ANALYSIS OF THE TIMING OF SPOKEN KOREAN WITH APPLICATION TO SPEECH SYNTHESIS

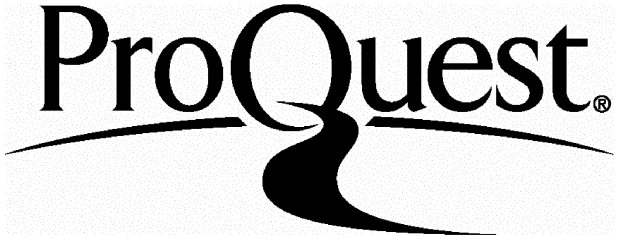
Hyunsong Chung

Department of Phonetics and Linguistics

University College London

A thesis submitted to the University of London
for the degree of Doctor of Philosophy

2002



ABSTRACT

The thesis describes new analysis and modelling of Korean segmental duration. It takes into account contemporary approaches to duration modelling, as used in English and Japanese synthesis to build predictive models of segment duration in context which could be used in Korean language text-to-speech (TTS) systems. It also analyses those models to learn more about which factors and which structures are most important in Korean prosody. The thesis concentrates on the duration modelling of a news-reading speech style; using a corpus of 670 read sentences collected from one speaker of standard Korean. The duration of each segment and its phonological context were extracted from the corpus. Statistical modelling explored the relationship between the context features and the realised duration. Based on previous research on timing, Sums-of-Products models and Classification And Regression Tree (CART) models were applied and evaluated on the data. Objective quality of the modelling was evaluated by root mean squared prediction error (RMSE) and the correlation coefficient between actual and predicted durations in reserved test data. The best performance result was obtained from a CART model with an RMSE of 25.11 ms and a correlation of 0.77; a result which was comparable with other published results on Korean segment durations. Analysis showed that prosodic phrase features have the greatest influence on segment duration, among them, the accentual phrase final position feature. In terms of segmental context, surrounding nasals were shown to have consistent shortening effect, while vowels seemed to be affected by the degree of glottal opening of adjacent consonants. Other segmental effects were less consistent. Perceptual tests show a slight listener preference for durations calculated from a CART model in this thesis compared to durations calculated from a commercial Korean TTS system.

ACKNOWLEGMENTS

My Ph.D. work was partly funded by the Overseas Research Students awards scheme (ORS) from the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom (CVCP) between 1998 and 2000 (ORS/C20/4 and ORS/C17/3). Funding for travel to several conferences to present some of the results from this thesis was also partly supported by the UCL Graduate School in 1999 and in 2001 and by the International Speech Communication Association (ISCA) in 2001. I would like to express my deepest gratitude to my supervisor, Dr Mark Huckvale, for his invaluable comments and challenges during my studies and in the preparation of this thesis. He has been an excellent supervisor, a good friend, and even been a good teacher of English to me. My thanks are also due to Professor Neil Smith, who has given me helpful and friendly advice whenever I consulted him with my academic and personal matters. Without the great support from the staff of the department, it would not have been possible to complete my thesis. My thanks go to Dr Andrew Faulkner, Dr Valerie Hazan, Dr Paul Iverson, Professor Stuart Rosen, David Cushing, Martyn Holland, Steve Nevard, and Warwick Smith. Two fellow students in the "boys room", Abbas Haydari and Gordon Hunter, and also Alex Fang deserve special thanks for their good humour and encouragement. Gordon especially contributed much to the scripts for text processing which were used in chapter 4. Dr Tae-yeoub Jang and Weonhee Yun helped me a lot in carrying out the letter-to-phoneme conversion and the phone alignment of the phonetic transcriptions to the speech signals in chapter 4. I would like to thank Professor Gyeongseog Gim in Busan National University—it was his original idea to work together to design the MBROLA Korean language diphone database which was used for the perceptual evaluation in chapter 6. I would like to give my special thanks to Professor Kook Chung of the Hankuk University of Foreign Studies, who was my MA thesis supervisor in Korea. He led me into the world of linguistics and encouraged me to extend my knowledge to phonetics and speech science. Despite the long distance between Seoul and London, he was always there whenever I needed his help. This thesis is dedicated to my wife, Sun Young and my parents, Kyung Sik and Chan Jo. Sun Young has shared every moment of happiness and all difficulties with me. Without her sacrifice, this thesis could not be produced. My parents have supported me by their prayers and encouragement. I am thankful for their endless love and much more.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1. INTRODUCTION

There has been a great increase in the commercial development and deployment of speech synthesis systems in recent years. This has been driven by two forces: the explosion in use of information services directed at mobile phone users, and the connections of a large number of ordinary home users to the Internet. This growth has increased concerns about the perceived naturalness of synthetic voices, since the acceptability of these information services seems to be closely tied to the quality of the synthetic speech. This has led to the emergence of "engineering" solutions to speech synthesis—solutions which are not based on cognitive or acoustic models of human speech production, but models which simply aim to sound like a human speaking. These solutions, called concatenative (Dutoit and Leich, 1993) or corpus-based synthesis (Hunt and Black, 1996; Black and Taylor, 1997; Taylor, 1999), have become possible because of the increasing power and capacity of contemporary digital computer systems. This type of speech synthesis relies on the creation of large, well-described and accurately-labelled databases of recorded speech. However these in turn require a good understanding of the prosodic and phonological structure of the language and of its phonetic interpretation. The reason for this is that systems locate sections of recorded speech in the database using linguistic descriptors and the linguistic descriptors must be applied reliably to the corpus and to the target text. These linguistic descriptors must be capable of describing the paradigmatic alternatives in prosody and segmental form, and the contexts within which prosodic and segmental variability takes place. The linguistic descriptors must also be aligned accurately to the signal. Thus these kinds of synthesis system change both the form and the motivation of experimental phonetics research. For example, there is now much more emphasis on prosody than on coarticulatory behaviour.

The challenge for research in experimental phonetics in the new millennium is to contribute to both better synthesis systems and a better understanding of human speech production. We see this in the research on prosodic phrasing, intonation modelling, and duration modelling that is actively under way in many languages.

By far the largest efforts in speech technology have been applied to the main Western languages and to Japanese. Only a small amount has so far been applied to the Korean language. Indeed there is anecdotal evidence that progress in Korean speech technology is behind other languages[1]. One particular area of concern is in the analysis and modelling of the prosody of Korean, particularly in the area of segmental durations. There have been very few studies in this area; and those that have been conducted are poorly suited to the issues in contemporary speech synthesis systems. This thesis sets out to perform a new analysis and modelling of Korean segmental duration. These studies are based on previous work where possible, but extended to take into account the demands of contemporary approaches to duration modelling as used in English and Japanese synthesis. However, this thesis does not just try to build the best predictive model of segment duration in context. It also seeks to learn more about which factors and which structures are most important in Korean prosody. The outcome of this work is both a better model of Korean timing for use in synthesis, and a better understanding of the Korean language.

In chapter 2, linguistic aspects of timing in English and Korean are reviewed. The linguistic factors considered for English in this chapter are: (1) the property of the target segment; (2) the property of surrounding segments; and (3) prosodic effects.

For the Korean language, we first describe its phonological system and then its phonological pattern in terms of phonemes and allophones. The distinctive features of each phoneme are also investigated, which are used for the processing described in chapter 4. Chapter 2 also describes the phrase structure of Korean as well as its syllable structure. Phrase structure decisions are shown to be based on acoustic distinctiveness and phonological evidence. Some arguments about Korean syllable structure are also introduced. Finally, factors affecting the duration of Korean are investigated, such as inherent duration, prosodic effects, syllable structure, and contextual effects of surrounding segments.

In chapter 3, approaches to duration modelling which have been applied in TTS systems are introduced. In an overview of the general structure of TTS systems, three approaches for TTS are briefly described: rule-based synthesis, concatenative synthesis, and corpus-based synthesis. For the modelling of timing, three modelling methods are investigated: sequential rule systems, CART (Classification and Regression Tree) models, and sums-of-products models. These three models are compared and their advantages and disadvantages are discussed. Chapter 3 also investigates the use of linguistic information in duration modelling. While all contextual effects are descriptively important in the linguistic viewpoint, only some are statistically significant factors useful in duration modelling. This is because not all factors make a contribution to improving the "naturalness" of TTS systems.

Chapter 4 describes the design of the training corpus, used for investigating linguistic aspects of the duration pattern and modelling the duration of Korean. This chapter describes the methodology of collecting the speech corpus, the database processing and

the database annotation. In database processing, the process of converting raw text to phonetic transcription is described. The phonological rules for constructing a pronunciation dictionary are also investigated. For database annotation, automatic forced phone alignment and manual checking are used. The principles to decide the segment boundaries and phrase boundaries are described. Chapter 4 also shows how the transcriptions are automatically processed using the ProXML scripting language (Huckvale, 1999) into feature strings. The output of this procedure is a database of durations which are input to statistical processing for analysis of the duration pattern.

Chapter 5 describes the fitting of duration models to the corpus. Some of the statistical duration modelling approaches described in chapter 3 are applied to the processed database of Korean language. In particular, a sums-of-products model and a CART approach are evaluated. The objective quality of the modelling is evaluated by root mean squared prediction error (RMSE) and the correlation coefficient between actual and predicted durations in reserved test data. The outcome of modelling is new insights into the linguistic characteristics of the timing of the Korean language. This chapter investigates how the linguistic features and prosodic hierarchy interact to affect the duration value in different contexts.

In chapter 6, the subjective evaluation of two duration models is performed with a new Korean language speech synthesis diphone database, "HANMAL (HN 1.0)". Informal perceptual tests shows a listener preference for durations calculated from models in this thesis compared to durations calculated from a commercial Korean TTS system.

Chapter 7 reviews the contributions this thesis has made to the study of the timing of spoken Korean and makes suggestions for further work.

NOTES

[1] It was revealed in a daily Korean newspaper article (The Internet Hankyoreh, January 25, 2001) that the quality of speech recognition and speech synthesis systems for the Korean language is behind the standard of such technology in the major world languages. Microsoft Korea decided not to add speech synthesis and recognition into the new MS Office package, *Office 10*, while versions for other languages, such as major western languages, Chinese and Japanese include them.

2. PHONETIC/PHONOLOGICAL INFLUENCES IN TIMING

2.1 Analysis of Timing in English

This chapter discusses previous phonetic and phonological analyses of timing in spoken language. Phoneticians and linguists have long been interested in describing the inherent durations of segments and the effects of context on those durations using either a phonological or an experimental approach. One of the first studies of this kind is Stetson's (1951) investigation on the interaction between the duration of the final foot of a (nonsense) word and its position in the breath group. He found that the final foot of a breath group, whether it is trochaic or iambic, tends to be lengthened in a four syllable nonsense word.

Since then, many researchers have continued the investigation of the interaction between context and segment duration. Most studies have paid attention to the effects of prosodic factors such as the number of syllables in a word; the location of stress; accent; emphasis; boundaries between words, between phrases, between clauses, and between sentences; and the influence of word importance and meaning/content (Pickett, 1980). Segment duration in English has been one of the most popular areas for research. Those features which have been used in exploring the duration of English segments can be summarised as follows:

(2.1)

a. inherent duration of segment: manner, voicing and place features of the segment
b. features of surrounding segments
c. structure of the containing syllable: open/closed, position in onset/coda
d. stress status of the syllable: stressed, unstressed, accented
e. phrase boundary effect: sentence, clause, phrase, word, pause
f. frequency of the word

g. function word/content word distinction
h. speech mode: tempo, style (dialogue, reading)
i. size of the phrase in syllable numbers or in segment numbers

Lehiste (1970) suggested that the duration of sounds may be conditioned by the following factors: place and manner of articulation of the segment itself; identity of preceding and following segmental sounds; suprasegmental factors (especially by stress); and position of the sound within a higher-level phonological unit. Umeda (1975) suggested the following factors for duration variation of vowels in American English: positional conditions, consonantal conditions, word prominence, function word/content word distinction, and stress. Umeda investigated three positional conditions: prepausal condition, monosyllabic condition, and polysyllabic condition. Klatt (1976) suggested the following phonetic factors: inherent phonological duration, stress, influence of a postvocalic consonant on vowel duration, and consonant clusters. Klatt also suggested that such syntactic factors as clause boundary, phrase boundary, and word boundary have an effect. Umeda (1977) claimed that consonant durations are a function of the following factors: (1) position of the consonant in the word, (2) its relation to lexical stress and morpheme boundary (if any) within the word, (3) whether it is in a postpausal position, (4) whether it is in a prepausal position, (5) content-function status of the word, and (6) effect of adjacent consonants both inside the word and across word boundaries. Umeda (1977) also considered that certain factors which had been claimed to be important by other researchers were not important enough according to her data. They were: (1) number of syllables in the word, (2) phrase-final position without a silence following, and (3) identity of the vowel preceding or following a consonant. We will review the influence of these factors in the following sections.

2.1.1 Inherent duration

In English, vowels are longer than nonvocalic sonorants such as liquids, nasals and glides, and the sonorants tend to be longer than obstruents (fricatives, affricates, and stops). According to Crystal and House (1988a), Lehiste (1979; quoted by Crystal and House, 1988a), and Parmenter and Treviño (1935; also quoted by Crystal and House, 1988a), in the obstruent class, the affricates are longest and the fricatives and stops have about the same length. Crystal and House also found that voiced segments are longer than matching voiceless ones. However, Lehiste (1970) argued that though fricatives are considered longer than plosives, it is not always the case when they are embedded within a sentence rather than in an isolated word. Lehiste said that the inherent duration is determined by the place and manner of articulation. Other factors being equal, labials are longer than alveolar and velar consonants, probably because lips move slower than tongue. Umeda (1977) found that in word-initial stressed condition, labials are the longest, followed by dentals and velars. For taps, flaps, and trills, Lehiste did not find any generalisation about the influence of the manner of articulation on their duration.

2.1.2 Contextual effect of surrounding segments

The majority of the analysis of the contextual effect of surrounding segments in English has been focussed on the effect of consonants following a vowel. It has been considered that the influence of the initial consonants upon the durations of the following vowel is negligible (Peterson and Lehiste, 1960). However, Fischer-Jørgensen (1964; quoted by Crystal and House, 1988a) argued that vowels following voiceless stops are shorter than vowels following voiced stops. Lehiste (1970) suggested that the duration of a vowel depends on the extent of the movement of the speech organs required in order to come

from the vowel position to the position of the following consonant. The greater the extent of the movement, the longer the vowel.

In respect to the effect of the voicing of following consonants, vowels are generally shorter when followed by voiceless consonants, and longer when followed by voiced consonants (Peterson and Lehiste, 1960; Klatt, 1976; Crystal and House 1988b). Klatt explained that this is due to the requirement for earlier glottal opening for the following voiceless segment. Lehiste (1970) and Halle and Stevens (1971) argued that the wide separation of the vocal folds during voiceless consonants can be achieved more rapidly than the more finely-adjusted smaller separation for a voiced consonant. However, Crystal and House (1988b) claimed that short vowels preceding obstruents have equal length whether the obstruents are voiced or not unless the vowel is stressed or in pre-pausal position. That is, vowels preceding voiceless obstruents are shorter than voiced ones only when vowels are stressed or in prepausal position. This is a good example of an interaction between contexts which is a particular problem in durational modelling.

In terms of the effect of the manner of following consonants, vowels are shortest when followed by plosives; and nasals have approximately the same influence as voiced plosives. Vowels are longest before voiced fricatives (Peterson and Lehiste, 1960). Lehiste (1970), House and Fairbanks (1953) and Umeda (1975) proposed a rank ordering of the relative influence of following consonants on the preceding vowels. They said that vowels preceding voiceless stops are shortest, followed by those vowels preceding voiceless fricatives, nasals, voiced stops and voiced fricatives in order. Lehiste (1970) pointed out that the shorter duration of vowels before nasals than before homorganic voiced plosives—different from Peterson and Lehiste (1960)—is due to the

special adjustment of the vocal folds which is needed to maintain vibrations during voiced plosives, though not all accents of English do this. Lehiste said that the influence of the manner of articulation of a consonant upon the duration of a preceding vowel seems to be largely dependent on the language, which could be interpreted to mean that it is a phonological process.

Any place-of-articulation effect has been considered a secondary influence on vowel duration (House 1961; Crystal and House, 1988a). There is some controversy over the effect of the place feature of following consonants. According to Lehiste (1970, 1976) and Maack (1953; quoted by Crystal and House, 1988a), vowels preceding labial consonants are shorter than those vowels preceding other consonants, while Luce and Charles-Luce (1985) and Crystal and House (1988a) found that vowels before bilabial consonants are longer than vowels before velar or alveolar consonants. Lehiste (1970) suggested that the vowel duration before a labial consonant is shortest, since two different articulators are involved in the sequence vowel + labial and there is no time delay in moving the tongue from vowel target to consonant target. On the other hand, /u/ was particularly long before /d/. Before /g/, /u/ had an intermediate value; the movement involved is relatively small, but the back of the tongue is not as mobile as the tip of the tongue and the closing process takes more time. Lehiste (1976) also argued that "vowel duration tends to increase as the point of articulation of the postvocalic consonants shifts farther back in the mouth." Maack (1953) claimed that the vowel in vowel + velar cluster is longest, the vowel in vowel + labial cluster shortest, and the vowel in vowel + dental cluster has a medium duration. On the other hand, Fischer-Jørgensen (1964) said that back vowels are longer than front vowels before labials and dentals, while back vowels are shorter than front vowels before velars. However, in

Crystal and House (1988a), the ordering by vowel length is the reverse of Maack (1953) and Lehiste (1976).

2.1.3 Prosodic effects

Each word has a primary stress derived by English stress rules (Chomsky and Halle, 1968; Liberman and Prince, 1977; Hogg and McCully, 1987). In this thesis, stress means the primary stress unless mentioned otherwise. For the same speech rate, vowels which have a primary stress are longer than unstressed vowels (Lehiste, 1970; Klatt, 1976). In this case, stressed vowels are approximately 50% longer than unstressed vowels (Lehiste, 1970). When the syllable is stressed, the lengthening effect is not as great for the consonants as for the vowels (Crystal and House, 1988b; Klatt, 1976); and the lengthening of stops is greater than other consonants (Crystal and House, 1988b). Intervocalic /t/ and /d/ following the stressed vowel are shortened (Umeda, 1977). Consonants in a word-initial stressed position are longer than those in other conditions except in postpausal position. In postpausal position, consonants whether or not they are in a stressed syllable become considerably shorter, sonorants are particularly affected and become much shorter than other consonants. Crystal and House (1988b) argued that a word-initial [s] preceding a stressed vowel is longer than a word-initial [s] in an unstressed syllable, and that prepausal [s] is as long as a word-initial [s] preceding a stressed vowel.

