contain the vowels they already know and "con-sonants" which correspond to configurations of their mouths and can only be uttered blended with vowels (when the consonant is not a sibilant), whether these precede the con-sonant or follow it. So one day when a baby is producing one of the syllables *ma, pa* or some similar pair which when repeated seems to be a word of that language (like mama or papa or dad) and if someone other than the baby monitoring himself, hears him, it is legitimate for that person who then calls the baby's parents to tell them "X is calling you." As the environment fusses over such occurrence the baby can make the observation not yet made, that he can hear, made by others, what he can make himself and then he can concentrate on finding in their flow of words these bits he can attempt by himself.

That certainly constitutes a bridge.

Now the baby has a reason for listening to the flow of words of others and to explore it for what he knows, while when he is alone he goes on with his own projects unconcerned with what others do. As his parents are prepared to echo his productions, every day more, he finds, as a reality, that they too can utter understandable sounds which for him present meaning as to their energy content.

Another proper discovery at hand is that in that field of flow of sounds *there are consistencies*. These are found in himself when he produces the same sounds or uses the same "algebras" denoted by the operations of *substituting* one sound for another, of *addition*, of adding a sound at either or both ends of a sound unit, of *reversing* a sequence of sounds, or of *inserting* a sound within a string of other sounds. But also he finds that independently of the energy qualities of voices, like pitch, timber, intensity, other people's voices carry something which will become more and more a reality provided this child uses a power of his mind called abstraction already known to him and used by him in so many learnings.

Abstraction (or the simultaneous stressing and ignoring of components in a perceptible situation) is certainly needed in order to give reality to that small proportion of the total

amount of energy carried in a sounded word, which is the actual energy of a word, i.e. that of the component sounds in it and of the stressed vowel if ever.

Babies spend time now peeling words out of voices and when they meet sounds they themselves can make they keep them within one category: that of those "common to me and them." It is the time thus spent which will become a solid bridge, but is also a beachhead for the conquest of "speaking."

All this is, of course, only a beginning in the detailed study of how each of us manages to shift from knowing himself and what happens within, to knowing what others do in the field of language with a process which has a history and has evolved collectively over generations. But this beginning is promising and has allowed us to speak intelligently of a universe into which we were not able or allowed to enter using only linguistic concepts and instruments.

* * *

We need now to consider the approaches open to babies for the conquest of "speaking" the environmental tongue.

References

Abercrombie, D. 1965. Parameters and phonemes. In *Studies in phonetics and linguistics* (London: OUP).

Abercrombie, D. 1967. The analysis of speech. In *Elements of General Phonetics*, 34-41 (Edinburgh: EUP).

Abramson, A. S. and N. Ren. 1990. Distinctive vowel length: duration vs. spectrum in Thai. *Journal of Phonetics* 18:79-92.

Adams, C. 1979. A physiological view of stress patterning and pause placement. In *English Speech Rhythm and the Foreign Learner* (The Hague: Mouton).

Adi-Japha, E. and N.H. Freeman. 2001. Development of differentiation between writing and drawing systems. *Developmental Psychology* 37 (1):101-114.

Aldridge, M. A., R. D. Stillman, and T. G. R. Bower. 2001. Newborn categorization of vowel-like sounds. *Developmental Science* 4 (2):220-232.

Alissandrakis, A., C. L. Nehaniv, and K. Dautenhahn. 2002. Imitation with ALICE: learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man, and Cybernetics: Part A: Systems and Humans* 32 (4):482-495.

Alku, P., E. Vilkman, and A. M. Laukkanen. 1998. Parameterization of the voice source by combining spectral decay and amplitude features of the glottal flow. *Journal of Speech, Language and Hearing Research* 41:990-1002.

Allard, F. and R. Young. 1990. The Silent Way. *The Language Teacher* 14 (6):27-29.

Allen, G. D. and S. Hawkins. 1980. Phonological Rhythm: Definition and Development. In *Child Phonology: Vol 1, Production* edited by Yeni-Komshian, G., J. Kavanagh, and C. A. Ferguson, 227-255 (New York: Academic Press).

Allen, G. D. 1985. How the young French child avoids the pre-voicing problem for word-initial voiced stops. *Journal of Child Language* 12:37-46.

Allen, J. S. and J. L. Miller. 1999. Contextual influences on the internal structure of phonetic categories: a distinction between lexical status and speaking rate. *Journal of the Acoustical Society of America* 106 (4):2242.

Allen, J. S., J. L. Miller, and D. DeSteno. 2003. Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 113 (1):544-552.

Anderson, S. 1981. Why phonology isn't 'natural'. *Linguistic Inquiry* 12:493-539.

Anisfeld, M., G. Turkewitz, and S. A. Rose. 2001. No compelling evidence that newborns imitate oral gestures. *Infancy* 2 (1):111-122.

Arbib, M. 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *The Behavioral and Brain Sciences* 28:105-167.

Arnott, S. R., M. A. Binns, C. L. Grady, and C. Alain. 2004. Assessing the auditory dual-pathway model in humans. *Neuroimage* 22:401-408.

Ashby, F. G. and W. T. Maddox. 2005. Human category learning. *Annual Review of Psychology* 56:149-178.

Bailey, T. M. and K. Plunkett. 2002. Phonological specificity in early words. *Cognitive Development* 17:1265-1282.

Bates, E. 1979. *The Emergence of Symbols.* (New York: Academic Press).

Beck, J. M. 1997. Organic variation of the vocal apparatus. In *The Handbook of Phonetics* edited by Hardcastle, W. H. and J. Laver, 256 (Oxford: Blackwell).

Beckman, M. E. 1986. *Stress and Non-Stress Accent.* (Dordrecht: Foris (Netherlands Phonetics Arch. No. 7)).

Beecher, M. D. and J. M. Burt. 2004. The role of social interaction in bird song learning. *Current Directions in Psychological Science* 13 (6):224-228.

Bell, A. G. 1916. *The mechanism of speech.* (New York: Funk & Wagnalls).

Bernstein Ratner, N. 1984. Phonological rule usage in mother-child speech. *Journal of Phonetics* 12:245-254.

Berry, J. 2004. Control of short lag voice-onset time for voiced English stops. *Journal of the Acoustical Society of America* 115:2465.

Bird, G. and C. Heyes. 2005. Effector-dependent learning by observation of a finger movement sequence. *Journal of Exp Psychology Hum Percept Perform* 31 (2):262-275.

Blake, J. and R. Fink. 1987. Sound-meaning correspondences in babbling. *Journal of Child Language* 14:229-253.

Blakemore, S-J., D. M. Wolpert, and C. D. Frith. 2002. Abnormalities in the awareness of action. *Trends in Cognitive Sciences* 6 (6):237-242.

Blevins, J. 2004. Diachronic phonology. In *Evolutionary Phonology*, 274-275 (CUP).

Bloom, L., L. Hood, and P. Lightbown. 1974. Imitation in language development: If, when and why. *Cognitive Psychology* 6:380-420.
Boë, L. J. 1999. Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis. In *ICPhS99*, 2501-2504 (San Francisco).
Boliek, C. A., T. J. Hixon, P. J. Watson, and W. J. Morgan. 1997. Vocalization and breathing during the second and third years of life. *Journal of Voice* 11 (4):373-390.
Bolinger, D. W. 1981. *Two kinds of vowels, two kinds of rhythm.* (Indiana University Linguistics Club).
Borg, G. 1998. *Borg's Perceived Exertion and Pain Scales.* (Champaign, IL: Human Kinetics).
Boucher, V. and M. Lamontagne. 2001. Effects of speaking rate on the control of vocal fold vibration: clinical implications of active and passive aspects of devoicing. *Journal of Speech, Language and Hearing Research* 44:1005-1014.
Bradlow, A. R., D. Pisoni, R. Akahane-Yamada, and Y. Tohkura. 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America* 101 (4):2299-2310.
Bril, B. and Y. Brenière. 1993. Posture and independent locomotion in early childhood: learning to walk or learning dynamic postural control? In *The development of coordination in infancy* edited by Savelsbergh, G. J. P., 337-358 (Amsterdam: North Holland).
Browman, C. P. and L. Goldstein. 1992. Articulatory phonology: an overview. *Phonetica* 49:155-180.
Bruner, J. S., J. J. Goodnow, and G. A. Austin. 1956. *A Study of Thinking.* (New York: Wiley).
Bucella, R., S. Hassid, R. Beeckmans, A. Soquet, and D. Demolin. 2000. Pression sous-glottique et débit d'air buccal des voyelles en français. In *XXIIIèmes Journées d'Etudes sur la Parole* (Aussois).
Buder, E. H. and C. Stoel-Gammon. 2002. American and Swedish children's acquisition of vowel duration: Effects of vowel identity and final stop voicing. *Journal of the Acoustical Society of America* 111 (4):1854-1864.
Byrne, R. W. and A. E. Russon. 1998. Learning by imitation: a hierarchical approach. *The Behavioral and Brain Sciences* 21:667-721.
Byrne, R. W. 2003. Imitation as behaviour parsing. *Phil.Trans.R.Soc.Lond.Series B* 358:529-536.
Call, J. and M. Carpenter. 2002. Three sources of information in social learning. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 211-228 (Cambridge, MA: MIT Press).
Cambier-Langeveld, T. and A. E. Turk. 1999. A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics* 27:255-280.
Carpenter, M. and J. Call. 2002. The chemistry of social learning. *Developmental Science* 5 (1):22-24.
Carter, A. 1998. You don't say! The phonetic manifestation of unpronounced syllables. *Texas Linguistic Forum* 41:15-23.
Catford, J. and D. Pisoni. 1970. Auditory vs. articulatory training in exotic sounds. *Modern Language Journal* 54:477-481.
Catford, J. C. 1977. *Fundamental Problems in Phonetics.* (Edinburgh University Press).
Catford, J. C. 1985. 'Rest' and 'open transition' in a systemic phonology of English. In *Systemic Perspectives on Discourse Vol 1* edited by Greaves, W. S. and J. D. Benson (Normal NJ: Ablex).
Chambers, D. and D. Reisberg. 1985. Can mental images be ambiguous. *Journal of Experimental Psychology: Human Perception and Performance* 11:317-328.
Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22:129-159.
Cho, T. and P. Ladefoged. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27:207-229.
Chouinard, M. M. and E. V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* 30:637-669.
Clark, E. V. and H. H. Clark. 1977. First sounds in the child's language. In *Psychology and language: an introduction to psycholinguistics*, 375-404 (New York: Harcourt Brace Jovanovich).
Clark, E. V. and B. F. Hecht. 1983. Comprehension, production and language acquisition. *Annual Review of Psychology* 34:325-349.
Clark, E. V. 1993. *The lexicon in acquisition.* (CUP).
Clark, E. V. 2003. *First language acquisition.* (CUP).
Clark, H. H. and S. E. Haviland. 1974. Psychological processes as linguistic explanation. In *Explaining Linguistic Phenomena* edited by Cohen, D., 91-124 (New York: Wiley).
Clark, H. H. and B. C. Malt. 1984. Psychological constraints on language: a commentary on Bresnan and Kaplan and on Givón. In *Method and Tactics in Cognitive Science* edited by Kintsch, W., J. R. Miller, and P. G. Polson, 191-214 (Hillsdale, NJ: LEA).
Coen, M. H. 2006. Self-supervised acquisition of vowels in American English. Submission to AAAI 06 (American Association for Artificial Intelligence).
Coleman, J. 1998. Cognitive reality and the phonological lexicon: a review. *Journal of Neurolinguistics* 11 (3):295-320.