With respect to the relations between phrase boundaries and segment length, the sentence final boundary has a clear lengthening effect. Klatt (1976) suggested the following syntactic factors of American English that influence the durational structure of a sentence: lengthening at clause and phrase boundaries and lengthening at word

boundaries. Prepausal syllable lengthening is observed in his data. He said that a syllable or syllables in utterance final position are lengthened, with the lengthening usually applied to the vowel, and any postvocalic sonorant or fricative consonants.

Umeda (1975) and Crystal and House (1988b) suggested that the presence of a pause is an important factor in the effect of a boundary. According to them, prepausal syllables are longer than nonprepausal syllables. Vowels in prepausal conditions are significantly longer than those in nonprepausal conditions whether they are stressed or not. Consonants in prepausal positions have a greater effect on the durations of preceding vowels than those in other positions. Consonants are longest when they are in prepausal position, while postpausal consonants are shorter than other consonants, and among them, postpausal sonorants are shortest (Umeda, 1977).

In non-phrase final position, some (Umeda, 1975; Klatt, 1976) consider word-final syllables to be slightly lengthened. Word-initial consonants are longer than word-final consonants unless the word is in phrase-final position. Otherwise, word-final consonants in phrase-final position are the longest, whereas word-medial consonants are the shortest. Plosives are not much affected by phrase-final lengthening (Klatt, 1976). Word-initial stops are longer than word-final stops (Crystal and House, 1988b). On the other hand, in investigations on unstressed vowels in polysyllabic words such as schwas or reduced vowels, Umeda (1975) found that vowels in word-final positions are longer than those in non-word-final positions and more affected by following consonants. The durations of unstressed vowels in non-word-final conditions behave similarly.

2.1.4 Word frequency and function word/content word distinction

Umeda (1975) argued that since frequent words are more predictable, so the durations of their vowels are reduced compared to the same vowels in less frequent words. Vowels in function words were found to be shorter than those in content words. Consonants in content words were found to be longer than those in function words except in word-final condition.

2.1.5 Tempo

There are two issues in terms of tempo and the segment durations: (1) overall rate which changes all durations proportionately, (2) differential effect of rate on segments of different type. In her experiment on the relationship between tempo and the syllable duration, Lehiste (1970) found that in English when the tempo of the utterance becomes faster, the unstressed syllables shorten more than the stressed syllables. But this is not always the case in other languages. She argued that in some other languages, the extent of shortening might be proportional over the whole utterance.

2.1.6 Segment and syllable numbers

Lehiste (1970) suggested that in some languages, there is a tendency for a word to maintain a constant duration, so the segmental length tends to decrease as the number of segments in the word increases. On the other hand, Crystal and House's (1990) investigation on the duration of syllables and stress groups in connected speech of American English revealed that the average duration of a stress group has a quasilinear dependency on the number of syllables and the number of phones, while the average duration of a syllable has a quasilinear dependency on the number of phones in the syllable. Furthermore, the linear factors were functions of stress.

2.1.7 Non-linear analysis

Ogden and Local (1992, 1996), Local and Ogden (1997), Ogden, Local, and Carter (1999) suggested that the durational variation within an utterance is not only determined by the segmental level, but also by its position within a prosodic hierarchy. In their speech synthesis systems, YorkTalk and ProSynth, timing changes are made at every level of phonological representation. Each non-linear phonological component takes part in the phonetic interpretation process, that includes timing, not just the lowest level segmental units. Timing changes are implemented by processes of syllable overlay and temporal compression. Syllable overlay is used to implement consonant ambisyllabicity, while temporal compression modifies syllable duration at different places in the metrical structure When a whole syllable is compressed, plosives and affricates are changed less than other constituents. This timing analysis tries to take into account all levels of linguistic structure.

2.1.8 Conclusions

The preceding summary of research into the timing of English highlights the range of problems facing any study of timing. Firstly, there are very many factors affecting timing, not just the intrinsic properties of a segment but also its context at many levels. Furthermore those factors can interact in complex ways so that it is difficult to study each independently. Secondly we see authorities in disagreement, such as the different opinions of Peterson and Lehiste (1960) and Fischer-Jørgensen (1964) on the effect of following segment. Thirdly, many researchers have failed to perform adequate empirical analysis of their hypotheses, such that they remain useless for synthesis. Fourthly, some researchers have used artificial materials, so that it is hard to determine the general applicability of their results. Finally, researchers have not managed to produce a

comprehensive account that combines all factors and interactions, although perhaps YorkTalk comes closest. This is perhaps due to the lack of an overall procedural framework for phonetic interpretation, such as the one developed in YorkTalk.

2.2 Analysis of Timing in Korean

In Korean, much of the previous research on duration has been concentrated on the analysis of the phonological vowel length contrast. This is described in section 2.2.3. In English, which has a lexical stress for each word and a pitch accent in each phrase, the interactions between segment duration and stress/pitch accent have been the main focus of linguistic investigation. However, in the Korean language, which only has a phrasal accent, most discussion on Korean timing has been focussed on the effect of phrasal boundaries and the effect of surrounding segments. Though there are some mismatches in the use of terminology, many studies agree that the Intonational Phrase (IP) and the Accentual Phrase (AP) are important phrasal units for describing rhythm and intonation in Korean (Koo, 1986; Lee, 1990; Jun, 1993; Chung et al., 1997). On the other hand, Lee (1996a) described the rhythm pattern using only the AP boundary and the Phonological Word (PW) boundary; while Han (1964) explored word duration using only the Utterance (UTT) boundary. The details of the phrasal hierarchy in Korean is described in 2.2.2 after the introduction of general phonological pattern in Korean in section 2.2.1. We show that boundaries can be defined on the basis of the pitch pattern and the occurrence of pauses at the phonetic level. Most scholars agree that Korean has a phrase accent and it has a phrase boundary tone at the end of an AP. However, there is some controversy over whether Korean has a pitch accent. Though the prosodic phrase is generally agreed to be the domain of phonological rules, there is not clear agreement about which phrase is the domain of which rules. Though many agree that boundaries

have a significant effect on duration, there are different views on which boundaries have the greatest effect or whether some boundary effects are statistically irrelevant. The relative importance of boundary-final and boundary-initial contexts is also an area of debate, while a final issue is which segments are affected by the boundary. A discussion of all these boundary effects is presented in section 2.2.4. There is much more agreement with the effects of surrounding segments in Korean. The durational variation between open syllables and closed syllables is well known. Studies on the effect of surrounding segments in English argue that post-vocalic consonants affect the vowel nucleus more than pre-vocalic consonants. However, in the Korean language, most studies argue that pre-vocalic consonants have the greatest effect. Section 2.2.5 investigates these segmental effects.

2.2.1 Phonological pattern of Korean

2.2.1.1 Consonants in Korean

Though there are some arguments about the Korean phonological pattern and its phonetic symbolisation, this thesis assumes that Korean language has the following consonants:

Table 2-1.
Consonants in Korean.

	Bilabial	Alveolar	Postalveolar	Velar	Glottal
Plosive	p^h p p'	t^h t t'		k^h k k'	
Nasal	m	n		ŋ	
Affricate			ts^h ts ts'		
Fricative		s s'			h
Liquid		l			

However, it is not possible to find all of the minimal pairs which could explain this phonological pattern. Some of the available minimal pairs in each phonological category are described below.

Plosives are seen in the following minimal pairs:

(2.2)

a. /p͡hal/ "arm" : /p͡al/ "feet"
   /palɯta/ "honest" : /p͡'alɯta/ "fast"
   /p͡hul/ "grass" : /p͡'ul/ "a horn"

b. /t͡hal/ "mask" : /t͡al/ "moon" : /t͡'al/ "daughter"

c. /k͡hi/ "height" : /k͡i/ "spirit" : /k͡'i/ "meals"

Korean plosives and affricates have a three way distinction: aspirated, tense, and lax. Aspirated plosives have strong aspiration when the closure releases. Lax plosives still have some aspiration, but the length of it is rather short, while tense plosives usually do not have aspiration. The feature [spread glottis] could be used as a distinctive feature for aspirated plosives, [constricted glottis] for tense plosives, and [stiff vocal folds] for aspirated and tense plosives. So aspirated plosives share [+spread glottis, -constricted glottis, +stiff vocal folds] features, lax plosives share [+spread glottis, -constricted glottis, -stiff vocal folds] features, and tense plosives share [-spread glottis, +constricted glottis, +stiff vocal folds]. In Korean plosives, all plosives share the feature [-slack vocal folds]. The combination of [+stiff vocal folds] and [-slack vocal folds] prevents the vocal folds from vibrating, so these plosives are voiceless (Halle and Stevens, 1971). In other words, this is the reason why the lax plosives which have [-stiff vocal folds] undergo voicing assimilation in many cases.

Nasals are seen in the following minimal pairs:

(2.3)

/tsam/ "sleeping" : /tsan/ "a glass" : /tsaŋ/ "chapter"

In Korean, the velar nasal /ŋ/ only appears on the coda position in the syllable structure.

Affricates are seen in the following minimal pairs:

(2.4)

/ts^{h}ata/ "cold" : /tsata/ "to sleep" : /ts'ata/ "salty"

Fricatives are seen in the following minimal pairs:

(2.5)

a. /si/ "city" : /s'i/ "a seed"
b. /s'anɯl/ "cool" : /hanɯl/ "sky"
c. /jʌŋsa/ "a consul" : /jʌŋha/ "below zero celsius"

There are two way distinctions in Korean fricatives: tense and lax. Tense fricatives usually do not have aspiration. Lax fricatives have frication and aspiration. Tense fricatives share [-spread glottis, +constricted glottis] and lax fricatives share [+spread glottis, -constricted glottis] features.

The lateral /l/ can be seen in minimal pairs with an alveolar stop or an alveolar nasal.

(2.6)

a. /kil̪o/ "turning point" : /kit̪o/ "prayer"
b. /tal̪/ "moon" : /tan̪/ "column"

Typically the consonant phonemes listed above could be realised using the allophones shown in Table 2-2. Those allophones which appear optionally are not included in this table.

Table 2-2.
Consonant allophones in Korean.

	Bilabial	Alveolar	Postalveolar	Velar	Glottal
Plosive	p^h p p' b	t^h t t' d		k^h k k' g	
Nasal	m	n ɲ		ŋ	
Affricate			ts^h ts ts' dz		
Fricative		s ɕ s' ɕ'			h ɦ
Liquid		l ɾ ʎ			

The lax plosives /p/, /t/, /k/ and lax affricate /ts/ are voiced between voiced sounds.

(2.7)

a. /napi/ "butterfly" [nab̲i]
b. /p^hato/ "wave" [p^had̲o]
c. /toŋkul/ "cave" [toŋg̲ul]
d. /itse/ "now" [id̲ze]

The alveolar fricatives /s/ and /s'/ are palatalised before a front close vowel /i/ or any diphthongs starting with /j/.

(2.8)

a. /sikan/ "time" [ɕigan]
b. /s'ial/ "an egg for breeding" [ɕ'ial]

The alveolar nasal /n/ becomes palatalised before a front close vowel /i/ or any diphthongs starting with /j/.

(2.9)

a. /jenilkop/ "six or seven" [jeɲilgop]

The alveolar lateral /l/ becomes a flap [ɾ] intervocalically, and becomes a palatal approximant [ʎ] when it is preceded by another /l/ and followed by front close vowel /i/ or any diphthongs starting with /j/.

(2.10)

a. /soli/ "sound" [soɾi]
b. /tallita/ "to run" [talʎida]

2.2.1.2 Vowels in Korean

This thesis assumes the following Korean monophthongs.

Table 2-3.
Monophthongs in Korean.

	Front	Central	Back	
			Unround	Round
Close	i		ɯ	u
Close-mid	e			o
Open-mid	ɛ		ʌ	
Open		a		

In a modern standard Korean, all front vowels are phonologically unrounded. Older users of Seoul dialect still have front rounded vowels /y/ and /ø/ in their phonological pattern. However, general modern standard Korean users replace them with diphthongs /wi/ and /we/ respectively. In the back vowels, the one to the right represents a rounded vowel. Some researchers replace the two unrounded back vowels /ɯ/ and /ʌ/ with central vowels /ɨ/ and /ə/. Traditional scholars such as Lee (1955), Lee (1956), Choi (1983), Han (1964), Kim (1974), and Heo (1985) argued that there are vowel length differences which are phonemic in Korean. Although Koo (1986) and Jun (1998) said that the contrast of phonemic vowel length does not exist in the modern Seoul dialect which is considered the basis of modern standard Korean. This thesis follows the latter claim, because the spoken data used in the experiments show that the speaker does not make a phonemic difference in his accent. So it is assumed that there is no need to mark vowel length in our phonemic transcriptions of Korean.

Korean also has the following rising diphthongs.

(2.11)

a. /ja/, /jʌ/, /jo/, /ju/, /je/, /jɛ/
b. /wa/, /wʌ/, /we/, /wi/, /wɛ/
c. /ɯi/

(2.12) provides an illustration of some of these vowel contrasts.

(2.12)

a. /o̲ta/ "to come" : /i̲ta/ "to put on the head"
b. /kɯ̲m/ "a line" : /ki̲m/ "Kim (one of Korean family names)"
c. /nu̲n/ "eyes" : /no̲n/ "paddy"
d. /na̲/ "myself" : /ne̲/ "Yes."
e. /tɛ̲/ "a bamboo" : /to̲/ "province"
f. /wa̲ntsʌn/ "perfect" : /wʌ̲ntsʌn/ "the original text"
g. /ju̲ki/ "brassware" : /jʌ̲ki/ "here"
h. /we̲saŋ/ "credit" : /ju̲saŋ/ "oily"
i. /ɯi̲sa/ "a doctor" : /i̲sa/ "house removal"

2.2.2 Prosodic structure of Korean

In this thesis, both the contextual effects of preceding/following segments and the effects of prosodic phrase boundaries are investigated. In order to clarify any interactions between segmental contexts and phrasal contexts, the prosodic boundaries need to be determined purely on the basis of syntactic structure in combination with pragmatic information. There are three approaches to determining the prosodic structure larger than a syllable: the end-based approach (Selkirk 1986), the relation-based approach (Nespor and Vogel, 1986; Hayes, 1989), and the intonational approach (Beckman and Pierrehumbert, 1986; Jun, 1993) In the end-based approach, the prosodic structure larger than a syllable is determined from the syntactic structure by referring to the edge of a maximal projection. In the relation-based approach, boundaries above the foot and below the Intonational Phrase are posited with reference to the head-complement relation or to the c-command relation. Above the syllable, the phonological structure is

assumed to be hierarchically organised observing the Strict Layer Hypothesis (Selkirk, 1984, 1986; Nespor & Vogel, 1986) which means that any given non-terminal unit of the prosodic hierarchy is composed from one or more units at the immediately lower category, and a unit of a given level of the hierarchy is wholly contained within the superordinate unit of which it is a part. For English, Selkirk (1984) suggested the intonational phrase (IP), the phonological phrase (PhP), and the prosodic word (Wd) as prosodic units, while the Foot (Ft) and the Syllable (Syl) need a separate model of the relationship between the syntax and phonology. Each prosodic constituent is defined by the left or right edge of a lexical or syntactic category. In Selkirk and Shen (1990), this claim is formulated as:

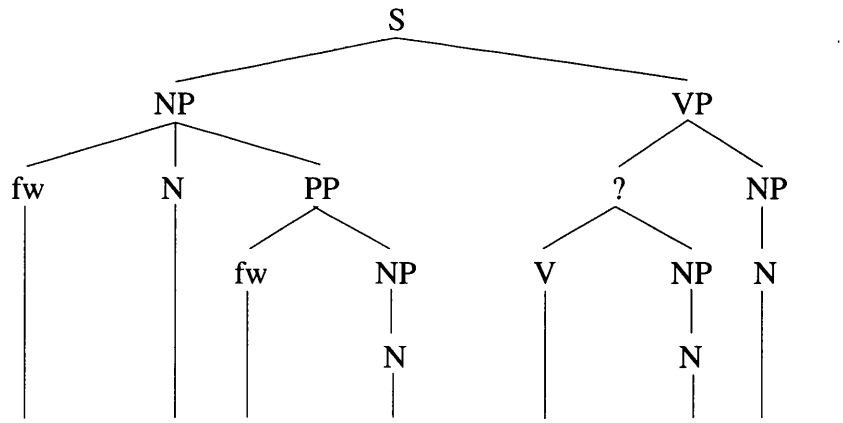
(2.13)

> The Syntax-Phonology Mapping
>
> For each category C^n of the prosodic structure of a language there is a two-part parameter of the form
>
> C^n: {Right/Left; X^m}
>
> where X^m is a category type in the X-bar hierarchy.
>
> A syntactic structure-prosodic structure pair satisfies the set of syntax-phonology parameters for a language if and only the Right (or Left) end of each constituent of type X^m in the syntactic structure coincides with the edge of constituent(s) of type C^n in the prosodic structure.

According to this principle, the right edge of the lexical word coincides with the right edge of the prosodic word, the right edge of the phrasal category coincides with the right edge of the phonological phrase, and the right edge of the clause category coincides with the right edge of the intonational phrase. This structure could be schematised as follows:

(2.14)

a.



b. .....]ω .....]ω .....]ω .....]ω .....]ω
.....]Xmax .....]Xmax ...]Xmax

c. ( ) ( ) ( ) ( ) ( ) PWd
( ) ( ) ( ) PPh

In this figure, 'fw' is "function word", which are not treated as "real" words but are included in the next domain. Thus the right edge of the word is the boundary of a prosodic word and the right edge of an XP is the boundary of a phonological phrase in English. The choice of right end or left end is a language-specific parameter.

Nespor & Vogel (1986) proposed seven units for the prosodic hierarchy: the phonological utterance (*U*), the intonational phrase (*I*), the phonological phrase (ϕ), the clitic group (*C*), the phonological word (ω), the foot (Σ) and the syllable (σ). Each category has its own formation rules. Though these rules indirectly refer to (morpho-)syntactic notions in their definitions, the phonological hierarchy does not necessarily correspond to the syntactic structure. Nespor and Vogel's phonological hierarchy is defined in terms of mapping rules representing the interface between phonology and other components of the grammar, so in some languages a phonological category X^i

might not exist because it does not interact with phonological rules. They said that levels of the prosodic hierarchy are the domain of the phonological rules. Nespor & Vogel said that the terminal category of the prosodic hierarchy is the syllable. They exclude segments, onsets, and rhymes from the prosodic hierarchy, because they are not organised in accordance with the principles governing all the other units above the syllable level, and do not serve as the domain of application of phonological rules. They are not claiming that onset and rhyme constituents have no role in phonology, but rather that they cannot be considered constituents of the prosodic hierarchy. Due to the violation of the Strict Layer Hypothesis caused by ambisyllabicity and feature sharing[1], they assume that segments, or at least their positions, are not the constituents of the prosodic hierarchy, but rather the central core of the phonological representation operating as the common intersection of all the subsystems.

As one of the intonational approaches, Jun (1993) proposed that Korean uses the Intonational Phrase (IP) and the Accentual Phrase (AP) as the units for intonation and rhythm. She suggested that the IP corresponds to that as in prosodic phonology, but that the AP is based on the tonal pattern. This AP is of the same level as the phonological phrase of prosodic phonologists, but is different in that the boundary is determined by the tonal pattern. She (1993, 212) argued "At least in Korean, an AP cannot include a prosodic word 'ω' to the preceding word(s), if 'ω' is the last prosodic word of the AP and the left element of the syntactic branching constituent." She suggested the following Accentual Phrasing rules.

(2.15)

a. Every prosodic word may be an AP.

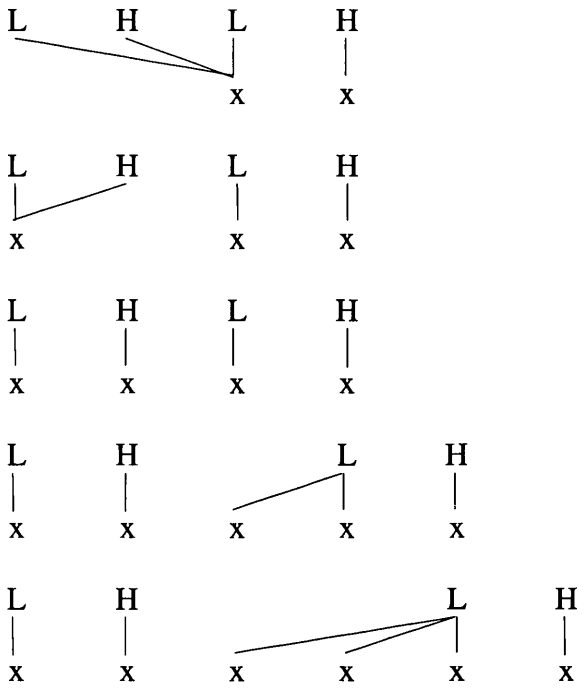
b. A focussed word must be the left-most word in an AP.
c. An AP can include any number of prosodic words as long as:
    i. the last prosodic word is not the left element of a branching constituent
    ii. all the prosodic words are not focused

In terms of the tonal pattern of the AP in Seoul Korean, Jun said that it is realised as a rising contour (=LH) when the phrase is short, but realised as a rise-fall-rise contour (=LHLH) when the phrase is longer than three or four syllables. However in fast speech, the first H tone may not be realised, even though the phrase has more than five syllables. In a long phrase, the first H is lower than the final H tone. Following De Jong (1994), she said that this first H tone is realised near the offset of the first syllable of the AP. Jun assumes that this first H tone is part of the underlying tone pattern (LHLH) of the AP in Seoul, but is not always realised on the surface due to phonetic undershoot.

Jun proposed that the final H is associated with the last syllable of the Accentual Phrase and that the penultimate syllable has a L tone. These L tones are realised in a following pattern down to a low "target" value, after which they remain flat. Jun said that because the last H tone is very salient in Seoul Korean, it has a function of demarcating the AP boundary. Jun argued that since the initial H tone is not so typical in Seoul Korean, the most significant factor to demarcate the AP is the phrase final rising tone, LH.

Her statement could be interpreted as follows:

(2.16)



This interpretation shows that the underlying tonal melody of a Korean AP is LHLH and mapped in a right-to-left procedure. The final LH tones are always associated with the last two syllables in the Accentual Phrase, respectively. The penultimate tone, L is left spreading until it reaches a syllable that is already associated. Each association observes the Well-formedness Condition (Goldsmith, 1976), that no tone and syllable is allowed to be stranded without being linked and that association lines do not cross. So the first L is linked to the first syllable, otherwise it is (left) spreading and linked to the first available syllable observing Well-formedness Condition. The second tone, H is linked to the second syllable, otherwise it is (left) spreading and linked to the first available syllable without violating the Well-formedness Condition.

Jun (1993) said that an IP is constituted of one or more AP tonal patterns with an IP boundary tone at the end of it. IP boundaries may be followed by a pause. When the AP

boundary tone overlaps with an IP boundary tone, either a small initial hump or a large initial rise precedes the IP boundary tone. At the same time, the last rising tone of the AP is replaced by the tonal pattern of the IP.

In contrast, Lee (1990) argued that the Korean Intonation system consists of rhythm units and intonation groups, which in turn define the utterance prosody. In his argument, the 'tune' of the intonation group consists of one or more 'phrasal tones'[2] of rhythm units and one 'boundary tone' on the final syllable of the intonation group. Lee (1996a) said that the rhythm unit is accompanied by a pause and breaks the flow of the rhythm. The phrasal tone can have a level, falling, rising, or rise-falling tone. The boundary tone can be a low level, mid level, high level, high fall, low fall, full rise, low rise, fall-rise, or rise-fall tone.

When the proposals of Lee (1990) and Jun (1993) are compared, the rhythm unit is somewhat similar to the AP and the intonation group is somewhat similar to the IP. Indeed, Lee (1996a) agreed with this. The difference lies in the inventory of tones.

In any case, it seems that the final LH tone is the prominent factor in the AP or rhythm unit boundary. The boundary tones used in Lee's (1990) and Jun's (1993) models have many types. Jun suggested that the IP can have boundary tones such as L, H, LH, HL, LHL, and HLH.

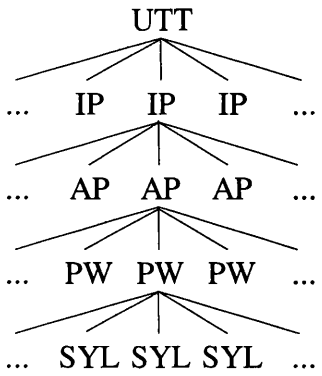
Koo (1986) argued that sentences can be decomposed into intonational phrases by (1) the insertion of pauses and (2) by tonal changes; but that pauses are not obligatory at the end of intonational phrases. He said that intonational phrasing is determined by syntactic

structure and focus structure and the tonal pattern before the intonational phrase boundary can be: (1) "rise-large fall" followed by pause, (2) "large rise" followed by pause, (3) "large rise-large fall" followed by pause, and (4) "large rise-large fall" not followed by pause.

Conceptually, these models of prosodic structure are clear and easy to follow. When the syllable is final to the AP and the at the same time is final to the IP, the final rising tone of the AP is replaced by the tones of the higher level, the IP boundary tone. However, as Jun (1993) described, when the actual speech signal is analysed, unless the pause is considered a cue for the IP boundary, it is not an easy task to identify which is the AP boundary or the IP boundary, because in many cases the boundary tone accompanies the overlapping H tone at the end of the IP which shares the property of the last tone of the AP.

In this thesis, we will use "AP" to denote the accentual phrase or the rhythmic unit; and "IP" to denote the Intonational Group. We assume that the AP can be demarcated by a phrase final tonal pattern of LH, and that the IP can be demarcated by a clear pause whether or not it ends with any kind of boundary tone. Chung et al. (1993) argued that IP boundaries obligatorily accompany pauses. This principle seems to be reliable enough to annotate the boundaries of IPs. The complete prosodic hierarchy thus becomes: Utterance (UTT), Intonational Phrase (IP), Accentual Phrase (AP), and Phonological Word (PW). The Syllable (SYL) is the terminal node of this prosodic structure. In this thesis, UTT is always a whole sentence. This hierarchical structure can be schematised as follows:

(2.17)

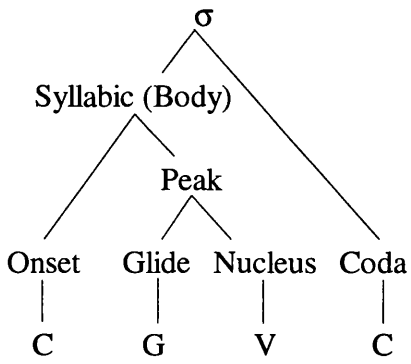


Because of the lack of a lexical stress system in standard Korean (Koo, 1986), the metrical foot is not used. In Korean, the phonological word is different from the morphological word. Each element of a compound word could be a PW and even a prefix could be a PW. A PW can contain one or more suffixes, case particles or enders. As mentioned above, each prosodic structure is the domain of phonological rules. Phonologically, the IP is the rule domain of n-lateralisation, aspiration, obstruent nasalisation, spirantisation, s-palatalisation, and resyllabification; AP is the rule domain of voicing assimilation, plosive nasalisation, nasal assimilation, and tone deletion; PW is the rule domain of t-palatalisation, n-lateralisation, l-nasalisation, obstruent tensification, n-deletion, and n-insertion; SYL is the rule domain of consonant cluster simplification, aspiration, and neutralisation. n-lateralisation and aspiration have two different rule domains to solve the ordering paradox arising from l-nasalisation and consonant cluster simplification respectively. Discussions on the interaction between the phonological hierarchy and the phonological rules of Korean can be found in Chung et al. (1997), Lee (1996a), Jun (1993), Kwack (1992), and Oh (1989),

Because this thesis investigates segmental as well as prosodic effects on segmental duration, we also need to consider the internal structure of the syllable. There has been debate whether Korean syllable structure is flat structure or has an onset-rhyme, or has a moraic structure. One of the interesting accounts of the Korean syllable structure is Moon's (1996).

Moon suggested that the Korean syllable has the following structure.

(2.18)



C stands for a consonant, V stands for a vowel, and G stands for a glide. He argued that the close relation between the glide and the nucleus compared to the onset and glide can be found in some speech errors, as follows.

(2.19)

a. /kolkjʌk/ "structure" → [kjʌlkok]
b. /kaŋp^{h}jak/ "stubborn" → [kjaŋp^{h}ak]
c. /kaisalja/ "Caesarea" → [kaisjara]

Also when two vowels merge into one vowel, they commonly undergo a glide formation rule. In this case, the first vowel must be /i/, /u/, or /o/ and the second vowel must be /ʌ/ or /a/.

(2.20)

a. /tsɯlkiʌla/ "Enjoy yourself." → [tsɯlkjʌɾa]
b. /kiʌkata/ "to crawl" → [kjʌkata]
c. /k'uʌ/ "to borrow" → [k'wʌ]
d. /muʌnja/ "What?" → [mwʌnja]
e. /s'oala/ "Shoot!" → [s'waɾa]
f. /poa/ "to see" → [pwa]
g. /iluʌtsiʌja/ "to be accomplished" → [iɾuʌtɕjʌja]

The constraint on the glide formation rule applying to two adjacent vowels could be evidence for the above syllabic structure in which syllable body branches into onset and peak and the peak branches into glide and nucleus.

Moon does not think that the syllable final coda constitutes a rhyme with a nucleus in Korean. He said that in Korean there are few phonological rules in which the rhyme functions as a unit. In many cases, onset, peak or coda independently play a role in the phonological rules; or the whole syllable becomes a unit; or onset and peak constitute a unit.

In Korean reduplication, some words contain reduplicated whole morphemes; some morphemes contain reduplicated last syllables; some contain reduplicated onset+peak. For example:

(2.21)

a. /huɯtɯl/ "swinging" → [huɯtɯlhuɯtɯl]
/mallaŋ/ "soft" → [mallaŋmallaŋ]
/pasak/ "crispy" → [pasakpasak]
/pintuŋ/ "lazily" → [pintuŋpintuŋ]

b. /ʌpʌŋ/ "ignorantly" → [ʌpʌŋpʌŋ]
/ʌlt'ʌl/ "confused" → [ʌlt'ʌlt'ʌl]
/ususu/ "in a multitude" → [usususu]
/salɯlɯ/ "gently" → [saɾɯɾɯɾɯ]

c. /tuŋsil/ "floating lightly" → [tutuŋsil]
/t'ekul/ "rolling over and over" → [t'et'ekul]
/tʌlk^{h}ʌŋ/ "to keep rattling" → [tʌtʌlk^{h}ʌŋ]
/tʌŋsil/ "dancing joyfully" → [tʌtʌŋsil]

d. /tɯlɯŋ/ "with a snore" → [tɯɾɯɾɯŋ]
/asak/ "crunching" →[asasak]
/tʌŋtʌŋ/ "a tum-tum" → [tʌŋtʌtʌŋ]

There is no reduplication rule in Korean which copies the peak+coda to make a reduplication word. This is evidence that in Korean onset and peak constitute a unit and that the coda is an adjunct to the syllable.

In terms of temporal structure, if a vowel is more affected by the preceding consonant than by the following consonant, then it could be assumed that onset and peak constitute a unit of the syllable body. If the opposite were true, the nucleus and coda could be taken to constitute a rhyme unit. Because the evidence is still inconclusive, we adopt the widely-used syllable structure form which is constituted of an onset and rhyme, the rhyme constituted of a nucleus and coda. But we allow our duration modelling to separate out the influences of pre-vocalic and post-vocalic consonants.

In summary, this section has justified a hierarchical prosodic structure of UTT, IP, AP, PW, SYL, ONSET, NUC, RHYME, and CODA for Korean. These structures are the main framework within which the durational pattern will be investigated in chapters 4 and 5.

2.2.3 Inherent duration

While general studies of English on vowel duration have been focused more on contextual effects, much of the work on the duration of vowels in Korean has been on the phonemic property of vowel length, which is a controversial issue in modern standard Korean. Vowel length in the Korean language has long been considered phonemic. Lee (1955) presents examples such as /kwaŋtsu/ "wide field" /kwa:ŋtsu/ "Kwangju (one of metropolitan cities in Korea)". Lee (1956) said that though vowel duration distinguishes the meaning of many words in Korean and the contrastive vowel duration is recognised by many Koreans, there is no written symbol to explain the distinction in Korean characters. Martin (1951) treated the Korean long vowel as a sequence of two short vowels and argued that vowel length is distinctive. However, Martin shows that for some speakers there are few contrasts of long and short vowels; while for others there are many. Many speakers do not use long vowels in words that have long vowels for some speakers. Even for a speaker making a maximum use of length distinctions, long vowels are usually restricted to the first few syllables of a word. Choi (1983) also suggested that vowel duration is a significant feature. Choi said "In Korean there are certain words whose meanings are distinguished by the difference between a long and short sound. However, these words are spelled in the same way in the writing forms." In Han (1964), based on her experimental auditory test result, she uses /:/ as a phonemic feature in Korean. According to her result, native speakers distinguish long and short

vowels on phonological cues along with more than 90% accuracy. She stated that vowels /i/, /a/, /ʌ/, /o/, /u/ commonly occur with extra duration: /i:/, /a:/, /ʌ:/, /o:/, and /u:/ respectively. Finding that the overall average of the contrast between short and long vowels was approximately 2.51, Han suggested that this may be regarded as the norm for the long-short contrast of Korean vowels. Han's data shows that because the inherent durations of /i/ and /u/ are much shorter than /a/ and /o/ and long /i:/ and /u:/ are almost as long as long /a:/ and /o:/, the ratios of /i:/ vs. /i/ (2.88) and /u:/ vs. /u/ (2.82) are much higher than those of /a:/ vs. /a/ (2.09) and /o:/ vs. /o/ (2.07).

For inherent vowel durations, Han (1964) found that the vowel [i] is always shorter than other vowels; [ɯ] and [u] are considerably shorter than [a], [ʌ] and [o]; [a] is usually longer than any other vowel except when it follows [s]. Han found that the long-short vowel contrast does not appear in the following contexts: after tense and aspirated consonants, before voiceless stops, in closed syllables, and in non-initial syllables. She implies that in these contexts (long) vowels tend to be short.

Kim (1974) argued that "in Korean, vowel length is phonemic even though the functional load of this feature is relatively light." Heo (1985) said that in Seoul dialect, which is generally considered to be a "standard Korean", the duration of vowels are part of speaker's linguistic intuitions, so it functions as a distinctive feature in determining lexical meanings. However, he also said that the younger generation of less than 60 years does not consider the chroneme (Jones, 1976) as a distinctive feature in a standard Seoul Korean, while the generation older than 60 years has the chroneme (=duration phoneme)[3]. Heo shows following examples:

(2.22)

a. /kam/ "going" /ka:m/ "a persimmon"
b. /kul/ "an oyster" /ku:l/ "a cave"
c. /kil/ "a road" /ki:l/ "long"
d. /tal/ "the moon" /ta:l/ "a reed"
e. /tol/ "an anniversary" /to:l/ "a stone"
f. /mal/ "a horse" /ma:l/ "a talk"
g. /mɛ/ "a whip" /mɛ:/ "a hawk"
h. /multa/ "to pay" /mu:lta/ "to spoil"
i. /pam/ "night" /pa:m/ "a chestnut"
j. /pal/ "a foot" /pa:l/ "a bamboo curtain"
k. /pɛ/ "a vessel" /pɛ:/ "double"
l. /pʌl/ "a set" /pʌ:l/ "a bee"
m. /pjʌl/ "to be classified by" /pjʌ:l/ "a star"
n. /son/ "hand" /so:n/ "descendants"
o. /sol/ "a pine" /so:l/ "a brush"
p. /sul/ "liquor" /su:l/ "a tassel"
q. /santa/ "to buy" /sa:nta/ "to live"
r. /tsul/ "a rope" /tsu:l/ "a file"

He also argued that many more examples of these distinctions can be found, especially in Sino-Korean words as in (2.24).

(2.23)

a. /kɯnkan/ "the basis" /kɯ:nkan/ "a recent publication"
b. /tansik/ "single match" /ta:nsik/ "a fast"
c. /ku:tso/ "rescue" /kutso/ "structure"
d. /pjʌ:ŋ/ "illness" /pjʌŋ/ "glasses"

In summary, Heo claimed that in Seoul and Kyonggi dialect, the length of vowels has a distinctive function, especially among the older generation. Heo also stated that long vowels are pronounced a bit higher in pitch so that pitch is considered a redundant feature not a distinctive feature in Korean.

Koo (1986) said that although many previous studies argue that the vowel length is phonemic, "it seems that a contrast of long and short vowels is no longer meaningful in distinguishing lexical words consisted of identical phoneme sequences in modern Korean." Jun (1998) argued that the phonemic long and short vowel contrast among the younger generation less than 50 years old is about to disappear, though some still distinguish them.