Collins, B. and I. M. Mees. 1996. *The phonetics of English and Dutch.* (Leiden: Brill).
Cooker, H. 1963. Time relationships of chest wall movements and intraoral pressures during speech. PhD thesis, U o Iowa.
Coupland, N. and H. Giles. 1988. The communicative contexts of accommodation. *Language and Communication* 8 (3):175-182.
Crelin, E. S. 1987. *The Human Vocal Tract: Anatomy, Function, Development and Evolution.* (New York: Vantage Press).
Crystal, T. H. and A. S. House. 1988. The duration of American-English vowels: an overview. *Journal of Phonetics* 16:263-284.
Crystal, T. H. and A. S. House. 1988. The duration of American-English stop consonants: an overview. *Journal of Phonetics* 16:285-294.
Cummins, F. 2003. Rhythmic grouping in word lists: competing roles of syllables, words and stress feet. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 325-328 (Barcelona: Causal Productions).
Cutler, A. and D. A. Swinney. 1987. Prosody and the development of comprehension. *Journal of Child Language* 14:145-167.
Davis, B. L. and P. MacNeilage. 1990. Acquisition of correct vowel production: a quantitative case study . *Journal of Speech and Hearing Research* 33 (1):16-27.
Davis, B. L. and P. MacNeilage. 1995. The articulatory basis of babbling. *Journal of Speech and Hearing Research* 38:1199-1211.
Davis, B. L. and P. MacNeilage. 2000. An embodiment perspective on the acquisition of speech perception. *Phonetica* 57:229-241.
Davis, B. L., P. F. MacNeilage, C. L. Matyear, and J. K. Powell. 2000. Prosodic correlates of stress in babbling: an acoustical study. *Child Development* 71 (5):1258-1270.
Davis, B. L., P. F. MacNeilage, and C. L. Matyear. 2002. Acquisition of serial complexity in speech production: a comparison of phonetic and phonological approaches to first word production. *Phonetica* 59:75-107.
de Boysson-Bardies, B., P. Halle, L. Sagart, and D. Durand. 1989. A cross-linguistic investigation of vowel formants in babbling. *Journal of Child Language* 16:1-17.
de Boysson-Bardies, B. 1999. *How Language Comes to Children.* (Cambridge, MA: MIT Press).
de Jong, K. J. 1991. An articulatory study of vowel duration changes in English. *Phonetica* 48:1-18.
de Jong, K. J. 1994. The correlation of P-center adjustments with articulatory and acoustic events . *Perception and Psychophysics* 56 (4):447-460.
de Jong, K. J. 2001. Effects of syllable affiliation and consonant voicing on temporal adjustment in a repetitive speech-production task. *Journal of Speech, Language and Hearing Research* 44:826-840.
Demiris, Y. and G. Hayes. 2002. Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 327-361 (Cambridge, MA: MIT Press).
Dennett, D. C. 1987. *The Intentional Stance.* (Cambridge MA: MIT Press/Bradford Books).
Di Paolo, M. and A. Faber. 1990. Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2:155-204.
Donegan, P. 2002. Normal vowel development. In *Vowel disorders* edited by Ball, M. J. and D. Gibbon, 1-35 (Oxford: Butterworth).
Doupe, A. J. and P. K. Kuhl. 1999. Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience* 22:567-631.
Duffy, R. J. and J. R. Duffy. 1975. Pantomime recognition in aphasics. *Journal of Speech and Hearing Research* 18 (1):115-132.
Durlach, N.I. and L.D. Braida. 1969. Intensity perception: a preliminary theory of intensity resolution. *Journal of the Acoustical Society of America* 46:372-383.
Edwards, E. 1979. *Drawing on the Right Side of the Brain.* (London: Harper Collins).
Eerola, O., J-P. Laaksonen, J. Savela, and O. Aaltonen. 2003. Perception and production of the short and long Finnish [i] vowels: individuals seem to have different perceptual and articulatory templates. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 989-992 (Barcelona: Causal Productions).
Ejiri, K. 1998. Relationship between rhythmic behavior and canonical babbling in infant vocal development. *Phonetica* 55:226-237.
Elbers, L. and F. Wijnen. 1992. Effort, production skill, and language learning. In *Phonological development: models, research, implications* edited by Ferguson, C. A., L. Menn, and C. Stoel-Gammon, 337-368 (Timonium, MA: York Press).
Emmorey, K. 2005. Signing for viewing: some relations between the production and comprehension of sign language. In *Twenty-First Century Psycholinguistics* edited by Cutler, A., 293-309 (LEA).

Engstrand, O., K. Williams, and F. Lacerda. 2003. Does babbling sound native? Listener responses to vocalizations produced by Swedish and American 12- and 18-month olds. *Phonetica* 60:17-44.
Epstein, M. A. 2003. Voice quality and prosody in English. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 2405-2408 (Barcelona: Causal Productions).
Faber, A. and C. T. Best. 1994. The perceptual infrastructure of early phonological development. In *The reality of lingustic rules* edited by Lima, S. D., R. L. Corrigan, and G. K. Iverson, 261-280 (Amsterdam: John Benjamin).
Fant, G., A. Kruckenberg, S. Hertegård, and J. Liljencrants. 1997. Sub- and supraglottal pressures in speech. *Phonum (Umea)* 4:25-28.
Faw, W. R. 1997. Outlining a brain model of mental imaging abilities. *Neuroscience and Biobehavioral Reviews* 21 (3):283-288.
Ferguson, C. A. and C. B. Farwell. 1975. Words and sounds in early language acquisition. *Language* 51 (2):419-439.
Finnegan, E. M., E. S. Luschei, and H. T. Hoffman. 1999. Estimation of alveolar pressure during speech using direct measures of tracheal pressure. *Journal of Speech, Language and Hearing Research* 42:1136-1147.
Finnegan, E. M., E. S. Luschei, and H. T. Hoffman. 2000. Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech. *Journal of Speech, Language and Hearing Research* 43:934-950.
Fischer-Jørgensen, E. 1984. Some basic vowel features, their articulatory correlates and their explanatory power in phonology. *ARIPUC (Annual Report of the Institute of Phonetics, University of Copenhagen)* 18:255-276.
Fischer-Jørgensen, E. 1990. Intrinsic F0 in tense and lax vowels with special reference to German. *Phonetica* 47:99-140.
Flavell, J. H., F. L. Green, E. R. Flavell, and J. B. Grossman. 1997. The development of children's knowledge about inner speech. *Child Development* 68 (1):39-47.
Flege, J. E. and J. M. Hillenbrand. 1987. A differential effect of release bursts on the stop voicing judgments of native French and English listeners. *Journal of Phonetics* 15:203-208.
Flege, J. E. and W. Eefting. 1988. Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *Journal of the Acoustical Society of America* 83 (2):729-740.
Fonagy, P., M. Target, G. Gergely, J. A. Allen, and A. W. Bateman. 2003. The developmental roots of Borderline Personality Disorder in early attachment relationships: a theory and some evidence. *Psycholanalytic Inquiry* 23 (3):412-459.
Fowler, A. E. 1991. How early phonological development might set the stage for phoneme awareness. *Haskins Laboratories Status Report on Speech Research* 105/106:53-64.
Fowler, C. A. 1979. "Perceptual centers" in speech production and perception. *Perception and Psychophysics* 25 (5):375-388.
Fowler, C. A., P. Rubin, R. E. Remez, and M. T. Turvey. 1980. Implications for speech production of a general theory of action. In *Language Production: Volume 1 Speaking and Talking* edited by Butterworth, B., 373-420 (London: Academic Press).
Fowler, C. A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology, General* 112:386-412.
Fowler, C. A. 1985. Current perspectives on language and speech production: a critical overview. In *Speech Sciences* edited by Daniloff, R., 47-72 (San Diego: Collyhill Press).
Fowler, C. A. 1996. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America* 99 (3):1730-1741.
Fowler, C. A., J. M. Brown, L. Sabadini, and J. Weihing. 2003. Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *Journal of Memory and Language* 49:396-413.
Fox, A. 2000. *Prosodic features and prosodic structures.* (Oxford: OUP).
Fry, D. B. 1968. The phonemic system in children's speech. *British Journal of Disord.Comm* 3:13-19.
Fudge, E. 1999. Words and feet. *Journal of Linguistics* 35:273-296.
Fujisaki, H. 1993. From information to intonation. In *Proceedings 1993 International Symposium on Spoken Dialogue* (Waseda University).
Galef, B. G. 1988. Imitation in animals: history, definition and interpretation of data from the psychological laboratory. In *Social Learning* edited by Zentall, T., 3-28.
Gandour, J., B. Weinberg, and D. Rutkowsky. 1980. Influence of postvocalic consonants on vowel duration in esophageal speech. *Language and Speech* 23:149-158.
Gattegno, C. 1972. *Teaching foreign languages in schools: the Silent Way.* Second ed. (New York: Educational Solutions).
Gattegno, C. 1973. *The Universe of Babies.* (New York: Educational Solutions).

Gattegno, C. 1975. *The Mind Teaches the Brain.* (New York: Educational Solutions).
Gattegno, C. 1976. *The Common Sense of Teaching Foreign Languages.* (New York: Educational Solutions).
Gattegno, C. 1985. *The Learning and Teaching of Foreign Languages.* (New York: Educational Solutions).
Gattegno, C. 1987. *The Science of Education. Part 1: Theoretical considerations.* (New York: Educational Solutions).
Gattegno, C. 1998. *Caleb Gattegno: Autoportrait impressioniste.* Edited by Granchamp, G. (Gaillard, France: Eveil).
Gattis, M., H. Bekkering, and A. Wohlschläger. 2002. Goal-directed imitation. In *The imitative mind: development, evolution and brain bases* edited by Meltzoff, A. N. and W. Prinz, 183-205 (Cambridge: CUP).
Gauffin, J. and J. Sundberg. 1989. Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research* 32:556-565.
Gentile, A. M. 1987. Skill acquisition: action, movement, and neuromotor processes. In *Movement science: foundations for physical therapy in rehabilitation* edited by Carr, J. H., 117-141 (Rockville: Aspen).
Gentilucci, M. and A. Negrotti. 1994. Dissociation between perception and visuomotor transformation during reproduction of remembered distances. *Journal of Neurophysiology* 72 (4):2026-2030.
Gergely, G. and J. S. Watson. 1996. The social biofeedback theory of parental affect-mirroring. *International Journal of Psycho-Analysis* 77:1181-1212.
Gergely, G., H. Bekkering, and I. Király. 2002. Developmental psychology: Rational imitation in preverbal infants. *Nature* 415:755.
Gerken, L. A. 2002. Early sensitivity to linguistic form. *Annual Review of Language Acquisition* 2:1-36.
Gerrits, E. and M. E. H. Schouten. 2004. Categorical perception depends on the discrimination task . *Perception and Psychophysics* 66 (3):363-376.
Gick, B. and I. Wilson. 2002. Excrescent schwa and vowel laxing: cross linguistic responses to conflicting articulatory targets. In *Laboratory Phonology 8* ().
Gick, B. 2002. An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica* 59:38-48.
Gigerenzer, G. and P. M. Todd. 1999. *Simple heuristics that make us smart.* (OUP).
Gimson, A. C. 1989. *An Introduction to the Pronunciation of English.* Edited by Ramsaran, S. 4 ed. (London: Edward Arnold).
Goldenberg, G. and S. Hagmann. 1997. The meaning of meaningless gestures: a study of visuo-imitative apraxia. *Neuropsychologia* 35 (3):333-341.
Goldstein, L. and C. A. Fowler. 2003. Articulatory phonology: a phonology for public language use. In *Phonetics and Phonology in Language Comprehension and Production* edited by Schiller, N. O. and A. S. Meyer, 159-207 (Mouton de Gruyter).
Goldstein, M. H., A. P. King, and M. J. West. 2003. Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc Natl Acad Sci USA* 100 (13):8030-8035.
Goldstein, U. 1980. *An articulatory model for the vocal tract of growing children.* (PhD dissertation, MIT).
Goodale, M. A. and G. K. Humphrey. 2001. Separate visual systems for action and perception. In *Handbook of Perception* edited by Goldstein, E. B., 311-343 (Oxford: Blackwell).
Goodell, E. W. and M. Studdert-Kennedy. 1993. Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: a longitudinal study. *Journal of Speech and Hearing Research* 36:707-727.
Gopal, H. S. 1996. Generalizability of Current Models of Vowel Duration. *Phonetica* 53:1-32.
Gordeeva, O. B. 2005. Language interaction in the bilingual acquisition of sound structure. PhD thesis, Queen Margaret University College.
Goto, H. 1971. Auditory perception by normal Japanese adults of the sounds 'l' and 'r'. *Neuropsychologia* 9:317-323.
Gottlieb, G. L., D. M. Corcos, and G. C. Agarwhal. 1989. Strategies for the control of voluntary movements with one mechanical degree of freedom. *The Behavioral and Brain Sciences* 12:189-250.
Gowen, E. and C. Miall. 2006. Eye-hand interactions in tracing and drawing tasks. *Human Movement Science* 25:568-585.
Grabe, E., B. Post, and I. Watson. 1999. The acquisition of rhythmic patterns in English and French. In *ICPhS99*, 1201-1204 (San Francisco).
Greenlee, M. 1980. Learning the phonetic cues to the voiced-voiceless distinction: a comparison of child and adult speech perception. *Journal of Child Language* 7:459-468.
Gregory, R. W. 1970. *The Intelligent Eye.* (London: Weidenfeld & Nicolson).