In this thesis, we follow the Koo's (1986) and Jun's (1998) analysis that the lexical contrast between long vowels and short vowels is neutralised in modern standard Korean. The speaker of our experimental data described in chapter 4 does not distinguish the traditional long vowels from short ones even in isolated words.

It is unfortunate that other than the phonemic contrast of vowel length, only a couple of articles are available about other matters of the inherent durational property of vowels. Among them, Koo (1998) found that the front, mid and low vowels tend to be longer than high and back vowels.

2.2.4 Prosodic effects

It is generally agreed that phonological units in sentence-final or IP-final position are longer than in any other positions (Han, 1964; Kim, 1974; Chung et al., 1997; Jun, 1993; Lee and Koo, 1997). Han (1964) measures word durations. Her results show that the duration of the sentence-final word is longest, usually shortened by only 10-20% from its citation form, while in sentence-initial position the duration is reduced by 30-40% on average. A word in sentence-medial position is shortened by 40%. Han said that consonants are more affected than vowels when the duration of a word is reduced. She

said that the temporal compression of the word is usually carried out by the great reduction of consonant duration, with the vowel being shortened moderately.

Kim (1974) found that vowels in UTT-final syllables and PW-initial syllables are longer than those in other positions, with vowels in UTT-final syllables longer than those in PW-initial syllables. Kim also argued that segments in phrase-initial syllables are longer than those in phrase-medial syllables.

Chung et al. (1997) found that there is a greater lengthening effect in IP-final position than other prosodic phrase boundaries. They said that an IP-final syllable is 60% longer than the average duration of all syllables and an IP-final vowel is 87% longer than the average duration of all vowels. Jun (1993) said that IP-final syllables are lengthened when followed by a pause.

Overall, there is some disagreement about which syllable/vowel in the AP or PW is most affected in duration. Chung et al. (1997) said that the lengthening effect in AP-final position is not significant. Jun (1993) argued that though the AP-initial segment is lengthened, the AP-final syllable is not noticeably lengthened. On the other hand, Lee (1990, 1996a) said that AP-final syllables are lengthened when followed by pause. In his discussion of the phonetic variation of vowel length, Lee argued that a vowel in an open syllable is longer in a rhythm unit (=AP) final position than in other positions, other things being equal.

Han (1965) analysed the duration of aspiration of Korean obstruents in word-initial position and word-medial position. Han found that after pause, the aspiration of three

obstruents /p^h/, /t^h/, and /k^h/ usually last more than 100 ms. However, in word-medial position, the duration of the aspiration becomes shorter and closer to the aspiration duration of word-initial lax stops. In terms of lax stops, /p/, /t/, /k/, the average duration of aspiration in word-initial position is approximately 40 ms. The aspiration of these also becomes shorter or even disappears in word-medial position. Between vowels there is even a tendency to become voiced. Based on her perceptual testing, Han argued that the difference of the aspiration duration between aspirated stops and lax stops should be between 80 and 60 ms, if they are to be perceived as different phonemes.

Lee (1996) argued that differently from other languages, Korean CVC/VC syllables are longest in phrase-initial position, followed by phrase-final position, then phrase-medial position. CV/V syllables are longest in phrase-final position, followed by phrase-initial position then phrase-medial position. Vowels in CV structures are longest in word-final syllables, while they become shorter in word-initial or word-medial syllables. Though vowel duration is affected by syllable structure, consonant duration is not affected. In word-initial syllables, /s/ is longest, followed by nasals, then aspirated, tensed and voiceless obstruents. In word-medial syllables, aspirated and tense obstruents have their average duration, while other consonants become shorter. Voiceless obstruents tend to be shorter than other consonants, because they are likely to be voiced between voiced sounds.

Kang (2000) measures the mean duration of Korean tense and lax fricatives, /s'/ and /s/ both word-initially and word-medially produced in isolation. She found that the lax fricative /s/ in word-initial position is 50% longer than that in word-medial position, while tense fricative /s'/ has a similar duration in word-initial or in word-medial position.

Using carrier-phrase sentences, Lee and Koo (1997) measure the duration of the final syllable before four boundaries: UTT, IP, AP, and PW. They use 117 sentences recorded by 6 different female speakers at three different speeds: fast, slow, and normal. They found that at any speed, the syllable before the UTT boundary is longest and is least influenced by speaking rate. At fast speed, the syllable before IP and PW boundaries has medium duration and is shortest before the AP boundary. At normal speeds, syllables before IP, AP, and PW boundaries have a similar duration. At slow speeds, syllables before a PW boundary are shortest and the duration of those before UTT, IP, and AP boundaries are similar. At slow speeds, the syllable before an UTT boundary is 11% longer than at normal speed; the syllable before an IP boundary is 32% longer; the syllable before an AP boundary is 28% longer; and the syllable before a PW boundary is 21% longer. At fast speeds, the syllable before an UTT boundary is 7% shorter than at normal speeds; the syllable before an IP boundary is 25% shorter; the syllable before an AP boundary is 35% shorter; and the syllable before a PW boundary is 18% shorter.

In her experiment on the vowel duration of /a/, Han (1964) found that vowels tend to be short in a closed syllable and in a non-initial syllable. The occurrence of phonemic /:/ is limited to closed syllables since in CV position, the phonemic vowel length is neutralised. Han observes that vowel [a] is longest when it is spoken in citation form between junctures without any preceding or following consonants. The average duration of [a] in isolation is 308 ms which is 2.4 times longer than the average duration of [a] in a CVC syllable, and 1.16 times longer than that of a CV syllable. When the vowel [a] is in a CV syllable, it is almost as long as [a] in isolation. She concludes that [a] in an open syllable

is approximately 2.1 times longer than that in closed syllables. She also said that a vowel is usually longer in a monosyllabic word than in a polysyllabic word.

Lee (1990) stated that vowels are longer in open syllables than in closed syllables, other things being equal.

(2.24)

a. [sa:ɾam] "human being" > [sa:lda] "to live"
b. [poda] "to see" > [ponnɯŋ] "instinct"

Koo (1998) said that when the vowel duration in V syllable structure is set as 100%, vowels in CV syllable structure are shortened to 79.3% and those in CVC syllable structure to 38.5%.

Lee (1996) said that the more syllables the phrase has, the shorter the segments are. The magnitude of the effect is greater in vowels than in consonants. Lee (1990) said that a vowel is longer in an AP with fewer syllables than in one with more syllables, other things being equal. He argued that this tendency is evidence for a stress-timing hypothesis concerning Korean rhythm. For example:

(2.25)

a. [sa:ɾam] "human being" > [sa:ɾamida] "be a human being"
b. [saɾang] "love" > [saɾangsɯɾʌpda] "be lovely"

This is closely related to the discussion of the nature of the timing unit in spoken Korean. Yun (1998) suggested that the duration of each linguistic unit (syllable, word, foot, and

sentence) is determined by the number and type of phonemes in Korean, in contradiction to the syllable-timed hypothesis (Martin, 1951), the stressed-time hypothesis (Lee, 1982), and the word-timed hypothesis (Kim, 1994). Yun argued that Korean is a phone-timed language. His argument implies that the timing is mainly determined by the properties of the segments in each linguistic unit, not by the stress or syllable numbers.

Han and Oh (1999) found that there is a speaker sex factor in determining the duration of IP-final syllables. The duration of IP-final syllables by male speakers is longer than that of UTT-final syllable; however female speakers tend to lengthen UTT-final syllables more than IP-final syllables.

Lee (1990) argued that the vowel and the coda in an accented syllable is longer than in the unaccented syllable.

(2.26)

a. ['si̲:tsaŋ sa:ɾam] > [si̲:tsaŋ 'sa:ɾam]

In terms of speaker variation, Han (1964) said that when seeking structural patterns or settings of a language, comparisons must first be made within one person's speech. The morphological or syntactic environment is another factor which affects vowel length. There are few works available on the duration of Korean consonants. Han and Ross (1968) investigated the duration pattern of Korean affricates. They measure the fricative portion of three affricates: /ts^{h}/, /ts/, and /ts'/ in word-initial position. They found that the fricative portion of aspirated affricate /ts^{h}/ is considerably longer than that of /ts/, and that the fricative portion of /ts/ is slightly longer than /ts'/.

In this section, previous work on the interactions between prosodic boundaries and segment duration have been described. Some said that an IP boundary lengthens the preceding syllable and others said that it is the AP boundary that lengthens the preceding syllable. Some argue that an IP-initial syllable is lengthened and others argue that it is AP-initial syllables. Some argue that the boundary effect is different depending on the syllable structure or the sex of the speaker. There is also debate on the relative changes in duration of vowels and consonants when syllables are lengthened. Some of these issues are addressed in the experimental work described in chapters 4 and 5. We investigate which phrase boundary type has more influence and which segments/syllables in a phrase are most affected by boundaries.

2.2.5 Contextual effects of surrounding segments

Han (1964) found that aspirated stops such as [t^h] and aspirated affricates such as [ts^h] shorten the following vowel considerably. After the lax stop [t] and affricate [ts] the vowel was near its average of 127 ms, while after the nasal [n] it was found to be much longer than the average. Han and Ross (1968) also found that the duration of a vowel after /ts'/ is lengthened.

However, different from many results in English, Han's experimental data did not show a simple pattern in the effect of the following consonant on the vowel in question.

Kim (1974) investigated the effect of consonants on the following vowel both in UTT-initial position and in UTT-medial position. In UTT-initial position, he found that the influence of the initial consonants on the duration of the following vowel is considerable.

He also found that when vowel is preceded by aspirated consonants such as /p^h/, and /s/, it is shortest. The vowel is longest after /p'/ and /m/, but less so after /p/ and /ts/. He said that as the aspiration portion of the preceding consonant becomes longer, the duration of the following vowel is shorter. He explained that in order to produce an aspirated sound, the vocal folds need to open widely, so this causes a delay in producing the following vowel. He suggested that the influence of any consonant on the duration of the following vowel is a result of the degree of glottal opening inherent in the articulation of the consonant.

In his experiment examining the effect of syllable initial consonants on the duration of following vowels in UTT-medial position, Kim (1974) found that the vowel is shortest after /p^h/ and the vowel after /s/ is shorter than the vowels after other consonants. However, the different effects of the two pairs /p, ts/ and /p', m/ are neutralised. He explained that because intervocalic weak plosives (lax stops) tend to be voiced in Korean, the effects of /p/, /t/, /k/, /ts/, /p'/, and /m/ become less clearly distinct from each other, caused by neutralisation of the difference in the extent of the glottal opening in the different cases.

Lee (1996) also found that preceding consonants considerably affect following vowel duration. Vowels are longest after nasals or liquids, followed by voiced plosives, then aspirated obstruents/tensed obstruents/fricatives/affricates.

Kang (2000) found that a vowel following a lax fricative /s/ is approximately 30% shorter than that following a tense fricative /s'/ in word-initial position when the word is pronounced in isolation, while the difference is neutralised in word-medial position. A

vowel before a tense fricative is 30% shorter than that before a lax fricative in word-medial position. On the other hand, a vowel following an aspirated alveolar stop /t^h/ is approximately 40% shorter than that following a lenis alveolar stops /t/ and 30% shorter than that following a tense alveolar stop /t'/ in word-initial position.

Lee (1996) found that nasals shorten the following consonants. Pre-consonantal vowels do not have much effect on the following consonants.

In his investigation on the effect of post-vocalic consonants, Kim (1974) argued that voicing, tenseness, or articulator movement have a greater effect on the preceding vowel duration than glottal opening. He found that though a vowel after /s/ is very short, it is not the case with a vowel before /s/, the duration of which is similar to that before /t/, or /k/. He explained that the shortening effect of /p^h/ and /p'/ relative to /p/ on the preceding vowel could be attributed either to their voicelessness or tenseness; and the lengthening effect of intervocalic /t/, /k/, and /ts/, to their voicedness or laxness. The shorter duration of the vowel before /p/ than before /t/ and /k/ could then be attributed to the bilabial articulation.

Lee (1990) said that a vowel is longer before a voiced consonant than before a voiceless consonant, other things being equal, though the difference in duration in this case is not as large as in English.

Kang (2000) found that a vowel preceding a tense alveolar stop or an aspirated alveolar stop is 60% shorter than that preceding a lax alveolar stop in word-medial position.

In this section, the durational effects of surrounding segments have been reviewed. Many researchers are in general agreement with the ranking of these effects. In pre-vocalic position, obstruents with aspiration have the greatest shortening effect on the following vowel; sonorants have the greatest lengthening effect on the following vowel. Other obstruents have smaller effects, with some obstruents lengthening the vowel and some shortening the vowel. Many studies show that the effect of post-vocalic consonant is small, but there appear to be some possible patterns. These problems need to be explored using phonological features, not just by using the name of the follwing segment. By doing this, the role of each phonological feature could be discussed. One of the aims of the experimental work described in chapters 4 and 5 is to obtain a clearer picture of the influence of surrounding segments on duration.

2.2.6 Summary

This review of Korean segment duration studies suggests that well-prepared speech data and feature selection could form the basis for an analysis of the Korean segmental duration pattern. Partly solved or controversial matters concerning segment duration could be detailed as follows:

(2.27)

a. Which type of phrase boundary has the most influence? UTT, IP, AP, PW boundaries were all claimed to have lengthening effects. However, more information was needed over which boundary is more important, the relative size of initial and final boundary effects, and whether syllables in post-initial or penultimate positions are also lengthened.

b. How does the structure of a syllable affect its constituents? In Korean, CVC, VC, V, and CV syllable structures can be observed. These structures are believed to have an influence on the segment duration. More information is required about how each syllable structure affects segment duration. The different behaviours of onset consonants and coda consonants also needs further study.

c. Which segmental features show a systematic effect on duration? In English, following segments have more influence than preceding segments. In Korean, it is claimed that preceding segments are more important than following segments.

Chapters 4 and 5 aim to provide data to address these controversies and, furthermore, to provide quantitative information useful for models for Korean text-to-speech.

NOTES

[1] Nespor and Vogel (1986) suggests three principles for the prosodic phonological representation. Because a syllable cannot contain 'one or more' onsets or 'one or more' rhymes, the syllable violates the principle 1. And in the ambisyllabic segment, the segment may at the same time a member of the rhyme of one syllable and the onset of the other, which violates the principle 2. And two segments may share a single feature, as in the harmony phenomena. The violation of principle 2 is the example of the strict layer hypothesis violation.

[2] Lee (1990) defines the phrasal tone as the pitch pattern overlaid on each rhythm unit excluding the last syllable of an intonation group.

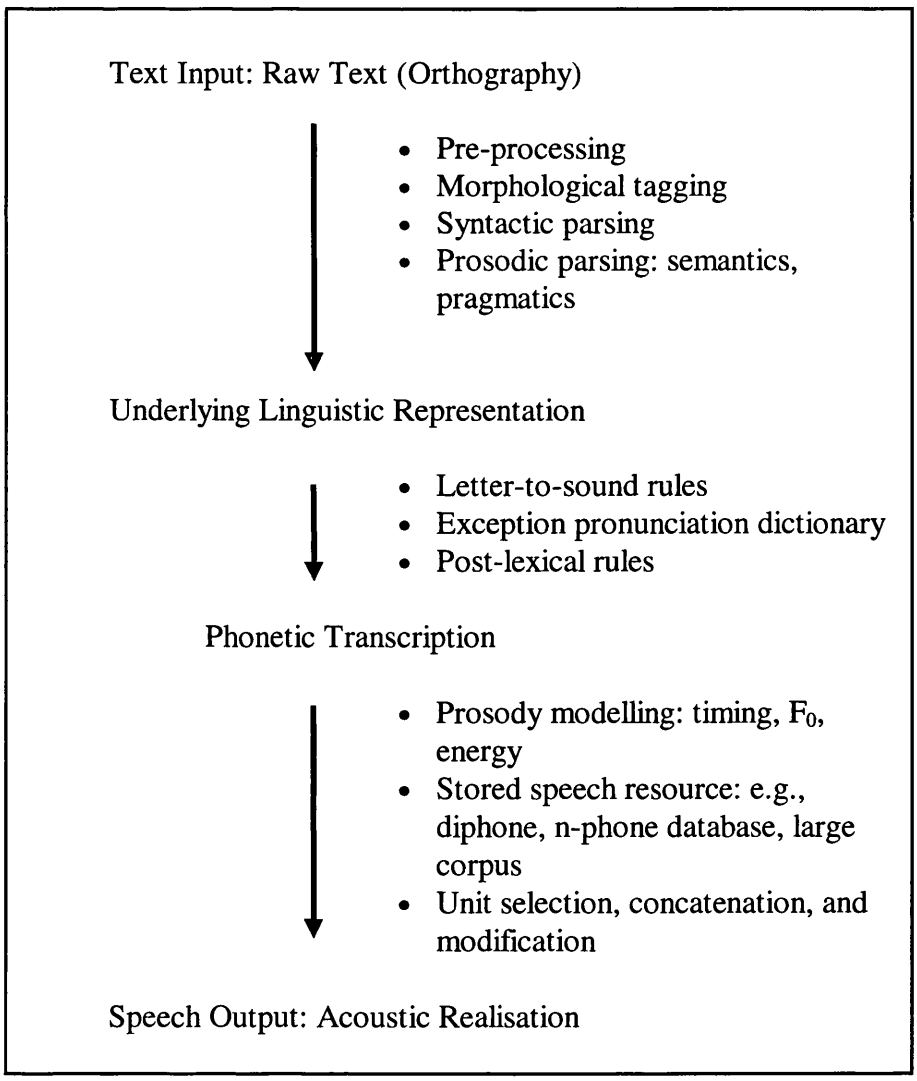
[3] Heo (1985) claims that the distinction among 'chrone', 'chroneme' and 'allochrone' is the same with the distinction among 'phone', 'allophone', and 'phoneme'.

3. TIMING IN TEXT-TO-SPEECH (TTS) SYSTEMS

3.1 Overview of Text-to-Speech

Thanks to the development of computer technology over the last few decades, linguists and engineers have been able to implement linguistic theories and descriptions within speech technology to create automatic text-to-speech (TTS) conversion systems in some world languages. The steps involved in converting text to speech are illustrated in Figure 3-1.

Figure 3-1.
Stages in Text-to-Speech conversion.



First, the input text is processed to an orthographic transcription of what has to be said. Numbers, abbreviations, and acronyms are transformed into plain text. Any special information about the type of material, or textual mark-up is taken into account. Often a lexicon or a morphological analyser is used to assign parts of speech information to each word; this aids later prosodic phrasing and prominence analysis. A syntactic analyser is then commonly used to group the words into clauses and grammatical phrases and to assign them syntactic functions. A subsequent stage of prosodic phrasing then breaks the syntactic constituents into performance units or prosodic phrases. When information structure and pragmatic information are added into the text through mark-up, these may be used to guide parsing and prosodic phrasing. The output of these analysis routines is a grouped and labelled word sequence.

The next stage of processing converts the word sequence into phonetic transcription. Pronunciations are assigned from a combination of a pronunciation dictionary and a set of letter to sound rules. The syntactic functions and parts of speech information may also inform this process. Post-lexical rules can be used, where the standard lexical form of pronunciation of a word is changed depending on the context that it appears in. The output of this stage is a hierarchical structure which describes both the phonological units required for the text and the phonological prosodic context in which they occur

The last stage of processing converts this phonological representation into a speech waveform. This involves the prediction of suitable intonation and timing, followed by means to realise the phonological units as sound. Different approaches to the second stage of processing are described in the sections 3.1.1 to 3.1.3 below. The prediction of intonation and timing can be made by rule, using knowledge of typical fundamental

frequency contours and durations for phonological events in different contexts (Klatt, 1973; van Santen, 1992; Local and Ogden, 1997). Recently, however there has been a change whereby the phonological description itself is used to search a database of known contours and timings to extract single known "cases" that match the prosodic pattern (Hunt and Black, 1996; Breen and Jackson 1998; Black and Taylor, 1999; Conkie, 1999). Depending on the choice of signal generation method, the predicted fundamental frequency and duration may be imposed on the generated sound by means of signal processing algorithms for prosody manipulation (e.g. Conkie, 1999). However some success has been made by systems that prefer a larger corpus of cases over signal modification (e.g. Hunt and Black, 1996; Black and Taylor, 1999). The prediction of a suitable timing for the phonological structure is at the heart of this thesis and is described in more detail in section 3.2. Even though some signal generation methods avoid modifying the timing of units during signal generation, it is still necessary to understand what phonological features influence timing in context.

3.1.1 Rule-based synthesis

Three approaches to signal generation have been widely used for TTS: rule-based synthesis, concatenative diphone synthesis, and corpus-based unit selection. Rule-based synthesis tends to be the preference of phoneticians and phonologists, who seek to encode a cognitive, generative model of human speech production. Rule-based synthesis usually uses a formant synthesiser for signal generation. Rules map phonological and phonetic properties to the control signals of a formant synthesiser. Holmes' (1973) parallel formant synthesiser consists of four parallel formants and a nasal formant, each excited by a mixture of voicing and noise. Klatt (1972) proposed "the desirability of using a hybrid synthesiser with cascaded formants (and an extra pole-zero pair for

mimicking nasalisation) for synthesis of sonorants, and parallel formants (with the same formant frequency values) for synthesis of obstruents." Klatt argued that the quantal theory of consonant place of articulation (Stevens, 1972) could be implemented directly by simple rules in such a synthesiser. The synthesiser includes components to simulate the generation of several different kinds of sound sources, components to simulate the vocal-tract transfer function, and a component to simulate sound radiation from the head (Klatt, 1980). The main problems with rule synthesis are lack of knowledge of how to map phonetic descriptions to formant parameters in a natural and coherent manner.

3.1.2 Concatenative synthesis

As opposed to rule-based synthesis, concatenative synthesis is based on the concatenation of recordings of elementary speech units to make a human-sounding synthetic speech signal. The evaluation of synthesis speech described in chapter 6 uses a concatenative method called MBR-PSOLA (Multi-Band Re-Synthesis Pitch Synchronous OverLap Add; Dutoit and Leich, 1993) for synthesis from a diphone database. Diphone units allows the modelling of some coarticulation effects across phones, and avoid the "targets" and "interpolation" metaphor used in rule synthesis. The concatenation process involves three procedures: (1) stretch or compress acoustic units; (2) attach successive acoustic units to each other; (3) impose an intonation contour. The information about timing imposed on the speech signal in (1) and (3) needs to be calculated using the prosodic component of the TTS system. When stage (2) takes place, there may be audible spectral discontinuities. In order to avoid this, this stage requires various forms of interpolation and smoothing. To minimise the distortions introduced by the diphone concatenation processing, care must be taken during the recording of the diphone

database. This is described in more detail in chapter 6 where a Korean diphone database is prepared.

3.1.3 Corpus-based synthesis

The corpus-based unit selection approach to signal generation is a recently developed technique, which is becoming widely used in speech synthesis (Hunt and Black, 1996; Taylor, 1999). This technique describes both the synthesis target and the components of the speech database as phonological trees, and uses a selection algorithm which finds the largest parts of trees in the database which match parts of the target tree. Often this method is used without explicit modification of pitch and timing during synthesis. The technique tries to avoid many of the errors made by prosody prediction modules by incorporating their operation implicitly in the selection process. In diphone synthesis a single diphone is used for every instance of that diphone in a target synthesis and its pitch and duration are modified by signal processing to match its target prosody. In unit selection synthesis, however there are many instances of each unit type, each with different pitch, durations and prosodic contexts. These are compared to the target and the most appropriate can be chosen. Furthermore the comparisons can be made using phonological features, thereby obviating the need to make explicit models of pitch and duration in Hertz and milliseconds. Unit selection algorithms are often successful at finding units of the appropriate pitch and duration specified in the target description. However, this technology requires an extremely large speech corpus, because it needs to be able to find a sequence of multi-segment units in the corpus that satisfies a number of requirements: (1) phone label match; (2) prosodic label match; (3) spectral match to adjacent units. Van Santen (1997) argued that this causes a problem of coverage and suggested that in corpus based synthesis, it is necessary to restrict synthesis to a single

task domain of limited vocabulary and sentence structure to satisfy the above three criteria.

3.1.4 Summary

Despite the success of corpus-based approaches to synthesis, the analysis of language timing is still an important endeavour. We still need an understanding of how segmental duration is affected by context. This understanding will help us to decide what features we need to index speech in the corpus, which features are most important in the unit selection matching function, what contexts need to be incorporated into the sentences recorded for the database, and what features have to be located and specified from the input text.

3.2 Modelling of Timing in TTS of English and Western Languages

As shown in Figure 3-1, the underlying linguistic representation in synthesis is symbolic, consisting of entities such as phoneme sequences, in combination with morphological, syntactic and prosodic information. The prosody prediction component computes the timing and pitch contour for the phrase. Prosody modelling refers to the equations involved in these computations, that is using the phonological structure to predict pitch and timing values in Hertz and milliseconds. Prosody modelling is one of the most important factors in determining the naturalness of synthesised speech (Horne, 2000). This section focusses on the development of the durational component of prosody modelling in text-to-speech conversion in English.

Following van Santen (1992), Campbell (2001) categorises current duration prediction systems into three classes: sequential rule systems, equation systems, and binary prediction trees. Such rule systems as Klatt's (1987) are considered sequential rule

systems which could be easily converted to equation systems such as van Santen's sums-of-products models (1992). CART models (Classification And Regression Tree; Breiman et al., 1984; Riley, 1992) are considered binary prediction tree systems which have been criticised by van Santen (1992) as just a collapsed form of lookup table. What these systems share is that they map symbolic input vectors provided by linguistic analysis routines onto acoustic quantities (duration), which may then used by the synthesis component to generate speech with the desired acoustic-prosodic characteristics. In the following sections, sequential rule systems (Klatt, 1987; Umeda, 1975), CART decision tree models (Breiman, 1984; Riley, 1992), and sums-of-products models (van Santen, 1992) are discussed for segmental duration prediction. Though neural networks have been used for duration modelling, this technique has the problem that its means of operation is not explicit, so that it does not give us any linguistic intuitions. Sums-of-products models and CART models were chosen in this thesis because they are believed to be widely used and representative of current duration prediction systems and because linguistic interpretation of their operation is possible.

3.2.1 Sequential rule systems

Klatt's (1973, 1979, 1987) duration model assumes that (1) each phonetic segment type has an inherent duration that is specified as one of its distinctive properties, (2) the effect of each phonological context can be expressed as a percentage increase or decrease in the duration of the segment, but (3) segments cannot be compressed shorter than a certain minimum duration. The duration of a segment can then be written as:

(3.1)

$$DUR = k(INHDUR - MINDUR) + MINDUR ,$$

where DUR is the output duration in ms, INHDUR is the inherent duration of the segment in ms, MINDUR is the minimum duration of the segment in ms (which for vowels is usually 45% of the inherent duration), and k is the scale factor determined by applying rules in contexts. Combining rules from previous researches, he proposed the following duration rules and contexts:

(3.2)

Rule 1. pause insertion before clause boundaries and before orthographic comma
Rule 2. clause-final lengthening
Rule 3. phrase-final lengthening
Rule 4. non-word-final shortening
Rule 5. polysyllabic shortening
Rule 6. non-word-initial consonant shortening
Rule 7. unstressed segment shortening
Rule 8. lengthening of emphasised vowels
Rule 9. shortening of vowels preceding voiceless consonants
Rule 10. shortening in consonant clusters
Rule 11. lengthening of stressed vowels or sonorants due to preceding aspirated plosives

Umeda (1975) proposed a similar duration rule for eight American English vowels as follows:

(3.3)

$$T = T_0 + S(K_1 + K_2 \times C),$$

where T is the output duration, T_0, K_1, and K_2 are constant values for each vowel, C is the consonantal context factor which depends on which consonant follows the vowel, and S is other factors such as the position of the vowel in a word and in a sentence, the word prominence, stress, speech rate, function word status, etc. This formula allows for interactions between segmental and prosodic features. Umeda argued that the duration of a vowel in word-medial position is little affected by its segmental context or its stress,

unlike Klatt (1979) who suggested a non-word-final shortening rule and a polysyllabic shortening rule. A description of the contextual effects described by Umeda were given in chapter 2.

In order to predict consonant duration, Umeda (1977) uses an additive model in which she added coefficients specified by various segmental and prosodic contexts to produce an estimated duration value. Thus the model has a large number of arbitrary constants to explain the various contextual effects. Umeda argued that consonant duration modelling is so complex that Klatt's (1973) model which used a fixed set of constants for all consonants could not predict the complexities of consonant duration.

YorkTalk (Local and Ogden, 1997) is a rule-based system that uses a non-linear model of timing. The basic timing unit is the syllable, which is modelled by the temporal interpretation function "overlay". Syllable overlay is calculated by using the distance of overlaid syllable and the distance of syllable in a monosyllabic utterance. The "distance" is a measurement of the separation between the onset and the coda in syllables. The same mechanism is applied to the temporal compression of prosodic feet. This distance has a direct relation with the following structural information: (1) Nucleus property: short or long; diphthong or monophthong; /ɑɪ ɔɪ aʊ/ vs. other diphthongs; (2) Coda property: simple or branching; (3) Rhyme property: heavy or light; voiced or voiceless; (4) Syllable strength: strong or weak; (5) position in Foot: initial, medial, or final; (6) the weight and strength of adjacent Syllables(s). Because in a single syllable, onsets and codas are 'overlaid' on syllable nuclei, the syllable end coincides with the rhyme end, nucleus end, and coda end. In polysyllabic words, the adjacent syllables are overlaid and the temporal compression is expressed as follows:

(3.4)

$$Syllable_n\ Start = Syllable_{n-1}\ End - Overlay$$

The YorkTalk model exploits a hierarchical metrical structure to describe the relationships between syllables. The model focusses on the temporal relations between syllables rather than on the durations of individual syllables or segments.

The strength of these rule systems lies in the fact that the rules are derived directly from linguistic analysis and phonological structure so that they are easy to understand and to use. Rules might be common across languages or at least make explicit the differences between languages. Rule systems also have weaknesses, however. They are incomplete in that they only cover some phenomena. Rules tend not to be tested on varied material such as sentences of different lengths. Though YorkTalk tries to create a declarative formulation of knowledge, generally rule interactions occur in the rule systems, which make them difficult to develop and extend. It is also not easy to adapt rule systems to changes in speaker, style, tempo, or genre.

3.2.2 Classification and regression tree (CART) modelling

The principle of the CART methodology was initially proposed by Breiman (1984) and it was applied to duration modelling by Riley (1992). CART analysis has become a common method for building classification models from simple feature data. This analysis was suggested by Riley (1992) as an alternative to heuristically-derived duration prediction rules for duration modelling in synthesis. CART trees partition a data set according to a binary tree of tests on feature values. For duration modelling, the nodes on the tree contain yes/no questions about the context features associated with a

segment, while leaves contain the mean duration of all training segments that end up in that partition. When the tree is being built, a set of values within one partition is split according to the available questions, and the split which minimises the variance of the data across two partitions is chosen. The tree building process terminates when partitions reach a minimum size, or when performance on some held out data reaches a maximum value.

Riley (1992) argued that CART is a promising method for duration modelling in that (1) the most significant features are selected based on statistics, (2) it provides "honest" estimates of its performance, (3) both categorical and continuous features are permitted, (4) humans can interpret and explore the result. In his analysis of 1,500 hand-segmented and labelled short utterances of English from a single speaker, Riley used the following features in a CART analysis:

(3.5)

a. Segment identity
b. Previous segment context (up to three segments to the left)
c. Following segment context (up to three segments to the right)
d. Stress (unstressed, primary, secondary)
e. Distance to the left boundary of the word in segments
f. Distance to the right boundary of the word in segments
g. Distance to the left boundary of the word in vowels
h. Distance to the right boundary of the word in vowels
i. Distance to the left boundary of the phrase in words
j. Distance to the right boundary of the phrase in words

Riley described features e-h as lexical position, and i-j as phrasing position. In order to make the CART approach more practical, Riley classifies the segment identity of each phone in terms of 4 features: consonant manner, consonant place, "vowel manner", and "vowel place", with each class having several values. The optimal regression tree had

about 250 nodes and predicted the durations of test data with an error of 23ms standard deviation. Unfortunately, the quality of the synthesised speech derived from the CART decision tree model was not noticeably better than that which was derived from rule based duration predictions. Riley suggested that this was due to insufficient training data in certain contexts or because of inadequate predictive power of the available features. Nevertheless, Riley argued that CART technique is valuable in that tree building and evaluation is rapid and that the technique may be easily applied to other feature sets, to other languages, to other speakers, and to other speaking rates. In other applications of CART to duration modelling, Deans, Breen, and Jackson (1999) showed that the performance of the CART decision tree model in the BT's Laureate TTS system was subjectively better than the rule-based method.

On his analysis of vowel duration in American English, House (1961) measured 12 vowels in a bisyllabic nonsense utterance [həC_iVC_i] and proposed an interesting diagram which is very similar to a CART decision tree as in Figure 3-2.

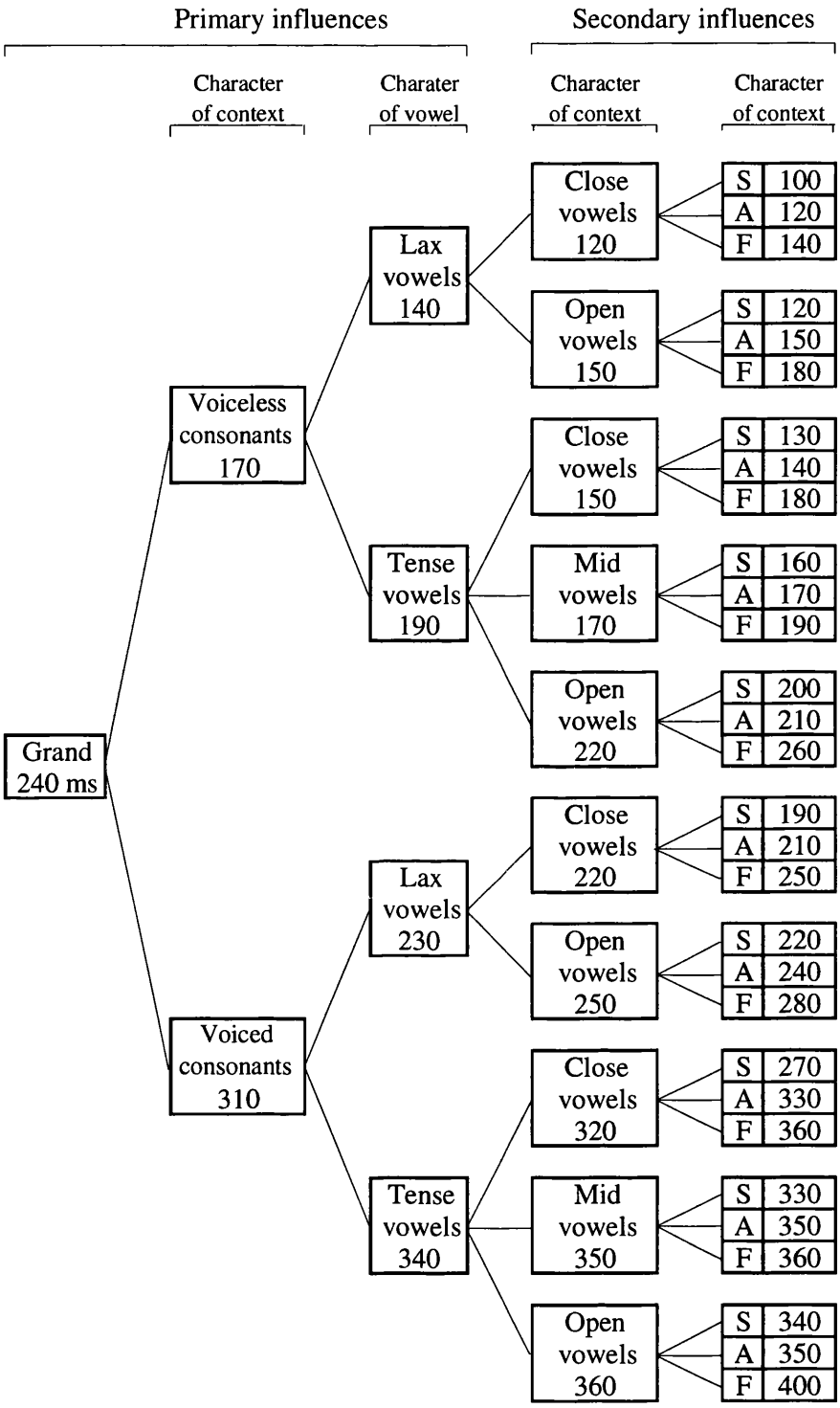
House reports the average durations of 12 vowels according to context. In the experiment, voiced and voiceless pairs of three stops, one affricate, and three fricatives form the preceding and following contexts of the vowel. House suggested that voicing of following consonants and tenseness of the target vowel have primary influences on vowel duration, which is a part of the phonology of the language and is learned by speakers of the language. The openness of the vowels and the manner of the following consonants have secondary influences on the vowel duration, which may be a function of their articulation. As shown in Figure 3-2, the voicing of the following consonants and the tenseness of the target vowel have a lengthening effect; as do the openness of the

target vowel and the presence of following fricatives. Though this experiment is rather limited in that it is restricted to nonsense words and does not use sonorant consonants, this diagrammatic representation of the linguistic feature interactions is easy to understand. CART analysis generates trees of a similar kind but in an automatic procedure with more rigorous evaluation.