Grèzes, J., N. Costes, and J Decety. 1998. Top-down effect of strategy on the perception of human biological motion: a PET investigation. *Cognitive Neuropsychology* 15:553-582.
Griffiths, T. D. and J. D. Warren. 2004. What is an auditory object? *Nature Reviews Neuroscience* 5:887-892.
Grush, R. 2004. The emulation theory of representation: motor control, imagery and perception. *The Behavioral and Brain Sciences* 27:377-442.
Guenther, F. H., M. Hampson, and D. Johnson. 1998. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* 105 (4):611-633.
Guenther, F. H., S. S. Ghosh, and J. A. Tourville. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96:280-301.
Guillaume, P. 1926. *Imitation in Children.* (tr. Halperin E.P. 1971) University of Chicago Press.
Hallé, P. A. and B. de Boysson-Bardies. 1996. The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development* 19:463-481.
Halliday, M. A. K. 1975. *Learning how to mean.* (London: Edward Arnold).
Hamlet, S. L. 1972. Vocal fold articulatory activity during whispered speech. *Archs Otolar.* 95:211-312.
Hanley, J. R., J. Kay, and M. Edwards. 2002. Imageability effects, phonological errors, and the relationship between auditory repetition and picture naming: implications for models of auditory repetition. *Cognitive Neuropsychology* 19 (3):193-206.
Hanley, J. R., G. S. Dell, J. Kay, and R. Baron. 2004. Evidence for the involvement of a nonlexical route in the repetition of familiar words: a comparison of single and dual route models of auditory repetition. *Cognitive Neuropsychology* 21 (2-4):147-158.
Hardcastle, W. J. and F. Gibbon. 2005. EPG as a research and clinical tool: 30 years on. In *A Figure of Speech: a Festschrift for John Laver* edited by Hardcastle, W. J. and J. Mackenzie Beck, 39-60 (Lawrence Erlbaum Associates).
Harris, J. G. 1999. States of the glottis for voiceless plosives. In *ICPhS99*, 2041-2044 (San Francisco).
Hart, B. and T. R. Risley. 1999. *The social world of children learning to talk.* (Baltimore, MD: Paul Brookes).
Hawkins, S. 1979. Temporal co-ordination of consonants in the speech of children: further data. *Journal of Phonetics* 7:235-267.
Hawkins, S. 1984. On the development of motor control in speech: Evidence from studies of temporal coordination. In *Speech and Language: Advances in basic research and practice, Vol 11* edited by Lass, N. J., 317 (Academic Press).
Hawkins, S. 1994. Speech development: acoustic/phonetic studies. In *The Encyclopaedia of Language and Linguistics* edited by Asher, B. E., 4178-4181 (Oxford: Pergamon).
Hawkins, S. 2003. Contribution of fine phonetic detail to speech understanding. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 293-296 (Barcelona: Causal Productions).
Heffner, R-M. S. 1964. *General Phonetics.* (Madison: University of Wisconsin Press).
Heinks-Maldonado, T. H., D. H. Mathalon, M. Gray, and J. M. Ford. 2005. Fine-tuning of auditory cortex during speech production. *Psychophysiology* 42 (2).
Heldner, M. 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics* 31:39-62.
Henry, L. A., J. E. Turner, P. T.Smith, and C. Leather. 2000. Modality effects and the development of the word length effect in children. *Memory* 8 (1):1-17.
Hewlett, N., F. Gibbon, and W. Cohen-McKenzie. 1998. When is a velar an alveolar? Evidence supporting a revised psycholinguistic model of speech production in children. *International Journal of Language and Communication Disorders* 33 (2):161-176.
Hewson, J. 1980. Stress in English: four levels or three? *Canadian Journal of Linguistics* 25 (2):197-203.
Hewson, J. 1998. Review of Lima et al. (ed.s) The Reality of Linguistic Rules. *Word* 49 (3):429-437.
Heyes, C. 2001. Causes and consequences of imitation. *Trends in Cognitive Sciences* 5 (6):253-261.
Heyes, C. 2005. Imitation by association: evolution and development. In *Perspectives on Imitation. Volume 1: Mechanisms of Imitation and Imitation in Animals* edited by Hurley, S. and N. Chater, 162-170 (Cambridge MA: MIT Press).
Heyes, C. M. 1994. Social learning in animals: categories and mechanisms. *Biological Review* 69:207-231.
Heyes, C. M. and E. D. Ray. 2000. What is the significance of imitation in animals? *Advances in the Study of Behavior* 29:215-245.
Heyes, C. M. and E. D. Ray. 2002. Distinguishing intention-sensitive from outcome-sensitive imitation. *Developmental Science* 5 (1):34-36.
Hickok, G. and D. Poeppel. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92:67-99.
Hixon, T. J., M. D. Goldman, and J. Mead. 1973. Kinematics of the chest wall during speech production: volume displacements of the rib cage, abdomen and lung. *Journal of Speech and Hearing Research* 16:78-115.

Hixon, T. J., M. D. Goldman, and J. Mead. 1976. Dynamics of the chest wall during speech production: function of the thorax, rib cage, diaphragm and abdomen. *Journal of Speech and Hearing Research* 19:297-356.
Hixon, T. J. 1987. *Respiratory function in speech and song.* (Boston, MA: Little, Brown).
Hixon, T. J. and G. Weismer. 1995. Perspectives on the Edinburgh study of speech breathing. *Journal of Speech and Hearing Research* 38:42-60.
Hixon, T. J. and J. D. Hoit. 2005. *Evaluation and management of speech breathing disorders.* (Tucson, AZ: Redington Brown).
Hogan, J. T. and A. J. Rozsypal. 1980. Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America* 67 (5):1764-1771.
Hoit, J. D., T. J. Hixon, P. J. Watson, and W. J. Morgan. 1990. Speech breathing in children and adolescents. *Journal of Speech and Hearing Research* 33:51-69.
Hoit, J. D., N. P. Solomon, and T. J. Hixon. 1993. Effect of lung volume on voice onset time (VOT). *Journal of Speech and Hearing Research* 36:516-521.
Hoppitt, W. and K. N. Laland. 2002. Neural network models of imitation. Poster presented at "Perspectives on imitation: From cognitive neuroscience to social science", Royaumont Abbey, May 2002.
Houde, J. F. and M. I. Jordan. 2002. Sensorimotor adaptation of speech I: compensation and adaptation. *Journal of Speech, Language and Hearing Research* 45:295-310.
Houston, D. M. 2005. Speech Perception in Infants. In *The Handbook of Speech Perception* edited by Pisoni, D. and R. E. Remez, 417-448 (Oxford: Blackwell).
Howard, D. and L. Nickels. 2005. Separating input and output phonology: semantic, phonological and orthographic effects in short-term memory impairment. *Cognitive Neuropsychology* 22 (1):42-77.
Howell, P. and D. J. Powell. 1984. Hearing your voice through bone and air: implications for explanations of stuttering behavior from studies of normal speakers. *Journal of Fluency Disorders* 9:247-264.
Howell, P. 1985. Acoustic feedback of the voice in singing. In *Musical Structure and Cognition* edited by Howell, P., I. Cross, and R. West, 259-286 (New York: AP).
Howell, P. and S. Sackin. 2002. Timing interference to speech in altered listening conditions. *Journal of the Acoustical Society of America* 111 (6):2842-2852.
Hsu, H-C. and A. Fogel. 2001. Infant vocal development in a dynamic mother-infant communication system. *Infancy* 2 (1):87-109.
Huang, C-T., C. Heyes, and A. Charman. 2002. Infants' behavioral reenactment of "failed attempts": exploring the roles of emulation learning, stimulus enhancement and understanding of intentions. *Developmental Psychology* 38 (5):840-855.
Hudgins, C. V. 1937. Voice production and breath control in the speech of the deaf. *American Annals of the Deaf* 82:338-363.
Huttenlocher, J. 1974. The origins of language comprehension. In *Theories in Cognitive Psychology: The Loyola Symposium* edited by Solso, R. L., 331-368 (Potomac, MD: LEA).
Ivry, R. 1996. Representational issues in motor learning: phenomena and theory. In *Handbook of perception and action: volume 2: motor skills* edited by Heuer, H. and S. W. Keele, 263 (London: Academic Press).
Jacquemot, C., E. Dupoux, and A-C. Bachoud-Lévi. 2007. Breaking the mirror: asymmetrical disconnection between the phonological input and output codes. *Cognitive Neuropsychology.*
Jakobson, R. 1972. *Child Language, Aphasia, and Phonological Universals.* (The Hague: Mouton).
Jakobson, R. and L. Waugh. 1979. *The Sound Shape of Language.* (Brighton: Harvester Press).
Jeannerod, M. 1994. The representing brain: neural correlates of motor intention and imagery. The Behavioral and Brain Sciences 17, 187-245.
Jeannerod, M. and P. Jacob. 2005. Visual cognition: a new look at the two-visual systems model. *Neuropsychologia* 43:301-312.
Jenkins, J. J. 1980. Research in child phonology: comments, criticism and advice. In *Child Phonology. Volume 2, Perception* edited by Yeni-Komshian, G., J. Kavanagh, and C. A. Ferguson, 217-228 (New York: Academic Press).
Jensen, C. 2004. Stress and accent: prominence relations in southern standard British English. PhD thesis, University of Copenhagen.
Johnson, K. 2005. Speaker normalization in speech perception. In *The Handbook of Speech Perception* edited by Pisoni, D. and R. E. Remez, 363-389 (Oxford: Blackwell).
Jones, J. A. and K. Munhall. 2000. Perceptual calibration of F0 production: Evidence from feedback perturbation . *Journal of the Acoustical Society of America* 108 (3):1246-1251.
Jones, J. A. and K. Munhall. 2003. Learning to produce speech with an altered vocal tract: The role of auditory feedback . *Journal of the Acoustical Society of America* 113 (1):532-543.