In the CART model, the training data set described by feature strings is trained until additional features make no significant improvement in performance. After the tree is built after this training, the tree is then 'pruned' by removing questions and pooling leaf nodes so that the performance of the tree on the held-out evaluation data set is maximised. Finally the correlation between observed and predicted durations and the root mean squared error of prediction is found for the training set and the test set.

The strengths of CART modelling come from the ease with which trees may be built from duration data and from the speed of classification of new data. It also shows good performance in subjective terms. CART models cope with complex interactions because it makes very few assumptions about the structure of the data. Also, in theory, it is possible to interpret the results of modelling. The weakness of CART models lies in the fact that it cannot interpolate between known contexts to find values for unknown contexts. This is particularly a problem when the data set is small or when the number of factors is large (van Santen, 1994). Another weakness of this model is that it relies on objective function for partitioning that may not be the best in a perceptual sense. We also need to find ad hoc means to terminate tree growth. Despite Riley's claim (1992), the interpretation of models can actually be quite difficult for large trees.

Figure 3-2.
Factor influences on vowel duration suggested by House (1961).



S = stop
A = affricate
F = fricative

3.2.3 Sums-of-products modelling

Sums-of-products duration modelling is the third main technique applied to duration modelling. The motivation of sums-of-products models is that certain data cannot be fitted by a simple rule model. For example, in Luce and Charles-Luce (1985), there is a case where the duration pattern of /ɪ/ vowels preceding voiced stops in non-phrase-final position and /a/ vowels preceding voiceless stops in non-phrase-final position is reversed when both vowels are in phrase-final position. In the former environment, /a/ is longer than /ɪ/, in the latter /ɪ/ is longer than /a/. Campbell (2001) said that this contradicts the independence of the factors 'voicing of the following consonants' and 'position in the phrase'. Van Santen (1997) explained this phenomenon as a violation of 'single factor independence', showing that the generalised additive model needs to be extended to a sums-of-products model. In the above example, if /i/ is assumed to be amplified by the phrase position factor and /a/ is not, the phenomena could be modelled as follows:

(3.6)

$$\mathrm{DUR}(V, \mathrm{VCE}, \mathrm{PP}) = \exp[[S_{1,1}(V) \times S_{1,2}(\mathrm{VCE})] + S_{2,1}(\mathrm{PP}) + S_{3,1}(V)]$$

where V is the identity of the target vowel, VCE is the voicing of the following consonant, and PP is the phrasal position of the syllable. In this formula, the vowel duration is made up of three added components in which the first is composed of a product of terms. $S_{1,1}$ and $S_{3,1}$ has different values for each vowel; $S_{1,2}$ has different values for each voicing status; $S_{2,1}$ has different values for each phrase position. Campbell (2000) explained that "the effects of vowel identity are distributed over two terms, only one of which involves the phrasal position factor; that is, splitting 'a part of

the vowel factor $S_{3,1}(V)$." For example, we can arrange that one of two vowels /ɪ/ and /a/ is not amplified in duration by the phrasal position factor.

Van Santen (1992) argued that previous studies on contextual effects on segmental duration have focussed more on theoretical issues and putative underlying processes rather than completeness of empirical description. He said that the first step to construct a duration rule system for a TTS system is to make a list of factors which describe the contexts of a segment. The second step is to produce a duration model to explain complex interactions. In his TTS duration model, van Santen tries to show the durational behaviour of a single speaker and produce a simple equation to predict the durations based on contextual factors. Van Santen (1992) also said that duration databases for statistical analysis commonly confound factors in that not all combinations of factors and levels occur with equal frequency. According to him, the factor confounding results in mean durations that correspond to the levels of the factor of interest (the critical factor) being affected by other factors (confounding factors). One such example is word-final lengthening of unstressed syllables. Because, in word-final position, vowels in English are more likely to be unstressed and stressed vowels are more likely to be longer than unstressed vowels, statistics show that word-final vowels are shorter than non-word final vowels when all vowels are analysed altogether. However, when stressed and non-stressed vowels are analysed separately, word-final vowels are longer than non-word final vowels. When a pair of interacting factors such as the vowel and stress factors need to be described, the quasi-minimal pairs technique can be used. Segment durations occurring with a combination of levels on confounding factors and with several levels on the critical factor are divided into "quasi-minimal" sets. If there are not enough duration events for all sets, a piecewise multiplicative correction method

can be introduced which assumes that the effect of the critical factor and the joint effects of the remaining factors combine multiplicatively. Van Santen gives the example of the interaction between the syllabic position factor and the stress factor. He argued that these interactions are better described by a multiplicative rule than an additive rule. However, such interactions are not necessarily completely multiplicative, so he uses the term 'piecewise'. Where the quasi-minimal sets and multiplicative correction methods have difficulties with factors that have many levels, he introduces sums-of-products models, which he calls "a special case of an additive-multiplicative models, consisting of the sum of a single product term and any number of single-factor terms."

According to sums-of-products models, the duration for a unit in the context combination described by the feature vector **d** is given by:

(3.7)

$$\mathrm{DUR}(\mathbf{d}) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j)$$

Here, K is a set of indices, each corresponding to a *product term.* I_i is the set of indices of factors occurring in the i-th product term. Van Santen suggested that major interactions between factors could be described as a complete multiplicative rule (a single product term) or a piecewise multiplicative rule (more than two terms) in a "sums-of-products" model. Otherwise, other interactions are described as additive in the model. The multiplicative interactions predict constancy when effect size is measured as a percentage, the additive interactions do it when it is measured in milliseconds.

In van Santen's (1992) experiment on vowel duration in American English, he used training data of 18,000 vowel segments and test data of 6,000 vowel segments was used with the following context factors:

(3.8)

a. Vowel identity (V), 9 levels
b. Accent (A), 3 levels
c. Syllabic stress (S), 3 levels
d. Prevocalic consonants (C_{pre}), 3 levels
e. Postvocalic consonants (C_{post}), 6 levels
f. Within-word position: Preceding syllable/segment counts (W_{pre}), 3 levels
g. Within-word position: Following syllable/segment counts (W_{post}), 5 levels
h. Utterance position (U), 4 levels

As for the lexical stress and pitch accent factors, van Santen found that in accented words secondary stressed vowels are significantly shorter than primary stressed vowels; while in unaccented words the difference is neutralised. This is a good example of the effect of feature interactions and justifies the principal claim of his model. In both accented and unaccented words, secondary stressed vowels are significantly longer than unstressed vowels. As for the effects of post-vocalic consonants which have been widely investigated by previous studies, he agreed with previous findings in that post-vocalic voiced obstruents lengthen the preceding vowels more than the post-vocalic voiceless obstruents and that fricatives lengthen preceding vowels more than stops. However, in contrast to previous studies he found that the place of articulation of post-vocalic consonants does not have a significant lengthening effect; a result which is also found in our experiments. Combining the voicing and manner features, he suggested the following rank order of post-vocalic effects: voiceless stops, voiceless affricates, liquids, voiceless fricatives, nasals, voiced stops, voiced affricates and voiced fricatives. As for

the effects of the pre-vocalic consonant, he found that though there are effects, they are much smaller than those of post-vocalic consonants.

In his study of the effect of syllable position on vowel duration, van Santen found that the only significant factor is the number of syllables between the vowel and the word end. The only boundary he investigates is the utterance boundary. His results show that vowels in utterance-final syllables are longest, followed by those in utterance-penultimate syllables. In 2- and 3-syllable words, the pre-boundary lengthening effect is not so significant.

Based on such observations of the data, he suggested the following seven-term sums-of-products model for English vowel duration.

(3.9)

$$\log[\mathrm{Dur}(A, S, V, C_{pre}, C_{post}, W_{pre}, W_{post}, U)] = [S_{1,1}(A) \times S_{1,2}(S)] + S_{2,1}(V) + S_{3,1}(C_{pre}) + S_{4,1}(W_{pre}) + S_{5,1}(W_{post}) + [S_{6,1}(C_{post}) \times S_{6,2}(U)]$$

This model shows two multiplicative interactions: the first between pitch accent and syllabic stress and the second between post-vocalic consonant and utterance position. The first interaction is described as one term $[S_{1,1}(A) \times S_{1,2}(S)]$ and the last two terms $S_{5,1}(C_{post}) + [S_{6,1}(C_{post}) \times S_{6,2}(U)]$. The deviation of this formula was based on the "covariance analysis method" (van Santen and Olive, 1990). Also note the fact that the model as a whole operates in the log duration domain.

In van Santen (1994), a sums-of-products model for intervocalic consonant duration was built using 20,840 consonant tokens, using the following factors:

(3.10)

a. Consonant identity (C): one for each
b. Stress levels of the surrounding vowels (combined stress; S)
c. Within-word position (WP): word-initial, medial, and final
d. Accent status of the word (A): accented vs. deaccented vs. cliticised
e. Phrasal position (PP): phrase-final vs. phrase-medial

Based on his observation on the consonant duration, he suggested the following sums-of-products model.

(3.11)

$$\log[DUR(C, S, WP, A, PP)] = S_{1,1}(C, WP) + [S_{2,1}(C, WP) \times S_{2,2}(S)] + S_{3,1}(A) + S_{4,1}(PP)$$

In this model, the magnitude of the interaction between the two factors $S_{3,1}(A)$ and $S_{4,1}(PP)$ is small, so they remain as additive terms.

In order to model the duration of consonants in clusters, van Santen (1994) used a simple multiplicative formula, because the data was too sparse. He investigated following factors of consonant clusters:

(3.12)

Factors for consonants in syllable onsets:

a. Class of following segment
b. Class of preceding segment × syllable boundary
c. Stress accent of last vowel × syllable boundary
d. Stress accent of next vowel

(3.13)

Factors for consonants in phrase-medial codas:

a. Class of following segment × syllable boundary
b. Class of preceding segment

c. Stress accent of last vowel
d. Stress accent of next vowel × syllable boundary

(3.14)

Factors for consonants in phrase-final codas:

a. Class of following segment × syllable boundary (including silence)
b. Class of preceding segment
c. Stress accent of last vowel
d. Syllable boundary

The overall correlations between observed and predicted durations based on the above sums-of-products models on test data are shown in Table 3-1. Van Santen's modelling, carried out on his own data set, seems to show very high performance—better than many other studies have obtained using CART modelling (but on a different data set). The conclusion is that sums-of-products models can make good predictions of duration but with the drawback that parameters must be estimated for the many interactions between contextual factors.

Table 3-1.
Performance result of duration modelling
in van Santen's (1994) sum-of-products model.

Category	Correlation
Vowels	0.908
Intervocalic consonants	0.903
Consonants in clusters	0.874

Based on van Santen's sums-of-products model, Febrer, Padrell, Bonafonte (1998) modelled the phone duration of vowels and consonants for Catalan TTS. For vowel duration modelling they used following descriptors:

(3.15)

a. Vowel identity (v)
b. Stress: stressed, unstressed (a)
c. Sentence position (p): prepausal, non-prepausal
d. Class of post-vocalic phone (c): voiced, voiceless
e. Manner of articulation of post-vocalic phone (t), 8 levels: silence, vowel, nasal, vibrant, plosive, approximant, fricative, lateral

And their model is as follows:

(3.16)

$$DUR(v, a, p, c, t) = S_{1,1}(v) + S_{2,1}(v,a) + [S_{3,1}(v) \times S_{3,2}(p) \times S_{3,3}(c) \times S_{3,4}(t)]$$

For consonant modelling, they used following factors:

(3.17)

a. Consonant identity (v), 1 level each consonant
b. Stress of the syllable (a), 2 levels: stressed, unstressed
c. Sentence position (p), 2 levels: prepausal, non-prepausal
d. Syllable position (r), 2 levels: onset, coda

Their model of consonant duration is as follows:

(3.18)

$$DUR(v, a, p, r) = S_{1,1}(v) + S_{2,1}(v,a) + [S_{3,1}(v) \times S_{3,2}(p) \times S_{3,3}(r)]$$

Unfortunately, Febrer, Padrell, and Bonafonte (1998) did not provide comparable performance results such as the correlation coefficient or the RMS prediction error on test data.

The strength of sums-of-products models is that with relatively few parameters, durations can be well estimated from training data. Unlike CART models, they naturally interpolate to unseen contexts. A sums-of-products formula is small so it is easy to apply and understand. The weakness of this approach to modelling is that it is difficult to unravel all interactions in training data and it needs a large corpus with a wide variety of contexts.

3.3 Modelling of Timing in TTS of Korean and Oriental Languages

There are very few published studies on the modelling of Korean prosody for TTS. In this section, a couple of timing models of Korean are reviewed.

In order to make a duration model, Lee (1996) suggested using the following factors for Korean segments:

(3.19)

a. syllable structure: V, GV, GVC, CV, CGV, VC, GVC, CVC, CGVC
b. surrounding context: 13 features for consonants, 4 features for vowels
c. position of the syllable in the word: initial, medial, final
d. number of syllables in the word
e. position of the syllable in the phrase: initial, medial, final
f. number of syllables in the phrase

Lee used a regression tree model for the statistical processing of the segment duration data set. For training data, he used just 15 sentences spoken by three males and four females, each sentence spoken in three different tempos: fast, slow, and normal. Another sentence with three different tempos is used for test data. He calculates the correlation between predicted and observed duration of syllables and that of segments. More than

75% of segments have actual durations within 25 ms of the predicted value. The quality of the prediction is as in Table 3-2.

Table 3-2.
Results of regression tree models of Lee (1996).

Tempo	Normal	Fast	Slow
Correlation (segment)	0.74	0.69	0.74
Correlation (syllable)	0.86	0.83	0.88

Lee and Oh's (1999a, b) tree-based modelling used the following features for the prediction of segmental duration:

(3.20)

a. preceding segment, observed segment, and following segment (45 segment categories)
b. the part-of-speech context corresponding to segmental context (23 word classes)
c. location of the syllable in PW and AP (initial, medial, and final)
d. the length of PW and AP in syllables

They trained on 240 sentences (15,037 segments) and tested on 160 sentences (9,494 segments). They carried out two separate performance tests: one with part-of-speech information and the other without it. Their performance of tree-based modelling of segmental duration is shown in Table 3-3.

Table 3-3.
Performance of tree-based modelling of segmental duration in Lee & Oh (1999a, b).

	Correlation	RMSE
Including part-of-speech	0.820	22.06 ms
Excluding part-of-speech	0.823	21.88 ms

RMSE = root mean squared prediction error

They concluded that part-of-speech information did not contribute to improving the performance. The number of nodes used in the tree was 73.

There is no published research on sums-of-products modelling for Korean language segment duration, so we review work in modelling of another oriental language: Japanese. In their modelling of segmental durations for Japanese TTS synthesis, Venditti and van Santen (1998) used the following factors:

(3.21)

a. current phone identity (id), 1 level each segment
b. length of current vowel (leng): phonological vowel length
c. preceding phone identity (prev): voiceless stop, voiceless fricative/affricate, voiced stop, voiced fricative/affricate, flap, nasal, glide, vowel
d. following phone identity (foll): voiceless stop, voiceless fricative/affricate, voiced stop, voiced fricative/affricate, flap, nasal, glide, vowel
e. left prosodic context (left_pos): major phrase-initial, minor phrase-initial, intonation phrase-initial, accentual phrase-initial, non-initial
f. right prosodic context (right_pos): major phrase-final, minor phrase-final, intonation phrase-final, accentual phrase-final, non-final
g. accent status (acc): accented, downstep accented, preceding an accent in an accented AP, following an accent in an accented AP, in an unaccented AP
h. syllable structure (syll): open or closed by a geminate or moraic consonant
i. special morpheme status (spec)

For vowel duration modelling, they apply a separate sums-of-products model to each of three vowel subgroups: non-initial/non-final, final, and initial. Here is one example of the sums-of-products model from the non-initial/non-final subgroup:

(3.22)

$$\begin{aligned}&\text{DUR(id, leng, prev, foll, left\_pos, right\_pos, acc, syll, spec)} = \\ &S_{1,1}(\text{id}) + [S_{2,1}(\text{leng}) \times S_{2,2}(\text{prev})] + [S_{3,1}(\text{leng}) \times S_{3,2}(\text{foll})] + S_{4,1}(\text{left\_pos}) + \\ &[S_{5,1}(\text{leng}) \times S_{5,2}(\text{right\_pos})] + [S_{6,1}(\text{leng}) \times S_{6,2}(\text{acc})] + \\ &[S_{7,1}(\text{foll}) \times S_{7,2}(\text{syll})] + S_{8,1}(\text{spec})]\end{aligned}$$

The performance of their model is shown in Table 3-4.

Table 3-4.
Results of Sums-of-Products Model for Japanese vowels in Venditti and van Santen (1998).

Vowel category	Correlation	RMSE
Non-init/non-final	0.87	10.7 ms
Final	0.88	9.3 ms
Initial	0.85	14.9 ms
All vowels	0.88	16.8 ms

For consonant modelling, they divided consonants into 8 subgroups due to the confounding of the factors of initiality/finality and syllabic structure with the factors for the surrounding phone context: CV non-initial (burst), CV non-initial (others), CV initial (all), CyV non-initial (all), CyV initial (all), CyV /y/, moraic N non-final, and moraic N final. They used a multiplicative model rather than an additive-multiplicative model. The model performance is shown in Table 3-5.

Table 3-5.
Results of Sums-of-Products Model for Japanese consonants in Venditti and van Santen (1998).

Consonant category	Correlation	RMSE
CV non-init (bursts)	0.79	5.0 ms
CV non-init (others)	0.93	7.9 ms
CV initial (all)	0.96	8.1 ms
CyV non-init (all)	0.93	8.8 ms
CyV initial (all)	0.89	10.8 ms
CyV /y/	0.90	3.6 ms
Moraic N non-final	0.54	13.2 ms
Moraic N final	0.90	7.5 ms
All consonants	0.94	12.5 ms

The impressive performance result of this experiment is probably because vowels and consonants were sub-divided into so many groups rather than being modelled together.