Jones, L. G. 1967. English phonotactic structure and first-language acquisition. *Lingua* 19:1-59.
Jones, S. S. 1996. Imitation or exploration? Young infants' matching of adult gestures. *Child Development* 67:1952-1969.
Jonsson, C-O., D. N. Clinton, M. Fahrman, G. Mazzaglia, S. Novak, and K. Sörhus. 2001. How do mothers signal shared feeling-states to their infants? An investigation of affect attunement and imitation during the first year of life. *Scandinavian Journal of Psychology* 42:377-381.
Karmiloff-Smith, A. 1992. *Beyond Modularity.* (MIT Press).
Keating, P. 1984. Universal phonetics and the organisation of grammars. *UCLA Working Papers in Phonetics* 59:35-49.
Keating, P. 2003. Phonetic and other influences on voicing contrasts. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 375-378 (Barcelona: Causal Productions).
Kehoe, M., C. Stoel-Gammon, and E. H. Buder. 1995. Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research* 38:338-350.
Kehoe, M. and C. Stoel-Gammon. 1995. An investigation of rhythmic processes in English-speaking children's word productions. In *ICPhS XIII, Volume 2* edited by Elenius, K. and P. Branderud, 702-705 (Stockholm).
Kehoe, M. M. and C. Stoel-Gammon. 2001. Development of syllable structure in English-speaking children with particular reference to rhymes. *Journal of Child Language* 28:393-432.
Kent, R. 1992. The biology of phonological development. In *Phonological development: models, research, implications* edited by Ferguson, C. A., L. Menn, and C. Stoel-Gammon, 65-89 (Timonium, MA: York Press).
Kent, R. D. 1981. Sensorimotor aspects of speech development. In *Development of Perception, Volume 1* edited by Aslin, R. N., J. R. Alberts, and M. R. Peterson, 162-185 (New York: Academic Press).
Kent, R. D., S. G. Adams, and G. S. Turner. 1996. Models of Speech Production. In *Principles of Experimental Phonetics* edited by Lass, N. J., 3-45 (St Louis: Mosby).
Kent, R. D. 2004. Models of speech motor control: implications from recent developments in neurophysiological and neurobehavioral science. In *Speech motor control* edited by Maassen, B., R. Kent, H. F. M. Peters, P. van Lieshout, and W. Hulstijn, 3-28 (Oxford: University Press).
Keri, S. 2003. The cognitive neuroscience of category learning. *Brain Research Reviews* 43 (1):85-109.
Khan, M. A., M. I. Garry, and I. M. Franks. 1999. The effect of target size and inertial load on the control of rapid aiming movements. *Experimental Brain Research* 124:151-158.
Killian, K. J. and S. C. Gandevia. 1996. Sense of effort and dyspnea. In *Respiratory sensation* edited by Adams, L. and A. Guz, 181-199 (New York: Dekker).
Kim, C. W. 1970. A theory of aspiration. *Phonetica* 21:107-116.
Kim, M. 2005. Acoustic characteristics of Korean stops in Korean child-directed speech. *Journal of the Acoustical Society of America* 117 (4.2):2458.
King, A. P., M. J. West, and M. H. Goldstein. 2005. Non-vocal shaping of avian song development: parallels to human speech development. *Ethology* 111:101-117.
Kingston, J., N. A. Macmillan, L. W. Dickey, R. Thorburn, and C. Bartels. 1997. Integrality in the perception of tongue root position and voice quality in vowels. *Journal of the Acoustical Society of America* 101 (3):1696-1709.
Klatt, D. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59 (5):1208-1221.
Kneil, T. R. 1972. Subglottal pressures in relation to chest wall movement during selected samples of speech. PhD thesis, University of Iowa.
Kochanski, G., E. Grabe, J. Coleman, and B. S. Rosner. 2005. Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America* 118 (2):1038-1054.
Koenig, L. L. 2000. Laryngeal factors in voiceless consonant production in men, women and 5-year-olds. *Journal of Speech, Language and Hearing Research* 43:1211-1228.
Kohler, K. J. 1984. Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica* 41:150-174.
Kokkinaki, T. and V. G. S. Vasdekis. 2003. A cross-cultural study on early vocal imitative phenomena in different relationships. *Journal of Reproductive and Infant Psychology* 2118 (2):85-101.
Konopczynski, G. 1995. A developmental model of acquisition of rhythmic patterns: results from a cross linguistic study. In *XIIIth ICPhS, Vol 4* edited by Elenius, K. and P. Banderud, 22-29 (Stockholm).
Kostyk, B. E. and A. Putnam Rochet. 1998. Laryngeal airway resistance in teachers with vocal fatigue: a preliminary study. *Journal of Voice* 12 (3):287-299.
Kraus, N. and T. Nicol. 2005. Brainstem origins for cortical 'what' and 'where' pathways in the auditory system. *Trends in Neurosciences* 28 (4):176-181.
Kroos, C., P. Hoole, B. Kuhnert, and H. G. Tillmann. 1997. Phonetic evidence for the phonological status of the tense-lax distinction in German. *FIPKM (Forschungberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München)* 35:17-25.

Kuhl, P. K. 1987. Perception of speech and sound in early infancy. In *Handbook of Infant Perception, Vol 2* edited by Salapatek, P. and L. Cohen, 275-382 (New York: AP).
Kuhl, P. K. 1991. Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech. In *Plasticity of Development* edited by Brauth, S. E., W. S. Hall, and R. J. Dooling, 79 (Cambridge MA: MIT Press).
Kuhl, P. K. and A. N. Meltzoff. 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America* 100 (4):2425-2438.
Kuhl, P. K. 2000. A new view of language acquisition. *Proc Natl Acad Sci USA* 97 (22):11850-11857.
Lacerda, F. 1993. Sonority contrasts dominate young infants' vowel perception. *Perilus (Institute of Linguistics, Stockholm University)* XVII:55-63.
Lacerda, F. 2003. Phonology: an emergent consequence of memory constraints and sensory input. *Reading and Writing* 16:41-59.
Ladd, D. R. 1993. Notes on the phonology of prominence. *Working Papers, Department of Linguistics, Lund University* 41:10-15.
Ladefoged, P. and N. P. McKinney. 1963. Loudness, sound pressure, and subglottal pressure in speech. *Journal of the Acoustical Society of America* 35 (4):454-460.
Ladefoged, P. 1967. *Three Areas of Experimental Phonetics.* (London: OUP).
Ladefoged, P. 1983. Cross-linguistic studies of speech production. In *The Production of Speech* edited by MacNeilage, P. F., 177-188 (New York: Springer-Verlag).
Ladefoged, P. 1984. 'Out of chaos comes order'; physical, biological, and structural patterns in phonetics. In *Proceedings of the Xth ICPhS* edited by Cohen, A. and M. P. R. van den Broecke, 83-95 (Dordrecht: Foris).
Laeufer, C. 1992. Patterns of voicing-conditioned vowel duration in French and English. *Journal of Phonetics* 20:411-440.
Laeufer, C. 1996. The acquisition of a complex phonological contrast: voice timing patterns of English final stops by native French speakers. *Phonetica* 53:117-142.
Laufer, M. Z. 1980. Temporal regularity in prespeech. In *Infant communication* edited by Murry, T. and J. Murry, 284-309 (Houston, TX: College Hill Press).
Laukkanen, A. M., P. Lindholm, and E. Vilkman. 1995. On the effects of various vocal training methods on glottal resistance and efficiency. *Folia Phoniatrica Logop* 47:324-330.
Levelt, W. J. M., A. Roelofs, and A. S. Meyer. 1999. A theory of lexical access in speech production. *The Behavioral and Brain Sciences* 22:1-75.
Lewis, M. M. 1951. *Infant Speech.* 2nd ed. (London: Routledge, Kegan, Paul).
Lewis, M. M. 1957. *How Children Learn to Speak.* (London: Harrap).
Liberman, A. M. and D. H. Whalen. 2000. On the relation of speech to language. *Trends in Cognitive Sciences* 4 (5):187-196.
Lieberman, P. 1967. *Intonation, Perception and Language.* (Cambridge, MA: MIT Press).
Lieberman, P. 1980. On the development of vowel production in young children. In *Child Phonology: Vol 1, Production* edited by Yeni-Komshian, G., J. Kavanagh, and C. A. Ferguson, 113-142 (NY: Academic Press).
Lieberman, P. 1984. *The Biology and Evolution of Language.* (Cambridge, MA: Harvard University Press).
Linell, P. 1982. The concept of phonological form and the activities ol speech production and speech perception . *Journal of Phonetics* 10:37-72.
Lisker, L. 1974. On 'explaining' vowel duration. *Glossa* 8:233-246.
Llisterri, J. 1995. Relationships between speech production and speech perception in a second language. In *ICPhS XIII, Volume 2* edited by Elenius, K. and P. Branderud, 92-99 (Stockholm).
Lock, A. 1980. *The Guided Reinvention of Language.* (London: Academic Press).
Locke, J. L. and K. J. Kutz. 1975. Memory for speech and speech for memory. *Journal of Speech and Hearing Research* 18:176-191.
Locke, J. L. 1979. The child's processing of phonology. In *Child Language and Communication: Minnesota Symposium on Child Psychology Volume 12* edited by Collins, W. A., 83-119 (Hillsdale, NJ: LEA).
Locke, J. L. 1983. *Phonological acquisition and change.* (NY: Academic Press).
Locke, J. L. 1986. Speech perception and the emergent lexicon: an ethological approach. In *Language Acquisition* edited by Fletcher, P. and M. Garman, 240-250 (CUP).
Locke, J. L. 1993. *The Child's Path to Spoken Language.* (Cambridge, MA: Harvard University Press).
Locke, J. L. 1996. Why do infants begin to talk? Language as an unintended consequence. *Journal of Child Language* 23:251-268.
Locke, J. L. and C. Snow. 1997. Social influences on vocal learning in human and nonhuman primates. In *Social influences on vocal development* edited by Snowdon, C. T. and M. Hausberger, 274-292 (New York: CUP).

Locke, J. L. 2001. First communion: the emergence of vocal relationships. *Social Development* 10 (3):294-308.
Löfqvist, A. 1997. Theories and models of speech production. In *The Handbook of Phonetics* edited by Hardcastle, W. H. and J. Laver, 405-426 (Oxford: Blackwell).
Luce, R. D. 1986. *Response Times.* (New York: OUP).
Mack, M. and P. Lieberman. 1985. Acoustic analysis of words produced by a child from 46 to 149 weeks. *Journal of Child Language* 12:527-550.
MacKain, K. S. 1982. Assessing the role of experience on infants' speech discrimination. *Journal of Child Language* 9:527-542.
MacKain, K. S. 1988. Filling the gap between speech and language. In *The Emergent Lexicon: the Child's Development of a Linguistic Vocabulary* edited by Locke, J. L. and M. D.Smith, 51-74 (New York: Academic Press).
MacKay, D. G. 1992. Constraints on theories of inner speech. In *Auditory Imagery* edited by Reisberg, D., 121-149 (Hillsdale, NJ: Erlbaum).
Macken, M. A. and D. Barton. 1980. The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language* 7:41-47.
Macmillan, N. A., R. F. Goldberg, and L. D. Braida. 1988. Resolution for speech sounds: basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America* 84 (4):1262-1280.
MacNeilage, P. and P. Ladefoged. 1976. The production of speech and language. Articulatory dynamics: segment durations. In *Handbook of Perception Vol VII: Language and Speech* edited by Carterette, E. C. and M. P. Friedman, 94-98 (London: Academic Press).
MacNeilage, P. 1998. The frame/content view of speech: what survives, what emerges. *The Behavioral and Brain Sciences* 21 (4):532-538.
MacNeilage, P. and B. L. Davis. 2000. Deriving speech from nonspeech: a view from ontogeny. *Phonetica* 57:284-296.
MacNeilage, P. F. and B. L. Davis. 1990. Acquisition of speech production: the achievement of segmental independence. In *Speech Production and Speech Modelling* edited by Hardcastle, W. J. and A. Marchal, 55-68 (London: Kluwer Academic Publ.).
Maddieson, I. 1997. Phonetic universals. In *The Handbook of Phonetics* edited by Hardcastle, W. H. and J. Laver, 622-624 (Oxford: Blackwell).
Maddox, W. T., F. G. Ashby, and E. M. Waldron. 2002. Multiple attention systems in perceptual organization. *Memory and Cognition* 30 (3):325-339.
Magill, R. A. 1993. *Motor Learning - Concepts and Applications (4th edition).* (Dubuque: Wm Brown).
Malécot, A. 1970. The Lenis-Fortis Opposition: Its Physiological Parameters. *Journal of the Acoustical Society of America* 47 (6):1588-1592.
Marcault, J-E. and T. Brosse. 1949. *L'éducation de demain.* Second ed. (Paris: Adyar).
Markova, G. and M. Legerstee. 2006. Contingency, imitation, and affect sharting: foundations of infants' social awareness. *Developmental Psychology* 42 (1):132-141.
Marshall, C. and S. Chiat. 2003. A foot domain account of prosodically conditioned substitutions. *Clinical Linguistics and Phonetics* 17 (8):645-657.
Marslen-Wilson, W. D. 1985. Speech shadowing and speech comprehension. *Speech Communication* 4:55-73.
Marteniuk, R. G. and S. K. E. Romanow. 1983. Human movement organization and learning as revealed by variability of movement, use of kinematic information and Fourier analysis. In *Memory and Control of Action* edited by Magill, R. A., 167-197 (North Holland).
Martin, R. C., M. F. Lesch, and M. C. Bartha. 1999. Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language* 41:3-29.
Mason, J. 1994. Researching from the inside in mathematics education - locating an I-You relationship. In *Proceedings of the XVIIIth International Conference for the Psychology of Mathematics Education Vol 1* edited by da Ponte, J. P. and J. F. Matos, 176-194 (Lisbon).
Mason, J. 2002. *Researching Your Own Practice: the Discipline of Noticing.* (London: RoutledgeFalmer).
Mattar, A. A. G. and P. L. Gribble. 2005. Motor learning by observing. *Neuron* 46:153-160.
Mattock, K., S. Rvachew, L. Polka, and S. Turner. 2005. A comparison of vowel formant frequencies in the babbling of infants exposed to Canadian English and Canadian French. *Journal of the Acoustical Society of America* 117 (4.2):2402.
McCarthy, R. and E. K. Warrington. 1984. A two-route model of speech production. Brain 107, 463-485.
McCune, L. and M. M. Vihman. 1987. Vocal Motor Schemes. *Papers and Reports in Child Language Development, Stanford University Department of Linguistics* 26:72-79.
McCune, L. 1992. First words, a dynamic systems view. In *Phonological development: models, research, implications* edited by Ferguson, C. A., L. Menn, and C. Stoel-Gammon, 313-321 (Timonium, MA: York Press).

McCune, L. and M. M. Vihman. 2001. Early phonetic and lexical development: a productivity approach. *Journal of Speech, Language and Hearing Research* 44:670-684.
McGowan, R. S. and A. Faber. 1996. Introduction to papers on speech recognition and perception from an articulatory point of view. *Journal of the Acoustical Society of America* 99 (3):1680-1682.
McLeod, P. and M. I. Posner. 1984. Privileged loops from percept to act. In *Attention and Performance X: Control of language processes* edited by Bouma, H. and D. G. Bouwhuis, 55-66 (London: LEA).
McLeod, P. and Z. Dienes. 1996. Do fielders know where to go to catch the ball or only how to get there? *Journal of Experimental Psychology: Human Perception and Performance* 22 (3):531-543.
McLeod, S., J. van Doorn, and V. A. Reed. 2001. Normal acquisition of consonant clusters. *American Journal of Speech-Language Pathology* 10:99-110.
McMurray, R., M. Spivey, and R. N. Aslin. 2000. The perception of consonants by adults and infants: categorical or categorized. *University of Rochester Working Papers in the Language Sciences* 1 (2):215-256.
Mead, J. and M. B. Reid. 1988. Respiratory muscle activity during repeated airflow interruption. *Journal of Applied Physiology* 64:2314-2317.
Meltzoff, A. N. and M. K. Moore. 1997. Explaining facial imitation: a theoretical model. *Early development and parenting* 6:179-192.
Menn, L. 1983. Development of articulatory, phonetic, and phonological capabilities. In *Language Production, Volume 2* edited by Butterworth, B., 3-50 (London: Academic Press).
Menn, L., K. L. Markey, M. Mozer, and C. Lewis. 1993. Connectionist modeling and the microstructure of phonological development: a progress report. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* edited by de Boysson-Bardies, B. and et al., 421-433 (Dordrecht: Kluwer).
Menn, L. and C. Stoel-Gammon. 1995. Phonological development. In *The handbook of child language* edited by Fletcher, P. and B. MacWhinney, 335-359 (Oxford: Blackwell).
Menn, L. 2000. Babies, buzzsaws and blueprints: commentary on review article by Sabbagh & Gelman. *Journal of Child Language* 27:753-755.
Menyuk, P., L. Menn, and D. Silverman. 1986. Early strategies for the perception and production of words and sounds. In *Language Acquisition* edited by Fletcher, P. and M. Garman, 198-222 (CUP).
Messer, D. 1994. Speech to children. In *The development of communication*, 231-236 (Chichester: John Wiley).
Messum, P. R. 2002. Learning and teaching vowels. *Speak Out! (Whitstable, IATEFL)* 29:9-27.
Messum, P. R. 2003. Phonetic consequences of the breath stream dynamics of spoken English. MPhil dissertation, University College London.
Messum, P. R. 2004. Autonomy, as soon as possible. *Speak Out! (Whitstable, IATEFL)* 32:12-23.
Middlebrooks, J. C. 2002. Auditory space processing: here, there or everywhere? *Nature Neuroscience* 5 (9):824-826.
Miller, N. E. and J. Dollard. 1941. *Social Learning and Imitation.* (London: Kegan, Paul).
Millikan, R. G. 2001. A theory of representation to complement TEC. *Behavioral and Brain Sciences* 24 (5):894-895.
Millikan, R. G. 2005. *Language: a biological model* (OUP).
Milner, A. D. and M. A. Goodale. (1995) *The visual brain in action.* Oxford University Press.
Mitchell, R. W. 2002. Imitation as a perceptual process. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 441-469 (Cambridge, MA: MIT Press).
Mitleb, F. M. 1984. Voicing effect on vowel duration is not an absolute universal. *Journal of Phonetics* 12:23-27.
Moerk, E. 1989. The fuzzy set called "imitations". In *The many faces of imitation in language learning* edited by Speidel, G. E. and K. E. Nelson, 277-303 (New York: Springer Verlag).
Mompeán González, J. A. 2004. Category overlap and neutralization: the importance of speakers' classifications in phonology. *Cognitive Linguistics* 15 (4):429-469.
Monsell, S. 1987. On the relation between lexical input and output pathways for speech. In *Language Perception and Production: Relationships between listening, speaking, reading and writing* edited by Allport, A., D. G. MacKay, W. Prinz, and E. Scheerer, 273-311 (London: Academic Press).
Moon, J. B., J. W. Folkins, A. E.Smith, and E. S. Luschei. 1993. Air pressure regulation during speech production. *Journal of the Acoustical Society of America* 94 (1):54-63.
Moore, B. R. 2004. The evolution of learning. *Biological Review* 79:301-335.
Moore, C. A. 2004. Physiologic development of speech production. In *Speech motor control* edited by Maassen, B., R. Kent, H. F. M. Peters, P. van Lieshout, and W. Hulstijn, 191-209 (Oxford: University Press).

Moosavi, S. H., G. P. Topulos, A. Hafer, R. Lansing, L. Adams, R. Brown, and R. Banzett. 2000. Acute partial paralysis alters perceptions of air hunger, work and effort at constant P(CO(2)) and V(E). *Respir.Physiol.* 122 (1):45-60.
Mooshammer, C., S. Fuchs, and D. Fischer. 1999. Effects of stress and tenseness on the production of CVC syllables in German. In *ICPhS99*, 409-412 (San Francisco).
Morrison, I. 2002. Knowing our imitations. *Trends in Cognitive Sciences* 6 (3):115-116.
Morrison, I. 2002. Genuine imitation. *Trends in Cognitive Sciences* 6 (9):367-368.
Morrison, I. 2002. Imitation: on the dot. *Trends in Cognitive Sciences* 6 (12):499-500.
Munhall, K., C. Fowler, S. Hawkins, and E. Saltzman. 1992. 'Compensatory shortening' in monosyllables of spoken English. *Journal of Phonetics* 20:225-239.
Munhall, K. 2001. Functional imaging during speech production. *Acta Psychologica* 107:95-117.
Murray, R. W. 2000. Syllable cut prosody in early middle English. *Language* 76 (3):617-654.
Naeser, M. A. 1970. *The American Child's Acquisition of Differential Vowel Duration.* (U o Wisconsin: W R & D Center for Cognitive Learning, Technical Report No. 144 (2 parts)).
Nasri, S., A. Namazie, J. Kreiman, J. A. Sercarz, B. R. Gerratt, and G. S. Berke. 1994. A pressure-regulated model of normal and pathologic phonation. *Otolaryngol Head Neck Surg.* 111 (6):807-815.
Nasri, S., A. Namazie, M. Ye, J. Kreiman, B. R. Gerratt, and G. S. Berke. 1996. Characteristics of an in vivo canine model of phonation with a constant air pressure source. *Laryngoscope* 106 (6):745-751.
Nathan, G. S. 1997. Review of Ritt 'Quantity Adjustment'. *Language* 73 (1):182-185.
Nathan, G. S. 1999. What functionalists can learn from formalists in phonology. In *Functionalism and formalism in linguistics, Vol 1* edited by Darnell, M., E. Moravcsik, F. Newmayer, M. Noonan, and K. Wheatley, 305-328 (Amsterdam: John Benjamins).
Nazzi, T. and J. Bertoncini. 2003. Before and after the vocabulary spurt: two modes of word acquisition? *Developmental Science* 6 (2):136-142.
Neath, I. and A. M. Surprenant. 2003. *Human Memory.* Second ed. (Thomson Wadsworth).
Nehaniv, C. L. and K. Dautenhahn. 2001. Like me? Measures of correspondence and imitation. *Cybernetics and Systems* 32:11-51.
Neisser, U. 1976. *Cognition and Reality.* (San Francisco: W.H. Freeman).
Neisser, U. 1987. Introduction: the ecological and intellectual bases of categorization. In *Concepts and Conceptual Development* edited by Neisser, U., 1-10 (CUP).
Neisser, U. 1994. Multiple systems: a new approach to cognitive theory. *European Journal of Cognitive Psychology* 6 (3):225-241.
Netsell, R., W. K. Lotz, J. E. Peters, and L. Schulte. 1994. Developmental patterns of laryngeal and respiratory function for speech production. *Journal of Voice* 8 (2):123-131.
Newson, J. 1979. The growth of shared understandings between infant and caregiver. In *Before speech: the beginning of interpersonal communication* edited by Bullowa, M., 207-222 (CUP).
Ní Chasaide, A. 1987. Glottal control of aspiration and of voicelessness. In *Proceedings of the XIth ICPhS, Vol*, 28-31 (6. Tallinn, Estonia).
Nicol, C. J. 1995. The social transmission of information and behaviour. *Applied Animal Behaviour Sciences* 44:79-98.
Nittrouer, S. 2005. Age-related differences in weighting and masking of two cues to word-final stop voicing in noise. *Journal of the Acoustical Society of America* 118 (2):1072-1088.
Nittrouer, S., S. Estee, J. H. Lowenstein, and J.Smith. 2005. The emergence of mature gestural patterns in the production of voiceless and voiced word-final stops. *Journal of the Acoustical Society of America* 117 (1):351-364.
Noble, J. and P. M. Todd. 2002. Imitation or Something Simpler? Modeling Simple Mechanisms for Social Information Processing. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 423-439 (Cambridge, MA: MIT Press).
Nolan, F. 1990. Who do phoneticians represent? *Journal of Phonetics* 18:453-464.
Nordström, P-E. 1977. Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics* 5:81-92.
Norman, J. 2002. Two visual systems and two theories of perception: an attempt to reconcile the constructivist and ecological approaches. *The Behavioral and Brain Sciences* 25 (1):73-96.
O'Connor, J. D. 1973. *Phonetics.* (Harmondsworth: Penguin).
Ochs, E. and B. B. Schieffelin. 1984. Language acquisition and socialization: three developmental stories. In *Culture Theory: Essays on mind, self and emotion* edited by Shweder, R. and R. LeVine, 276-320 (New York: CUP).
Ohala, J. 1970. Aspects of the control and production of speech. *Working Papers in Phonetics, UCLA* 15:156-164.
Ohala, J. 1976. A model of speech aerodynamics. *Report of Phonology Laboratory, Berkeley* 1:93-107.