3.4 Summary of Chapter 3

This chapter has described three approaches to speech synthesis and three approaches to duration modelling. Rule systems are now seen as a special kind of sums-of-products model, so they are no longer taken seriously because of the difficulty in developing and maintaining them. On the other hand, CART models are simple to build and use, with good performance. Sums-of-products models have shown excellent performance but seem rather tricky to build. They require complex data analysis to unravel the interactions between factors. Among the synthesis methods, formant and diphone-style synthesis require a numerical model that predicts durations in context; while corpus-based synthesis needs to know which factors are most important for unit-selection to get good prosody. So this thesis sets out to build a durational model of Korean using CART

and sums-of-products approaches to get reasonable performance within the limitations of the project time scale and to uncover the most important contextual factors. It is hoped that this will make a contribution to the understanding of Korean timing which is still rather undeveloped and under-researched compared to English and Japanese.

4. DESIGN OF CORPUS

4.1 Pilot Study

In order to obtain a general picture of the contextual effects on the segment duration and to investigate what factors affect vowel duration, a pilot study on Korean vowel duration was carried out using artificial carrier phrase sentences (Chung, Huckvale, and Gim, 1999).

4.1.1 Database

In order to investigate what factors affect vowel duration, a corpus of read speech was recorded and analysed. For this study, 600 artificial carrier phrase utterances were designed and recorded by a single speaker. The utterances systematically explored both syllable position and syllable composition within a sentence frame containing nonsense monosyllable pairs. For example: /ikʌsɯn V V solita/ was used to investigate inherent vowel duration; /ikʌsi CV(C) CV(C) solita/ for consonantal influences on vowel duration; /ikʌsi CV CVCVCVCV/ for prosodic influences on vowel duration. Three repetitions were recorded in an anechoic chamber on digital tape using 2 channels at 44,100 samples/sec/channel. Channel 1 was the speech signal from microphone, channel 2 was a Laryngograph signal. These were resampled to 16 kHz and transferred to disk. In order to make the speaker keep a consistent tempo, a prompting tool was used during the recording. Sentences were displayed on a monitor screen at five second intervals so that the speaker could read each sentence with a regular rhythm. A total of 1,054 syllables were hand-labelled and annotated. From these a table of vowel timing data was extracted, comprising the duration and a description of the segmental and phrasal context in which each vowel was found. The context was encoded as a set of 27 factors, each of

which could be said to be active or not for the vowel in question. The list of factors is given Table 4-1.

Table 4-1.
Factors used in the training corpus.

Factor	Factor description
f_0	phrase-final
f_1	phrase-initial
f_2	phrase-second
f_3	phrase-third
f_4	vowel after strong aspiration consonant
f_5	vowel after slight aspiration consonant
f_6	vowel after no aspiration consonant
f_7	vowel after fricative consonant
f_8	vowel after stop consonant
f_9	vowel after nasal consonant
f_{10}	vowel after affricate consonant
f_{11}	vowel after liquid consonant
f_{12}	vowel before stop consonant
f_{13}	vowel before fricative consonant
f_{14}	vowel before nasal consonant
f_{15}	vowel before liquid consonant
f_{16}	vowel after ambisyllabic consonant
f_{17}	vowel before ambisyllabic consonant
f_{18}	vowel after bilabial consonant
f_{19}	vowel after alveolar consonant
f_{20}	vowel after velar consonant
f_{21}	vowel after alveopalatal consonant
f_{22}	vowel before bilabial consonant
f_{23}	vowel before alveolar consonant
f_{24}	vowel before velar consonant
f_{25}	vowel after voiced segment
f_{26}	vowel before voiced segment

4.1.2 Parameter estimation of the timing model

The vowel durations and vowel contexts established from the training corpus were used to estimate the parameters of a simple multiplicative timing model based on Klatt (1987). The model estimates the duration of a vowel as a function of the identity of the vowel (v) and the context (c) in which it is found:

(4.1)

$$d(v,c) = d_{min}(v) + [d_{inh}(v) - d_{min}(v)] * F(c)$$

where $d_{min}(v)$ is the minimum duration of the vowel v; $d_{inh}(v)$ is the inherent duration of vowel v - i.e. the duration found in a 'neutral' context; and F(c) is a compression factor based on the context, independent of the vowel. F(c) in turn is calculated from:

(4.2)

$$F(c) = \prod_i f_i$$

where each compression factor f_i has a value that depends on one binary component of the context, for example f_0 represents the 'phrase-final syllable' context, which takes a value different to one in phrase-final contexts and a value equal to one elsewhere.

Although it is possible to hypothesise which contexts might influence vowel durations it is necessary to use an automated procedure to establish the relative importance of the compression factors and the optimal value for each factor. The procedure determined the best factors and the optimal factor values by minimising the squared error of prediction on the training data.

Input to the procedures was the 1,054 vowel duration measurements labelled according to the 27 different binary contexts hypothesised as being relevant for vowel duration.

Minimum and inherent durations were estimated from the distribution of durations for each vowel type, these are listed in Table 4-2.

Table 4-2.
Minimum and inherent duration of vowels.
$d_{min}(v)$ = minimum duration of the vowel (ms)
$d_{inh}(v)$ = inherent duration of the vowel (ms)

unit: ms

Segment	$d_{min}(v)$	$d_{inh}(v)$	Segment	$d_{min}(v)$	$d_{inh}(v)$
a	82	154	u	37	166
ja	89	190	wʌ	139	238
ʌ	79	168	we	118	230
e	71	179	wi	90	190
jʌ	144	240	ju	69	180
je	144	250	ɯ	68	161
o	51	175	ɯi	83	175
wa	138	232	i	48	164
jo	122	224			

For each hypothesised context in turn the best model comprising a single factor was found using a function minimisation procedure (Nelder and Mead, 1965). This process identified the most significant context and the optimal factor value for a model of a single factor. The context causing the greatest reduction in squared error was then accepted and the search repeated for the best two factor model by testing each of the remaining 26 contexts in turn. The best second factor is then chosen and the process repeated for a third factor and so on until the squared error fails to fall by a significant amount, in this case at about nine factors. The result of this procedure is shown in Table 4-3. The final model of 9 factors explains over 80% of the variance in the training data.

Table 4-3.
Factor distribution.

Number of Factors	Add Factor	Squared Error (ms^2)	Variance %
0		5,902,000	100
1	f_0	4,382,000	74.4
2	f_{12}	2,716,000	46.1
3	f_1	1,869,000	31.7
4	f_{14}	1,386,000	23.5
5	f_4	1,271,000	21.6
6	f_{25}	1,234,000	20.9
7	f_{13}	1,202,000	20.4
8	f_{17}	1,156,000	19.6
9	f_{15}	1,129,000	19.2

From this result, a simpler equation was produced to predict the vowel durations in the training data. The chosen 9 contexts were simplified using 5 phonological categories as follows:

(4.3)

F(c) = PP * CM * AS * VOC * AMB,

where:

PP (Phrasal Position Factor) =
1.72, if the vowel is in the phrase-final position (f_0),
0.93, if the vowel is in the phrase-initial position (f_1),
1, elsewhere.

CM (Consonant Manner Factor) =
0.31, if the vowel is before a stop consonant (f_{12}),
0.26, if the vowel is before a nasal consonant (f_{14}),
0.33, if the vowel is before a fricative consonant (f_{13}),
0.73, if the vowel is before a liquid consonant (f_{15}),
1, elsewhere.

ASP (Aspiration Factor) =
0.82, if the vowel is after a strong aspiration consonant (f_4),
1, elsewhere.

VOC (Voicing Factor) =
0.33, if the vowel is after a voiced consonant (f_{25}),
1, elsewhere.

AMB (Ambisyllabicity Factor) =
1.59, if the vowel is before an ambisyllabic consonant (f_{17}),
1, elsewhere.

4.1.3 Summary of the pilot study

In this pilot study of contextual effects on segment duration in artificial carrier phrase sentences of Korean, it was found that the manner of surrounding consonants had a more significant effect on the duration of vowels than their place of articulation. The largest lengthening context was when the vowel was in phrase-final position. It was this result which influenced the use of phrase factors and manner in the feature index used in the duration modelling of vowels in the main experiment.

4.2 Material of Main Corpus

The main corpus was designed and built using lessons learned from the pilot study. Firstly, to capture interactions between factors, and to build a more sophisticated model, it was necessary to collect a larger data set. The corpus also needed to be extended to the investigation of consonant durations. We also wanted to use more natural material including sentences with more than one prosodic phrase. This would allow the study of the effects of different phrase boundaries. Information about syllable structure and syllable position was also needed. Finally we needed to use more automatic annotation and data processing to cope with a larger corpus.

The main corpus consisted of 670 sentences spoken by one speaker in a news reading style. This choice was made since contextual factors have to be rich enough to capture

all aspects that affect timing (van Santen, 1997) and so it is necessary to avoid lists and sentences of repetitive structure. The size of 670 sentences was chosen because this was believed to be big enough to cover the majority of segmental contexts and was comparable to or bigger than those of similar studies[1]. One speaker was chosen for the database because, as pointed out in Han (1964), Lehiste (1970) and Lee (1990), among others, the speech data should be from one individual to obtain a coherent pattern of variation in context. News reading was chosen because it is believed that control over speaking style helps to reduce variability and news texts seemed most appropriate for speech synthesis applications, because they are factual and dense in information.

News scripts from two main Korean broadcasting stations were chosen: KBS (Korea Broadcasting System) and MBC (Munhwa Broadcasting Corporation). The script of the KBS 9 o'clock news broadcast on January 19, 2000 and that of the MBC 9 o'clock news broadcast on January 20, 2000 were downloaded from the internet. The KBS news script contained 412 sentences and the MBC news script contained 338 sentences. From these, the 670 sentences were chosen after the removal of speech errors by the speaker and those utterances which were incomplete or which seemed less grammatical. The sentences were divided into two data sets: 80% went into the training data set (42,103 segments in 535 sentences), while 20% went into the test data set (10,737 segments in 135 sentences). Besides the 670 sentences, extra evaluation data set was also prepared for the evaluation process of CART models (10,609 segments in 135 sentences).

4.3 Subject

The subject was a male speaker of modern standard Korean who had lived in Seoul, Korea for 16 years and had lived in London, England for the last 3 years. He was 20

years old and did not have any experience in this kind of recording. Because he is in the category of younger generation who uses modern standard Korean, he did not make phonological distinctions between short vowels and long vowels, between /e/ and /ɛ/, /we/ and /wɛ/, between /je/ and /jɛ/, between /wi/ and /y/, and between /we/ and /ø/. Though he was well informed of the phonological distinction between these vowel combinations, he did not distinguish them even in a citation form. Disappearance of the phonemic vowel-length contrast in modern standard Korean among younger generation has been described by Koo (1986) and Jun (1998).

4.4 Recording

The recordings were made in an anechoic chamber on digital tape using 2 channels at 44,100 samples/sec/channel. Channel 1 was the speech signal from microphone, channel 2 was a Laryngograph signal. They were resampled to 16 kHz and transferred to disk. The recordings were carried out in 12 sessions over a two-month time span. Though fewer sessions would have been ideal, the speaker found it difficult to maintain voice quality after 30 minutes of recording per day. In order to check for variations, the tempo and the fundamental frequency of the utterance were monitored by the author. The first 80% of the session recordings were allocated to the training data set and the last 20% to the test data set. Subsequent further recordings were made for the evaluation data. The speaker was prompted with a script displayed on a computer monitor, and sessions were recorded without interruption. The speaker was requested to read each sentence rapidly and fluently to simulate a real news reading style. Sentences containing misreadings and disfluencies were repeated until a fluent utterance was produced.

4.5 Phonetic Transcription

In order to label the phonological structure of the spoken utterances, it was first necessary to develop standards for transcription. To achieve this, the transcription inventory was defined, a pronunciation dictionary was built, sentence transcripts were produced, and the signal aligned with phonetic transcriptions. Phonetic transcriptions of the 670 sentences were generated from a dictionary, a set of phonological rules was applied[2], and then alignments were performed automatically and then hand-checked. First of all, by using a Romanisation program, the orthography of the individual words in Korean was automatically converted to their corresponding Roman characters. Before the automatic conversion, special characters such as English words, numbers and abbreviations were hand-checked and manually converted into Roman characters in advance. Then these Roman characters were sorted out into word units to construct a pronunciation dictionary.

Next, obligatory phonological rules as illustrated in Table 4-4 were applied to the underlying phonological representations to generate a standard pronunciation of each word. These rules are obligatory or nearly-obligatory, irrespective of speaker's identity, or speech rate.

Table 4-4.
Phonological rules used in the pronunciation dictionary.

Phonological rule	Conversion example	Meaning
Cluster simplification	/hɯlk/ → [hɯk]	'soil'
Coda neutralisation	/os/ → [ot]	'clothes'
Consonant nasalisation	/kukmin/ → [kuŋmin]	'people'
Tensing	/multsa/ → [mults'a]	'supplies'
Aspiration	/nohta/ → [not^{h}a]	'put + ending'
Flapping	/tolo/ → [toɾo]	'street'
Palatalisation	/kuti/ → [kutsi]	'obstinately'
Lateralisation	/ʌnlon/ → [ʌllon]	'mass media'

Then, one standard pronunciation for each word was constructed in the form of pronunciation dictionary for speech synthesis. The phonetic transcription of the sentence was then generated by concatenating phonetic transcriptions. These processes could be summarised as follows:

(4.4)

a. Korean (orthography):
바람과 햇님이 서로 힘이 더 세다고 다투고 있었습니다.

b. Underlying phonological representation:
/ palamkwa hɛsnimi sʌlo himi tʌ setako tat^{h}uko is'ʌs'sɯpnita /

c. Phonetic transcription derived from the pronunciation dictionary:
[pa**ɾ**amkwa hennimi sʌ**ɾ**o himi tʌ setako tat^{h}uko is'ʌ**ts'**ɯ**m**nita]

In the above transcription, the bold characters represent phonetic changes. In Korean, intervocalic /l/ obligatorily becomes flap /ɾ/. So /palamkwa/ in the first word became /paɾamkwa/ and /sʌlo/ in the third word became [sʌɾo]. The second word /hɛsnimi/ underwent three phonological processes: first, the vowel /ɛ/ in the first syllable became [e] phonetically by a vowel neutralisation rule; secondly, /s/ in the coda position of the

first syllable became neutralised to [**t**] in the coda position; finally, the neutralised [t] was assimilated to the following nasal. So the final output became [hennimi]. In the final word /is'ʌs'sɯpnita/, /s'/ in the coda position of the second syllable was neutralised to [t]. /s/ in the onset position of the third syllable became tensed after neutralised plosive [t]. And the plosive /p/ in the coda position of the third syllable was nasalised before the following nasal. The final output of this word was [is'ʌ**ts'**ɯ**m**nita]. In this example sentence, four phonological rules were applied in the pronunciation dictionary to produce the phonetic forms: flapping, neutralisation of the consonant in the coda position, consonant nasalisation, and tensing.

4.6 Manual-checking of Phone Alignments and Prosodic Phrases

The phonetic transcriptions of each sentence were then aligned to the speech signal of corresponding sentence. In order to carry out a forced phone alignment, the HVite program from the Hidden Markov Model Toolkit (Young et al., 1996) was used. Hand-labelled phone units were used to train models for the phone recognition. Finally the phone alignment, pronunciation and prosodic phrase labelling of every sentence was manually checked. The hand-checking of the phone alignment was carried out according to criteria based on guidelines published by Korea Telecom (Chung et al., 1995, 1997)[3]. The details are as follows.

(4.5)

a. Stops and affricates were annotated with information of the closure duration, the burst and aspiration. In post-pausal position, 50 ms of closure duration was arbitrarily included in the duration. The stop closure onset in pre-pausal position was defined as the point at which energy in the region of F2 and the higher formants ceases to be visible on the spectrogram display. The closure duration of the stop in pre-pausal position was assumed to be 20 ms. Aspiration was marked as the duration between the burst and the first glottal pulse in the vowel. When stops or affricates were preceded by another stop, 30 ms of closure duration before the consonants were arbitrarily allocated to the second

consonant and the rest closure duration was assigned to the preceding stop. When stops preceded fricatives, they were annotated from the end point of the stop to when a change of zero-crossing rate[4] was found between two consonants.

b. Fricatives were annotated when high-frequency energy appeared. When it was difficult to find high-frequency energy, a change in zero-crossing rate was used as a cue to define the onset or the offset of fricatives.

c. Nasal boundaries were defined as the points when formant frequencies showed a discontinuity and the amplitudes of the formants were decreased. When two nasals were geminated, their boundary was determined from any change in the energy, otherwise the mid-point was chosen.

d. In the lateral [l], in some cases the amplitude of F1 is decreased. The mid-point of this transition was assumed to be the onset of lateral [l]. The flap [ɾ] was easily detected, because it only appeared between vowels. When it appears, it had a 20 - 30 ms duration with an energy decrease and weaker formants.

e. Both diphthongs and monophthongs were considered unitary vowels. Whenever formant frequencies appeared after consonants, they were considered to indicate the vowel onset. A diphthong was treated as a glide and a monophthong. An energy change could be observed in the onset of glides, but in this experiment, no boundary was annotated between the glide and the rest of the vowel. Nasalised vowels were annotated as oral vowels. In nasalised vowels, the amplitude of F2 and F3 were decreased. When two vowels were adjacent to each other, the formant change was first investigated. If there was still a difficulty in distinguishing the boundaries, the energy change between two vowels was used. Otherwise, the boundary was annotated at the mid-point.

Manual-checking of the phone alignment also allowed a check on whether the real utterance was in agreement with the predicted pronunciation. There are phonetic changes beyond those incorporated in the dictionary which are optional or sensitive to prosodic phrase boundaries (Chung et al., 1997; Lee, 1996a; Jun, 1993; Kwack, 1992; Oh, 1989). The following phonetic changes were manually checked and annotations modified after the automatic phone alignment.

Table 4-5.
Optional phonetic changes found in hand-checking.