Ohala, J., C. J. Riordan, and H. Kawasaki. 1980. Investigation of pulmonary activity in speech. *Report of the Phonology Lab, University of Calif, Berkeley,* 5:89-95.
Ohala, J. 1990. Respiratory activity in speech. In *Speech Production and Speech Modelling* edited by Hardcastle, W. J. and A. Marchal (London: Kluwer Academic Publ.).
Ohala, J. 1992. Length of low and high vowels. In *Speech Perception, Production and Linguistic Structure* edited by Tohkura, Y., E. Vatikiotis-Bateson, and Y. Sagisaka, 307 (Oxford: IOS Press).
Ohala, J. and R. Sprouse. 2003. Effects on speech of introducing aerodynamic perturbations. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 2913-2916 (Barcelona: Causal Productions).
Öhman, S. E. G. 1967. Word and sentence intonation: a quantitative model. *STL-QPSR* 2-3:20-54.
Öhman, S. E. G. 1975. What is it that we perceive when we perceive speech? In *Structure and Process in Speech Perception* edited by Cohen, A. and S. Nooteboom, 36-47 (Berlin: Springer).
Oller, D. K., R. E. Eilers, R. Urbano, and A. B. Cobo-Lewis. 1997. Development of precursors of speech in infants exposed to two languages. *Journal of Child Language* 24:407-425.
Oller, D. K. and R. E. Eilers. 1998. Interpretive and methodological difficulties in evaluating babbling drift. *Revue PArole* 7-8:147-163.
Otomo, K. and C. Stoel-Gammon. 1992. The acquisition of unrounded vowels in English. *Journal of Speech and Hearing Research* 35:604-616.
Otomo, K. 2001. Maternal responses to word approximations in Japanese children's transition to language. *Journal of Child Language* 28:29-57.
Papoušek, M. and H. Papoušek. 1989. Forms and functions of vocal matching in interactions between mothers and their precanonical infants. *First Language* 9:137-158.
Parton, D. A. 1976. Learning to imitate in infancy. *Child Development* 47:14-31.
Patel, A. D., A. Löfqvist, and W. Naito. 1999. The acoustics and kinematics of regularly-timed speech: A database and method for the study of the P-center problem. In *Proc. ICPhS XIV*, 405-408 (San Francisco).
Paul-Brown, D. and G. Yeni-Komshian. 1988. Temporal changes in word revisions by children and adults. *Journal of Speech and Hearing Research* 31:630-639.
Pawlby, S. J. 1977. Imitative interaction. In *Studies in Mother-Infant Interaction* edited by Schaffer, H. R., 203-223 (London: Academic Press).
Perkell, J. S., F. H. Guenther, H. Lane, M. L. Matthies, P. Perrier, J Vick, R. Wilhelms-Tricarico, and M. Zandipour. 2000. A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics* 28:233-272.
Perruchet, P. and S. Pacton. 2005. Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences* 10 (5):233-238.
Peterson, G. E. 1957. Breath Stream Dynamics. In *Manual of Phonetics* edited by Kaiser, L, 139-148 (Amsterdam: North Holland).
Pickett, J. M., H. T. Bunnell, and S. G. Revoile. 1995. Phonetics of intervocalic consonant perception: retrospect and prospect. *Phonetica* 52:1-40.
Pike, K. L. 1943. *Phonetics.* (Ann Arbor: U o Michigan Press).
Pines, M. 1984. Reflections on mirroring. *International Review of Psycho-Analysis* 11:27-42.
Pines, M. 1985. Mirroring and child development. *Psycholanalytic Inquiry* 5:211-231.
Pollock, K. E., D. M. Brammer, and C. F. Hagerman. 1993. An acoustic analysis of young children's productions of word stress. *Journal of Phonetics* 21:183-203.
Pörschmann, C. 2000. Influences of bone conduction and air conduction on the sound of one's own voice. *Acustica* 86:1038-1045.
Port, R. and R. Rotunno. 1979. Relation between voice onset time and vowel duration. *Journal of the Acoustical Society of America* 66 (3):654-662.
Port, R. F. 1981. Linguistic timing factors in combination. *Journal of the Acoustical Society of America* 69 (1):262-274.
Port, R. F., J. Dalby, and M. L. O'Dell. 1987. Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America* 81 (5):1574-1585.
Porter, R. J. and J. F. Lubker. 1980. Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research* 23 (3):593-602.
Postma, A. 2000. Detection of errors during speech production: a review of speech monitoring models. *Cognition* 77:97-131.
Priestly, T. M. S. 1980. Homonymy in child phonology. *Journal of Child Language* 7:413-427.
Pulvermüller, F., M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, and Y. Shtyrov. 2006. Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci USA.*
Rakerd, B., W. Sennett, and C. Fowler. 1987. Domain-final lengthening and foot-level shortening in spoken English. *Phonetica* 44:147-155.

Reisberg, D., J. D.Smith, D. A. Baxter, and M. Sonenshine. 1989. "Enacted" auditory images are ambiguous; "pure" auditory images are not. 41A ed., 619-641 ().
Repp, B. H. 1979. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech* 22 (2):173-189.
Repp, B. H. 1984. Categorical perception: issues, methods, findings. In *Speech and Language: Advances in Basic Research and Practice vol 10* edited by Lass, N. J., 244 (Academic Press).
Rietveld, T., J. Kerkhoff, and C. Gussenhoven. 1999. Prosodic structure and vowel duration in Dutch. In *ICPhS99*, 463-466 (San Francisco).
Rigamonti, M. M., D. Custance, E. P. Previde, and C. Spiezo. 2005. Testing for localized stimulus enhancement and object movement reenactment in pig-tailed macaques and young children. *Journal of Comparative Psychology* 119 (3):257-272.
Rizzolatti, G., L. Fadiga, V. Gallese, and L. Fogassi. 1996. Premotor cortex and the recognition of motor actions. *Brain Res Cogn Brain Res* 3 (2):131-141.
Rizzolatti, G., L. Fogassi, and V. Gallese. 2006. The mirrors in the mind. *Scientific American* (November):30-37.
Robb, M. P. and K. M. Bleile. 1994. Consonant inventories of young children from 8 to 25 months. *Clinical Linguistics and Phonetics* 8 (4):295-320.
Rochat, P. 2001. Origins of self-concept. In *Blackwell Handbook of Infant Development* edited by Bremner, G. and A. Fogel, 191-212 (Oxford).
Romani, C. 1992. Are there distinct input and output buffers? Evidence from an aphasic patient with an impaired output buffer. *Language and Cognitive Processes* 7 (2):131-162.
Romanyshyn, R. D. 1982. *Psychological Life: From Science to Metaphor.* (Milton Keynes: OU Press).
Rosenbaum, D. A. and H. Krist. 1996. Antecedents of action. In *Handbook of perception and action: volume 2: motor skills* edited by Heuer, H. and S. W. Keele, 3-69 (London: Academic Press).
Rothenberg, M. 1968. *The breath-stream dynamics of simple-released-plosive production.* (Basel: S. Karger).
Rothenberg, M., D. G. Miller, R. Molitor, and D. Leffingwell. 1987. The control of air flow during loud soprano singing. *Journal of Voice* 1 (3):262-268.
Russell, N. K. and E. Stathopoulos. 1988. Lung volume changes in children and adults during speech production. *Journal of Speech and Hearing Research* 31:146-155.
Saffran, J. R. 2003. Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science* 12 (4):110-114.
Sampson, G. 1980. *Schools of Linguistics.* (London: Hutchinson).
Sapir, E. 1921. *Language.* (New York: Harcourt, Brace and World).
Schmuckler, M. A. 1993. Perception-action coupling in infancy. In *The development of coordination in infancy* edited by Savelsbergh, G. J. P., 137-173 (Amsterdam: North Holland).
Schutte, H. K. 1984. Efficiency of professional singing voices in terms of energy ratio. *Folia Phoniatrica* 36:267-272.
Schwartz, R. G., K. Petinou, L. Goffman, G. Lazowski, and C. Cartusciello. 1996. Young children's production of syllable stress: an acoustic analysis. *Journal of the Acoustical Society of America* 99 (5):3192-3200.
Scobbie, J. M., F. Gibbon, W. J. Hardcastle, and P. Fletcher. 1997. Longitudinal phonological and phonetic analyses of two cases of disordered /s/+stop cluster acquisition. In *Proceedings of the GALA 97 conference on language acquisition* edited by Sorace, A., C. Heycock, and R. Shillcock, 278-283 ().
Scott, S. K. 1998. The point of P-centres. *Psychological Research* 61 (1):4-11.
Scott, S. K. and I. S. Johnsrude. 2003. The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences* 26 (2):100-107.
Scott, S. K. and R. J. S. Wise. 2004. The functional neuroanatomy of prelexical processing in speech perception. *Cognition* 92:13-45.
Scott, S. K. 2005. Auditory processing — speech, space and auditory objects. *Current Opinion in Neurobiology* 15:197-201.
Scully, C. 1990. Speech production modelling with particular reference to English. PhD thesis, London.
Scully, C. 1992. L'importance des processus aerodynamiques dans la production de la parole. In *Actes 19èmes Journées d'Etude sur la Parole*, 7-12 (Bruxelles: Université Libre, Institut de Phonetique).
Shaffer, L. H. 1982. Rhythm and timing in skill. *Psychological Review* 89 (2):109-122.
Shanks, D. R. 2005. Implicit learning. In *Handbook of Cognition* edited by Lamberts, K. and R. Goldstone, 202-220 (London: Sage).
Sharf, D. 1964. Vowel duration in whispered and in normal speech. *Language and Speech* 7:89-97.
Shattuck-Hufnagel, S. 1986. Comment on Studdert-Kennedy 'Sources of variability in early speech development': Why we need more data. In *Invariance and variability in speech processes* edited by Perkell, J. and D. Klatt, 77-84 (Hillsdale, New Jersey: LEA).

Sheldon, A. and W. Strange. 1982. The acquisition of /r/ and /l/ by Japanese learners of English: evidence that speech production can precede speech perception. *Applied Psycholinguistics* 3:243-261.
Shelton, J. R. and A. Caramazza. 1999. Deficits in lexical and semantic processing: implications for models of normal language. *Psychonomic Bulletin and Review* 6 (1):5-27.
Shockley, K., L. Sabadini, and C. A. Fowler. 2004. Imitation in shadowing words. *Perception and Psychophysics* 66 (3):422-429.
Shuster, L. I., D. M. Ruscello, and A. R. Toth. 1995. The use of visual feedback to elicit correct /r/. *American Journal of Speech-Language Pathology* 4 (2):37-44.
Shuster, L. I. 1998. Linear predictive coding parameter manipulation/synthesis of incorrectly produced /r/. *Journal of Speech, Language and Hearing Research* 39:827-832.
Shuster, L. I. and J. D. Durrant. 2003. Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders* 36:1-11.
Shvachkin, N. K. 1973. The development of phonemic speech perception in early childhood. In *Studies of Child Language Development* edited by Ferguson, C. A. and D. I. Slobin, 91-127 (New York: Holt, Rinehart and Winston).
Sievers, G. E. 1901. Silbenbildung (on different types of syllable). In *Grundzüge der Phonetik*, 198-206 (Leipzig: Breitkopf & Härtel).
Simon, C. and A. Fourcin. 1978. Cross-language study of speech-pattern learning. *Journal of the Acoustical Society of America* 63 (3):925-935.
Skoyles, J. 1998. Speech phones are a replication code. *Medical Hypotheses* 50:167-173.
Skoyles, J. 2002. Mirror neurons and articulation, and the origin of speech. Evolution of Language 2002, Fourth International Conference, Harvard .
Skoyles, J. 2002. Language and imitation: Informational processing and the elementary units of speech. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 130-138 (Cambridge, MA: MIT Press).
Slifka, J. 2003. Respiratory constraints on speech production: Starting an utterance. *Journal of the Acoustical Society of America* 114 (6):3343-3353.
Sloman, S. A. and L. J. Rips. 1998. Similarity as an explanatory construct. *Cognition* 65:87-101.
Sluijter, A. C. M. and V. J. van Heuven. 1995. Effects of focus distribution, pitch accent and lexical stress on the temporal organisation of syllables in Dutch. *Phonetica* 52:71-89.
Sluijter, A. C. M. and V. J. van Heuven. 1995. Intensity and vocal effort as cues in the perception of stress. ESCA Eurospeech '95. 941-944. Madrid.
Sluijter, A. M. C. and V. J. van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100:2471-2485.
Sluijter, A. M. C., V. J. van Heuven, and J. J. A. Pacilly. 1997. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America* 101 (1):503-513.
Smith Hammond, C., D. W. Warren, R. Mayo, and D. Zajac. 1999. Respiratory responses to sudden pressure venting during stop consonant production. *Folia Phoniatrica et Logopaedica* 51:250-260.
Smith, A. and L. Goffman. 2004. Interaction of motor and language factors in the development of speech production. In *Speech motor control* edited by Maassen, B., R. Kent, H. F. M. Peters, P. van Lieshout, and W. Hulstijn, 227-252 (Oxford: University Press).
Smith, B. L. 1978. Temporal aspects of English speech production: a developmental perspective. *Journal of Phonetics* 6:37-67.
Smith, B. L. 1979. A phonetic analysis of consonantal devoicing in children's speech. *Journal of Child Language* 6:19-28.
Smith, B. L. and M. K. Kenney. 1999. A longitudinal study of the development of temporal properties of speech production: data from 4 children. *Phonetica* 56:73-102.
Smith, B. L. 2002. Effects of speaking rate on temporal patterns of English. *Phonetica* 59:232-244.
Smith, B. L., A. R. Bradlow, and T. Bent. 2003. Production and perception of temporal contrasts in foreign-accented English. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 519-522 (Barcelona: Causal Productions).
Smith, C. L. 1997. The devoicing of /z/ in American English: effects of local and prosodic context. *Journal of Phonetics* 25:471-500.
Smith, N. V. 1973. *The Acquisition of Phonology*. (CUP).
Smith, N. V. 2003. Are gucks mentally represented? *Glot International* 7 (6):164-166.
Smith, V. A., A. P. King, and M. J. West. 2000. A role of her own: female cowbirds, *Molothrus ater*, influence the development and outcome of song learning. *Animal Behaviour* 60:599-609.
Snow, C. E. 1977. The development of conversation between mothers and babies. *Journal of Child Language* 4:1-22.
Snow, C. E. 1988. The last word: questions about the emerging lexicon. In *The Emergent Lexicon: the Child's Development of a Linguistic Vocabulary* edited by Locke, J. L. and M. D.Smith, 341-353 (New York: Academic Press).

Solomon, N. P. and S. Charron. 1998. Speech breathing in able-bodied children and children with cerebral palsy: a review of the literature and implications for clinical intervention. *American Journal of Speech-Language Pathology.*

Solomon, N. P., K. D. R. Drager, and E. S. Luschei. 2002. Sustaining a constant effort by the tongue and hand: effects of acute fatigue. *Journal of Speech, Language and Hearing Research* 45:613-624.

Stathopoulos, E. T. and C. Sapienza. 1993. Respiratory and laryngeal function of women and men during vocal intensity variation. *Journal of Speech and Hearing Research* 36:64-75.

Stathopoulos, E. T. 1995. Variability revisited: an acoustic, aerodynamic and respiratory kinematic comparison of children and adults during speech. *Journal of Phonetics* 23:67-80.

Stathopoulos, E. T. 2000. A review of the development of the child voice: an anatomical and functional perspective. In *Child Voice* edited by White, P., 1-12 (Stockholm: KTH).

Stern, D. N. 1985. *The Interpersonal World of the Infant.* (London: Karnac Books).

Sternberg, S., S. Monsell, R. L. Knoll, and C. E. Wright. 1978. The latency and duration of rapid movement sequences: comparisons of speech and typewriting. In *Information processing in motor control and learning* edited by Stelmach, G. E. (New York: Academic Press).

Sternberg, S., C. E. Wright, R. L. Knoll, and S. Monsell. 1980. Motor programs in rapid speech: additional evidence. In *Perception and production of fluent speech* edited by Cole, R. A., 507-534 (Hillsdale, NJ: LEA).

Sternberg, S., R. L. Knoll, S. Monsell, and C. E. Wright. 1988. Motor programs and hierarchical organization in the control of rapid speech. *Phonetica* 45:175-197.

Stetson, R. H. 1951. *Motor Phonetics.* (Amsterdam: North Holland).

Stevens, K. 1998. *Acoustic Phonetics.* (Cambridge, MA: MIT Press).

Stevens, K. N., S. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America* 91 (5):2979-3000.

Stevens, K. N. 1997. Articulatory-acoustic-auditory relationships. In *The Handbook of Phonetics* edited by Hardcastle, W. H. and J. Laver, 490-501 (Oxford: Blackwell).

Stevick, E. W. 1980. *Teaching Languages: A Way and Ways.* (Newbury House).

Stevick, E. W. 1990. *Humanism in Language Teaching.* (OUP).

Stoel-Gammon, C., E. H. Buder, and M. M. Kehoe. 1995. Acquisition of vowel duration: a comparison of Swedish and English. In *XIIIth ICPhS, Volume 4* edited by Elenius, K. and P. Banderud, 30-36 (Stockholm).

Stoel-Gammon, C. and E. H. Buder. 1999. Vowel length, post-vocalic voicing and VOT in the speech of two-year-olds. In *XIVth ICPhS 99, Volume 3*, 2485-2488 (San Francisco).

Stoinski, T. S., J. L. Wrate, N. Ure, and A. Whiten. 2001. Imitative learning by captive Western Lowland gorillas in a simulated food-processing task. *Journal of Comparative Psychology* 115 (3):272-281.

Straight, H. S. 1980. Auditory versus articulatory phonological processes and their development in children. In *Child Phonology: Vol 1, Production* edited by Yeni-Komshian, G., J. Kavanagh, and C. A. Ferguson, 43-71 (NY: Academic Press).

Straight, H. S. 1982. The formulation-interpretation circuit: a linguistic motor for the creation of meaning. *Quaderni di Semantica* 3:123-128.

Straight, H. S. 1986. The importance and irreducibility of the comprehension/production dialectic. In *Language for Hearers* edited by McGregor, G., 69-90 (Oxford: Pergamon).

Straight, H. S. 1992. Processing: Comprehension and production. In *Encyclopedia of Linguistics* edited by Bright, W., 271-273 (Oxford: OUP).

Straight, H. S. 1993. Processualism in linguistic theory and method. In *Linguistics and philosophy* edited by Harré, R. and R. Harris, 199-216 (Oxford: Pergamon).

Studdert-Kennedy, M. 1986. Sources of variability in early speech development. In *Invariance and variability in speech processes* edited by Perkell, J. and D. Klatt, 58-84 (Hillsdale, New Jersey: LEA).

Studdert-Kennedy, M. 1987. The phoneme as a perceptuomotor structure. In *Language Perception and Production: Relationships between listening, speaking, reading and writing* edited by Allport, A., D. G. MacKay, W. Prinz, and E. Scheerer, 67-84 (London: Academic Press).

Studdert-Kennedy, M. 2000. Imitation and the emergence of segments. *Phonetica* 57:275-283.

Studdert-Kennedy, M. 2002. Mirror neurons, vocal imitation, and the evolution of particulate speech. In *Mirror Neurons and the Evolution of Brain and Language* edited by Stamenov, M. I. and V. Gallese, 207-227 (Amsterdam: John Benjamins).

Sundberg, J., I. Titze, and R. Scherer. 1993. Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency and mode of phonation on the voice source. *Journal of Voice* 7:15-29.

Sundberg, J. 1995. Subglottal pressure behaviour in singing and speech. In *XIIIth ICPhS, Volume 3* edited by Elenius, K. and P. Banderud, 98-104 (Stockholm).

Sundberg, J., M. Andersson, and C. Hultqvist. 1998. Effects of subglottal pressure variation on professional baritone singer's voice source. *TMH-QPSR* 1-2:1-8.
Sundberg, U. and F. Lacerda. 1999. Voice onset time in speech to infants and adults. *Phonetica* 56:186-199.
Suomi, K. 1993. An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics* 21:29-60.
Sutton, R. S. and A. G. Barto. 1998. *Reinforcement Learning.* (Cambridge, MA: MIT Press).
Sutton, R. S. 1999. Reinforcement learning. In *The MIT Encyclopedia of the Cognitive Sciences* edited by Wilson, R. and F. Keil (Cambridge, MA: MIT Press).
Swingley, D. and R. N. Aslin. 2002. Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science* 13 (5):480-484.
Swingley, D. 2005. 11-month-olds' knowledge of how familiar words sound. *Developmental Science* 8 (5):432-434.
Tang, J. and E. T. Stathopoulos. 1995. Vocal efficiency as a function of vocal intensity: a study of children, women and men. *Journal of the Acoustical Society of America* 97:1885-1892.
Thelen, E. 1995. Motor development: a new synthesis. *American Psychologist* 50 (2):79-95.
Thomas, N. J. T. 2001. "Re: Mental Images Query." Available from http://mailgate.supereva.com/sci/sci.psychology.consciousness/msg00436.html.
Thyme-Frøkjær, K. and B. Frøkjær-Jensen. 2001. *The Accent Method.* (Bicester: Speechmark).
Titze, I. R. 1994. *Principles of Voice Production.* (Englewood Cliffs: Prentice Hall).
Trevarthen, C. and K. J. Aitken. 2001. Infant intersubjectivity: research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry* 42 (1):3-48.
Tsushima, T. 2005. Relation between perception and production ability during a speech training course. *Journal of the Acoustical Society of America* 117 (4.2).
Tucha, O., L. Mecklinger, S. Walitza, and K. W. Lange. 2006. Attention and movement execution during handwriting. *Human Movement Science* 25:536-552.
Turk, A. E. and S. Shattuck-Hufnagel. 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics* 28:397-440.
Tye-Murray, N. and G. Woodworth. 1989. The influence of final-syllable position on the vowel and word duration of deaf talkers . *Journal of the Acoustical Society of America* 85 (1):313-321.
Uguzzoni, A., G. Azzaro, and S. Schmid. 2003. Short vs long and/or abruptly vs smoothly cut vowels. New perspectives on a debated question. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 2717-2720 (Barcelona: Causal Productions).
Underhill, A. 1994. *Sound Foundations.* (Oxford: Heinemann ELT).
Vallabha, G. K. and B. Tuller. 2004. Perceptuomotor bias in the imitation of steady-state vowels. *Journal of the Acoustical Society of America* 116 (2):1184-1197.
Van Dam, M. 2003. VOT of American English stops with prosodic correlates. *Journal of the Acoustical Society of America* 113 (4.2):2328.
van der Kamp, J. and G. J. P. Savelsbergh. 2000. Action and perception in infancy. *Infant Behavior and Development* 23:237-251.
van Heuven, V. J. and A. M. C. Sluijter. 1996. Notes on the phonetics of word prosody. In *Stress patterns of the world, Part 1 Background* edited by Goedemans, R., H. van der Hulst, and E. Visch (Leiden: HIL Publications).
Van Lancker, D., J. Kreiman, and D. W. Bolinger. 1988. Anticipatory lengthening. *Journal of Phonetics* 16:339-347.
van Santen, J. P. H. and C. Shih. 2000. Suprasegmental and segmental timing models in Mandarin Chinese and American English. *J.Acoust.Soc.Am.* 107 (2):1012-1026.
Velleman, S. L., L. Mangipudi, and J. L. Locke. 1989. Prelinguistic phonetic contingency: data from Down syndrome. *First Language* 9:159-174.
Veneziano, E. 1988. Vocal-verbal interaction and the construction of early lexical knowledge. In *The Emergent Lexicon: the Child's Development of a Linguistic Vocabulary* edited by Locke, J. L. and M. D.Smith, 109-147 (New York: Academic Press).
Vennemann, T. 2000. From quantity to syllable cuts: On so-called lengthening in the Germanic languages. *Rivista di Linguistica* 12 (1):251-282.
Vihman, M. M. 1993. Variable paths to early word production. *Journal of Phonetics* 21:61-82.
Vihman, M. M. 1996. *Phonological development.* (Cambridge, MA: Blackwell).
Vihman, M. M. and S. L. Velleman. 2000. Phonetics and the origins of phonology. In *Phonological Knowledge* edited by Burton-Roberts, N., P. Carr, and G. Docherty, 305-339 (OUP).
Vihman, M. M. 2002. The role of mirror neurons in the ontogeny of speech. In *Mirror Neurons and the Evolution of Brain and Language* edited by Stamenov, M. I. and V. Gallese, 305-314 (Amsterdam: John Benjamins).

Vihman, M. M. and S. Nakai. 2003. Experimental evidence for an effect of vocal experience on infant speech perception. In *15th ICPhS* edited by Solé, M. J., D. Recasens, and J. Romero, 1017-1020 (Barcelona: Causal Productions).
Vihman, M. M., S. Nakai, R. A. DePaolis, and P. A. Hallé. 2004. The role of accentual pattern in early lexical representation. *Journal of Memory and Language* 50:336-353.
Vinter, A. 1986. A developmental perspective on behavioral determinants. *Acta Psychologica* 63:337-349.
Vinter, A. 1999. How meaning modifies drawing behaviour in children. *Child Development* 70 (1):33-49.
Vogt, S. 2002a. Visuomotor couplings in object-oriented and imitative actions. In *The imitative mind: development, evolution and brain bases* edited by Meltzoff, A. N. and W. Prinz, 206-220 (Cambridge: CUP).
Vogt, S. 2002b. Dimensions of Imitative Perception-Action Mediation. In *Imitation in Animals and Artifacts* edited by Dautenhahn, K. and C. L. Nehaniv, 525-554 (Cambridge, MA: MIT Press).
von Euler, C. 1982. Some aspects of speech breathing physiology. In *Speech Motor Control* edited by Grillner, S., B. Lindblom, J. Lubker, and A. Penrose, 95-103 (London: Pergamon).
Walley, A. C. 1993. The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review* 13:186-350.
Walley, A. C. 2005. Speech perception in childhood. In *The Handbook of Speech Perception* edited by Pisoni, D. and R. E. Remez, 449-468 (Oxford: Blackwell).
Walsh, T. and F. Parker. 1982. Consonant cluster abbreviation: an abstract analysis. *Journal of Phonetics* 10:423-437.
Want, S. C. and P. L. Harris. 2002. How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children. *Developmental Science* 5 (1):1-13.
Want, S. C. and P. L. Harris. 2002. Social learning: compounding some problems and dissolving others. *Developmental Science* 5 (1):39-41.
Wardrip-Fruin, C. and S. Peach. 1984. Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants. *Language and Speech* 27 (4):367-379.
Warren, D. W. 1988. Aerodynamics of speech. In *Handbook of Speech - Language, Pathology and Audiology* edited by Lass, N. J. (Toronto: BC Decker).
Warren, D. W., R. M. Dalston, and E. T. Dalston. 1990. Maintaining speech pressures in the presence of velopharyngeal impairment. *Cleft Palate Journal* 27 (1):53-57.
Warren, D. W. 1996. Regulation of Speech Aerodynamics. In *Principles of Experimental Phonetics* edited by Lass, N. J., 46-92 (St Louis: Mosby).
Warren, J. E., R. J. S. Wise, and J. D. Warren. 2005. Sounds do-able: auditory-motor transformations and the posterior temporal plane. *Trends in Neurosciences* 28 (12):636-643.
Warren, P. 1999. Timing properties of New Zealand English rhythm. In *ICPhS99*, 1843-1846 (San Francisco).
Weismer, G. 1979. Sensitivity of VOT measures to certain segmental features in speech production. *Journal of Phonetics* 7:196-204.
Weiss, R. 1976. *The Perception of Vowel Length and Quality in German.* (Hamburg: Helmut Buske).
Welford, A. T. 1976. Changes of performance during long practice. In *Skilled performance: perceptual and motor skills*, 121-124 (Glenview Il: Scott, Foreman).
Wells, J. C. 1982. *Accents of English.* Vol. 3 (CUP).
Wells, J. C. 1990. *Longman Pronunciation Dictionary.* (Harlow: Longman).
Welty, E. 1983. *One Writer's Beginnings.* (New York: Warner Books).
Werker, J. F., C. T. Fennell, K. M. Corcoran, and C. L. Stager. 2002. Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy* 3 (1):1-30.
Werker, J. F. and H. H. Yeung. 2005. Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences* 9 (11):519-527.
Werker, J. F. and S. Curtin. 2005. PRIMIR: a developmental framework of infant speech processing. *Language Learning and Development* 1 (2):197-234.
West, R. and L. H. Turner. 2002. Communication accommodation theory. In *Introducing Communication Theory* (McGraw-Hill).
Westermann, G. and E. R. Miranda. 2002. Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research* 31 (4):367-375.
Westermann, G. and E. R. Miranda. 2004. A new model of sensorimotor coupling in the development of speech. *Brain and Language* 89:393-400.
Westwood, D. A., C. D. Chapman, and E. A. Roy. 2000. Pantomimed actions may be controlled by the ventral visual stream. *Experimental Brain Research* 130:545-548.
Westwood, D. A. and M. A. Goodale. 2001. Perception and action planning: getting it together. *The Behavioral and Brain Sciences* 24 (5):907-908.

Whalen, D. H., A. M. Cooper, and C. A. Fowler. 1989. P-center judgments are generally insensitive to the instructions given. *Phonetica* 46 (4):197-203.
White, L. S. 2002. *English speech timing: a domain and locus approach.* PhD thesis, University of Edinburgh.
Whiten, A. 2002. The imitator's representation of the imitated: Ape and child. In *The imitative mind: development, evolution and brain bases* edited by Meltzoff, A. N. and W. Prinz, 98-121 (Cambridge: CUP).
Whiten, A., V. Horner, C. A. Litchfield, and S. Marshall-Pescini. 2004. How do apes ape? *Learning and Behavior* 32 (1):36-52.
Whiten, A., V. Horner, and S. Marshall-Pescini. 2005. Selective imitation in child and chimpanzee: a window on the construal of others' actions. In *Perspectives on Imitation. Volume 1: Mechanisms of Imitation and Imitation in Animals* edited by Hurley, S. and N. Chater, 263-283 (Cambridge MA: MIT Press).
Whiten, A. 2005. The imitative correspondence problem: solved or sidestepped? In *Perspectives on Imitation. Volume 1: Mechanisms of Imitation and Imitation in Animals* edited by Hurley, S. and N. Chater, 220-222 (Cambridge MA: MIT Press).
Whiteside, S. P. and J. Marshall. 2001. Developmental trends in VOT: some evidence for sex differences. *Phonetica* 58:196-210.
Wijnen, F. and H. H. J. Kolk. 2005. Conclusions and prospects. In *Phonological encoding and monitoring in normal and pathological speech* edited by Hartsuiker, R. J., R. Bastiaanse, A. Postma, and F. Wijnen (Hove, UK: Psychological Press).
Willis, G. 1919. *The philosophy of speech.* (London: G. Allen & Unwin).
Wilson, C. and A. L. Woodward. 2002. A window to the structure of the mind: review of Meltzoff and Prinz "The Imitative Mind". *Trends in Cognitive Sciences* 6 (12):537-538.
Wilson, M. 2001. Perceiving imitatible stimuli: consequences of isomorphism between input and output. *Psychological Bulletin* 127 (4):543-553.
Wilson, M. 2004. Motoric emulation may contribute to perceiving imitable stimuli. *The Behavioral and Brain Sciences* 27 (3):424.
Winnicott, D. W. 1971. The mirror role of mother and family in child development. In *Playing and Reality* (London: Tavistock Clinic).
Wise, R. J. S., S. K. Scott, S. C. Blank, C. J. Mummery, K. Murphy, and E. A. Warburton. 2001. Separate neural subsystems within 'Wernicke's area'. *Brain* 124 (83):95.
Wolpert, D. M., K. Doya, and M. Kawato. 2003. A unifying computational framework for motor control and social interaction. *Phil Trans R Soc Lond* 358:593-602.
Wood, D. 1989. Social Interaction as Tutoring. In *Interaction in Human Development* edited by Bornstein, M. and J. Bruner, 59-80 (Hillsdale, NJ: LEA).
Wood, S. A. J. 1975. The weakness of the tongue-arching model of vowel articulation. *Working Papers, Department of Linguistics, Lund University* 11:55-108.
Wulf, G., N. McNevin, T. Fuchs, F. Ritter, and T. Toole. 2000. Attentional focus in complex skill learning. *Research Quarterly for Exercise and Sport* 71 (3):229-239.
Yamamoto, F. 1996. English speech rhythm studied in connection with British traditional music and dance. *Journal of Himeji Dokkyo University Gaikokugogakubo* 9:224-243.
Yando, R., V. Seitz, and E. Zigler. 1978. *Imitation: A developmental perspective.* (Hillsdale, NJ: LEA).
Yavas, M. 1998. *Phonology: Development and Disorders.* (San Diego: Singular).
Yoshikawa, Y., M. Asada, K. Hosoda, and J. Koga. 2003. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science* 14 (4):245-258.
Young, R. 1984. The Silent Way. In *New Approaches in Foreign Language Methodology: 15th AIMAV Colloquium* edited by Knibbeler, W. and M. Bernards, 99-104 (Department of Applied Linguistics, University of Nijmegan).
Young, R. 1990. Universaux dans l'enseignement et l'apprentissage du français et de l'anglais dans des situations pédagogiques diverses. PhD thesis, Université de Franche-Comté.
Young, R. 1995. Caleb Gattegno's "Silent Way": some of the reasons why. *Methoden der Fremdsprachenvermittlung (University of Mainz)*:55-74.
Young, R. 2000. Round table discussion (Nusbaum, Stevick, Thornbury, Young): Under what circumstances do we apply the word 'scientific' in language learning? *Prism (Paris) - a learning journal* 5:7-183.
Zemlin, W. R. 1988. *Speech and Hearing Science: Anatomy and Physiology.* (Prentice Hall (3rd Ed)).
Zesiger, P., M-D. Martory, and E. Mayer. 1997. Writing without graphic motor patterns: a case of dysgraphia for letters and digits sparing shorthand writing. *Cognitive Neuropsychology* 14 (5):743-763.
Zhang, G. 1996. Foot-timing and word-timing in English. PhD thesis, University of Delaware.
Zukow-Goldring, P. 1996. Sensitive caregiving fosters the comprehenstion of speech: when gestures speak louder than words. *Early development and parenting* 5 (4):195-211.