Phonological rule	Conversion example	Meaning
h-deletion	/koŋhaŋ/ → [koŋaŋ]	'airport'
Place assimilation	/hamnita/ → [hammita]	'do + respect + end'
monophthongisation	/hweɯɻi/ → [hwei]	'conference'
	/tosiɯɻi/ → [tosie]	'city + genitive'
Tensification	/jukwʌntsa/ → [juk'wʌntsa]	'electorate'
L-nasalisation	/ɯmunlon/ → [ɯmunnon]	'phonology'

Also the transcription needed to be enhanced with the insertion of pauses. The transcription of pauses and prosodic boundaries was also added during checking. As discussed in chapter 3, four prosodic boundaries are assumed in this thesis: utterance (UTT), intonational phrase (IP), accentual phrase (AP) and phonological word (PW). Each sentence has the default UTT boundary in the starting and the end point of the sentence. We did not use any special symbol for the UTT boundary. When we found a clear pause in the actual utterance, we not only marked the pause in the annotation file in the speech data, but also put an IP boundary in the pronunciation string. We used the diacritic symbol "/" for marking the IP boundary. Based on the fundamental frequency contour in the speech data, AP boundaries were also marked in the pronunciation. Each AP has an underlying tonal pattern of LHLH which is sometimes phonetically realised as LH in a short AP (Jun, 1998). The AP boundary was marked in the pronunciation by using the symbol "^". Then PW boundaries were indicated in the pronunciation using the symbol "\". The PW is a morphological and syntactic unit which is demarcated by one content or functional word with one or more suffixes, case particles, or endings. When more than two prosodic phrase structure occur in the same place, we only marked the higher prosodic structure. For example, each IP boundary was also an AP boundary and a PW boundary. An example of these processes is as follows:

(4.6)

a. Expected phonetic transcription derived from the pronunciation dictionary:
[nampuk^{h}anɯi hwalpalhan ints'ʌk kjorjunɯn hanpantoɯi tsʌntseŋɯl
kts'ehanɯnte kijʌhako itt'ako malhal su itsɯmnita]

b. Actual pronunciation and prosodic phrasing:
[nampuk^{h}ane ` hwalpaɾan \ ints'ʌk \ kjorjunɯn / hanpantoe \
tsʌntseŋɯl \ ʌkts'eanɯnte ` kijʌhako \ itt'ako ` maɾal \ su \ its'ɯmmita]

In this example, the diphthong genitive particle [ɯi] in the first and the fifth words was pronounced as the monophthong [**e**] in the actual utterance. In the second word [hwalpalhan], two phonological rules were applied: h-deletion and flapping. After the deletion of [h] in the third syllable, [l] in the second syllable became a flap [**ɾ**] between vowels. The same process also applied to the penultimate word [malhal], so it was pronounced [maɾal]. [h] in [ʌkts'ehanɯnte] was deleted in the actual utterance between vowels. In the final word [itsɯmnita], [n] in the penultimate syllable was assimilated to the preceding nasal, so it became [itsɯm**m**ita]. There was also a pause after [kjoryjunɯn].

4.7 Database Processing

The annotated transcription was processed into a hierarchical prosodic structure encoded in XML comprising UTT, IP, AP, PW, syllable, onset, rhyme, nucleus and coda nodes as well as segments, which are described using features. Table 4-7 and Table 4-8 show the segment features used in the XML script. In order to ease the database processing, each IPA symbol was converted to two Roman characters as in Table 4-6.

Table 4-6.
Phone to symbol conversion chart.

IPA	i	u	e	o	a	ʌ	ɯ	wa	we	wi	wʌ	ja	je
Roman	ii	uu	ee	oo	aa	vv	xx	wa	we	wi	wv	ya	ye
IPA	jo	ju	jʌ	ɯi	m	n	ŋ	l	ɾ	p^h	p	p'	t^h
Roman	yo	yu	yv	xi	mm	nn	ng	ll	rr	ph	p0	pp	th
IPA	t	t'	k^h	k	k'	ts^h	ts	ts'	s	s'	h		
Roman	t0	tt	kh	k0	kk	ch	c0	cc	s0	ss	h0		

Table 4-7.

Vowel features used in XML script.

	First timing slot in nucleus				Second timing slot in nucleus			
Vowel	LAB	COR	DOR	OPN	LAB	COR	DOR	OPN
ii	N	Y	N	N	N	Y	N	N
uu	Y	N	Y	N	Y	N	Y	N
ee	N	Y	N	Y	N	Y	N	Y
oo	Y	N	Y	Y	N	N	Y	Y
aa	N	N	N	Y	N	N	N	Y
vv	N	N	Y	Y	N	N	Y	Y
xx	N	N	Y	N	N	N	Y	N
wa	Y	N	Y	N	N	N	N	N
we	Y	N	Y	N	N	Y	N	Y
wi	Y	N	Y	N	N	Y	N	N
wv	Y	N	Y	N	N	N	Y	Y
ya	N	Y	N	N	N	N	N	N
ye	N	Y	N	N	N	Y	N	Y
yo	N	Y	N	N	N	N	Y	Y
yu	N	Y	N	N	Y	N	Y	N
yv	N	Y	N	N	N	N	Y	Y
xi	N	N	Y	N	N	Y	N	N

Table 4-8.
Consonant features used in XML script.

CNS	SON	NAS	LAT	SPR	CST	VOI	CNT	DEL	LAB	COR	ANT	DOR
mm	Y	Y	N	N	N	Y	N	N	Y	N	N	N
nn	Y	Y	N	N	N	Y	N	N	N	Y	Y	N
ng	Y	Y	N	N	N	Y	N	N	N	N	N	Y
ll	Y	N	Y	N	N	Y	N	N	N	Y	Y	N
rr	Y	N	N	N	N	Y	N	N	N	Y	Y	N
ph	N	N	N	Y	N	N	N	N	Y	N	N	N
p0	N	N	N	N	N	N	N	N	Y	N	N	N
pp	N	N	N	N	Y	N	N	N	Y	N	N	N
th	N	N	N	Y	N	N	N	N	N	Y	Y	N
t0	N	N	N	N	N	N	N	N	N	Y	Y	N
tt	N	N	N	N	Y	N	N	N	N	Y	Y	N
kh	N	N	N	Y	N	N	N	N	N	N	N	Y
k0	N	N	N	N	N	N	N	N	N	N	N	Y
kk	N	N	N	N	Y	N	N	N	N	N	N	Y
ch	N	N	N	Y	N	N	N	Y	N	Y	N	N
c0	N	N	N	N	N	N	N	Y	N	Y	N	N
cc	N	N	N	N	Y	N	N	Y	N	Y	N	N
s0	N	N	N	N	N	N	Y	N	N	Y	Y	N
ss	N	N	N	N	Y	N	Y	N	N	Y	Y	N
hh	N	N	N	Y	N	N	Y	N	N	N	N	N

CNS=Consonant, CST=CONSTR, CNT=CONT, DEL=DELR

The distinctive features used in this thesis are generally based on Chomsky and Halle (1964). However, for the place features of vowels and consonants, the thesis follows the idea of Clements' (1989) Unified Set Model[5] and Kim (1990)[6]. The selection of features for Korean in this thesis was carried out with descriptive convenience for duration modelling in mind. Each feature in the script has a binary value "Y/N" (Yes or No). For the consonant manner features, SON (sonorant), NAS (nasal), LAT (lateral), SPR (spread glottis), CONSTR (constricted glottis), VOI (voice), CONT (continuant), and DELR (delayed release) features were used. Other features except SPR and CONSTR are not different from other general phonological feature descriptions. In Korean, SPR and CONSTR features are necessary to distinguish among the aspirated, tense, and lax obstruents. For example, for the aspirated obstruents, *th/kh/ph/ch*, the combination of SPR='Y' and CONSTR='N' were used, because the wide opening of the glottis is

responsible for the aspiration. SPR='N' and CONSTR='Y' were used for tense obstruents, *tt/kk/pp/cc/ss*, because the tenseness of the glottis is the cause of this sound. Though there is still some amount of aspiration in producing tense obstruents, it is considered negligible in the phonological description. For lax obstruents, *t0/k0/p0/c0/s0*, SPR='N' and CONSTR='N' were used. In principle, there are significant amounts of aspiration in Korean lax obstruents, so it should be SPR='Y' and CONSTR='N', which is the same as the aspirated obstruent. According to Halle and Stevens (1971), the "stiff vocal cord" feature can distinguish the lax obstruent from the aspirated one in Korean. The aspirated obstruent has [+stiff vocal cord] feature and the lax obstruent is [-stiff vocal cord], so that aspirated obstruents are not voiced at all phonetically and lax obstruents become phonetically voiced in certain contexts. However, in this thesis only two features were used, SPR and CONSTR for the three-way distinction of Korean obstruent for the descriptive convenience. Thus lax obstruents had SPR='N'

For the consonant place features, LAB (labial), COR (coronal), ANT (anterior), and DOR (dorsal) were used. The ANT feature was introduced to distinguish between palato-alveolar obstruents and alveolar obstruents in Korean. It was assumed that the glottal fricative *hh* has a negative value in all the consonant place features.

This thesis generally followed Clements' (1989) idea that the place features for description of vowels should be based on the consonant place features. Thus the vowel place features were also based on LAB (labial), COR (coronal), and DOR (dorsal). For modelling, it was hoped that this might generate regularities across vowels and consonants. The LAB feature was used to describe rounded vowels and the COR feature was used to mark the frontness of the vowel. The DOR feature was used to

mark back vowels. High vowels were marked with OPN='N' feature and the other vowels have OPN='Y'. A "closed" feature was not used, though mid vowels could have been distinguished from the vowel *aa* by this feature. Because *aa* is the only central vowel in Korean[7], it could still be distinguished without using "closed" feature, since *aa* has the property of OPN='Y' and negative values for the other features. Other mid vowels have at least one positive value for the other features.

The syllable nucleus has two timing slots, though it is considered one segment. When the nucleus is occupied by a diphthong, the glide occupies the first slot and the rest of the vowel occupies the second. The features of the glide component are the same features used for the equivalent vowels. For example, the glide [j] shares the same feature values with the monophthong [i], the glide [w] with the monophthong [u], and the glide [ɰ] with [ɯ]. When the contextual effects between diphthongs and onset consonants were investigated, the feature values of the first timing slot of the diphthong were compared with those of onset consonants.

4.8 ProXML Processing

Each phonetic transcription was parsed into a hierarchical prosodic structure in which the symbolic transcription is replaced by feature descriptions stored in tree nodes. Each pronunciation was encoded as a metrical structure comprising syllables, onset, rhyme, nucleus and coda nodes as well as the segment, which are described using features. The prosodic annotation was a hierarchy of UTT, IP, AP and PW nodes. Attributes were added to the nodes in the hierarchy to reflect the prosodic information at all levels. After this process, the hierarchical structure was stored in extensible mark-up language XML (Huckvale, 1999). The hierarchical structure was then aligned to the checked

annotations in the speech signal. One example of the output after this process is shown in (4.7).

(4.7)

One example of the output of ProXML processing

```
<?xml version='1.0'?>
<!DOCTYPE KORSYNTH SYSTEM "korsynth.dtd" >
<KORSYNTH>
<UTT START="0.2367" STOP="2.2400">
    <IP START="0.2367" STOP="2.2400">
        <AP START="0.2367" STOP="1.1148" TYPE="PRE-NUCLEAR">
            <PW DUR="1" START="0.2367" STOP="1.1148" STRENGTH="STRONG">
                <SYL DUR="1" START="0.2367" STOP="0.3517" STRENGTH="WEAK"
                WEIGHT="HEAVY">
                    <RHYME CHECKED="N" DUR="1" START="0.2367" STOP="0.3517"
                    STRENGTH="WEAK" VOI="Y" WEIGHT="HEAVY">
                        <NUC CHECKED="N" DUR="1" LONG="Y" START="0.2367" STOP="0.3517"
                        STRENGTH="WEAK" VOI="Y" WEIGHT="HEAVY">
                            <VOC COR="N" DOR="Y" DUR="1" INHDUR="0" LAB="N" MINDUR="0"
                            OPN="N" START="0.2367" STOP="0.2942">x</VOC>
                            <VOC COR="Y" DOR="N" DUR="1" INHDUR="0" LAB="N" MINDUR="0"
                            OPN="N" START="0.2942" STOP="0.3517">i</VOC>
                        </NUC>
                    </RHYME>
                </SYL>
                <SYL DUR="1" START="0.3517" STOP="0.5359" STRENGTH="WEAK"
                WEIGHT="HEAVY">
                    <ONSET DUR="1" START="0.3517" STOP="0.4380" STRENGTH="WEAK">
                        <CNS CNSANT="N" CNSCOR="N" CNSDOR="Y" CNSLAB="N" CONSTR="N"
                        CONT="N" DELR="N" DUR="1" INHDUR="0" LAT="N" MINDUR="0" NAS="N"
                        SON="N" SPR="N" START="0.3517" STOP="0.4380" VOCCOR="N"
                        VOCDOR="Y" VOCLAB="Y" VOCOPN="N" VOI="N">k0</CNS>
                    </ONSET>
                    <RHYME CHECKED="N" DUR="1" START="0.4380" STOP="0.5359"
                    STRENGTH="WEAK" VOI="Y" WEIGHT="HEAVY">
                        <NUC CHECKED="N" DUR="1" LONG="Y" START="0.4380" STOP="0.5359"
                        STRENGTH="WEAK" VOI="Y" WEIGHT="HEAVY">
                            <VOC COR="N" DOR="Y" DUR="1" INHDUR="0" LAB="Y" MINDUR="0"
                            OPN="N" START="0.4380" STOP="0.4869">w</VOC>
                            <VOC COR="N" DOR="N" DUR="1" INHDUR="0" LAB="N" MINDUR="0"
                            OPN="Y" START="0.4869" STOP="0.5359">a</VOC>
                        </NUC>
                    </RHYME>
                </SYL>
```

The XML indicates the starting point and the end point of each prosodic phrase and segment. For example, <IP START="0.2367" STOP="2.2400"> implies that the

designated IP starts at 0.2367s and ends at 2.2400s in the seconds. These duration information can be found in all the segmental and prosodic phrase structures.

4.9 Generation of Training and Test Data for Modelling

For the modelling process, a feature string for each segment was automatically generated from the phonological structure using the ProXML scripting language (Huckvale, 1999). The script looked at each segment in turn and constructed a binary or n-ary feature string from the properties of the target segment, the properties of its neighbours and its position in the prosodic structure[8]. Each segment was annotated with the following features together with the actual duration:

(4.8)

a. phonemic identity of the target segment, e.g. segment name, or phonemic features of the target segment, i.e. major class features of the segment
b. phonemic features of the preceding and the following segments
c. syllable structure: position and structure of containing syllable
d. position of syllables in UTT, IP, AP and PW

Two groups of feature descriptions are prepared: one with general class features and the other with more sophisticated distinctive features. The first group of features, "Compact feature set", treats vowels and consonants separately and is used in CART and sums-of-products modelling. The second group, "Binary feature set", is used for the CART analysis only in order to investigate which distinctive features have most influences on duration.

Seven features for vowels were used in the "Compact feature set", each of which has a number of sub-levels.

Table 4-9.
Compact feature set for vowels.

Features	Sub-levels	Description	Features	Sub-levels	Description
id:	one			aspstp	aspirated stop
inherent property	for			aspaff	aspirated affricate
of the target segment	each segment			tnsstp	tense stop
man:	mono	monophthong		tnsaff	tense affricate
manner	di	diphthong		tnsfri	tense fricative
of the target segment				laxstp	lax stop
prev:	vow	vowel		laxaff	lax affricate
the property	nas	nasal		laxfri	lax fricative
of the preceding	lat	lateral		pause	pause
segment	fla	flap	**syll:**	cv	
	aspstp	aspirated stop	syllable	cvc	
	aspaff	aspirated affricate	structure	v	
	tnsstp	tense stop		vc	
	tnsaff	tense affricate	**left_pos:**	utt-init	UTT initial
	tnsfri	tense fricative	the distance	ip-init	IP initial
	laxstp	lax stop	from the syllable	ap-init	AP initial
	laxaff	lax affricate	to the left edge	pw-init	PW initial
	laxfri	lax fricative	of the phrase	non-init	Non-initial
	pause	pause	**right_pos:**	utt-fi	UTT final
foll:	vow	vowel	the distance	ip-fi	IP final
the property	nas	nasal	from the syllable	ap-fi	AP final
of the following	lat	lateral	to the right edge	pw-fi	PW final
segment	fla	flap	of the phrase	non-fi	Non-final

For consonants, features "id" and "man" were modified accordingly and a new syllable position feature "syllpo" was added as shown in Table 4-10.

Table 4-10.
Modified and additional features used
for consonants in the "Compact feature set".

Features	Sub-levels	Description
id: inherent property of the target segment	one for each segment	
man: the property of the target segment	stop aff fri nas lat fla	stop affricate fricative nasal lateral flap
syllpo: segment position in the syllable	on co	onset coda

Example feature strings from the "Compact feature set" for a vowel and a consonant are as follows.

(4.9)

a. 50 aa mono aspstp lat cvc pw-init non-fi
b. 60 nn nas vow laxstp cvc co utt-init non-fi

In the above example, the first number indicates ms duration of the segment (which will be replaced by a z-score in some experiments), the second character the name of the target vowel (a) or consonant (b) "id", the third the manner of the target segment "man", the fourth the property of the preceding segment "prev", the fifth the property of the following segment "foll", the sixth the syllable structure of the syllable of which the target segment is a member "syll", the seventh in (a) is the distance of the vowel to the left edge of the phrase "left_pos", and the eight in (a) is the distance of the vowel to the right edge of the phrase "right_pos". On the other hand the seventh in (b) is the segment position in the syllable "syllpo", the eight and the ninth in (b) is the distance of the

consonant to the left and right edge of the phrase. These output feature strings are used in the statistical analysis of Experiment 1 in chapter 5. One example set of feature strings for a complete sentence is given in Appendix 1.

In the "Binary feature set", each segment was annotated with a total of 69 features, describing the phonological contexts in more detail. For example, instead of just using the "initial" or "final" in the "left_pos" and "right_pos" parameters, the positions of the syllable in each phrase were divided into "first", "post-initial", "medial", "penultimate", and "last". The manner features of the preceding segment and the following segment were also differently classified. Rather than using "aspstp (aspirated stop)" feature, the feature was divided into two features "asp_ (aspiration)" and "stp_ (stop)". The reason for these classification is to investigate how individual features have influence on the duration of the segment. To find the overall performance of all segments, vowels and consonants were processed together in the binary feature set. Details of the features are described as follows.

Table 4-11.
The 69 features used in the "Binary feature set".

Feature	Description	Feature	Description
mono	monophthong	_lab	following labial
di	diphthong	_cor	following coronal
stp	plosive	_dor	following dorsal
aff	affricate	_glt	following glottal
fri	fricative	_hiV	following high vowel
nas	nasal	_mdV	following mid vowel
lat	lateral	_loV	following low vowel
fla	flap	CV	CV syllable structure
V_	preceding vowel	CVC	CVC syllable structure
vce_	preceding voiced segment	VC	VC syllable structure
nas_	preceding nasal	V	V syllable structure
lat_	preceding lateral	ON	onset
fla_	preceding flap	NUC	nucleus
stp_	preceding plosive	CODA	coda
aff_	preceding affricate	PW_1	first syllable in PW
fri_	preceding fricative	PW_2	post-initial syllable in PW
asp_	preceding aspiration	PW_m	medial syllable in PW
tns_	preceding tense consonant	2_PW	penultimate syllable in PW
lab_	preceding labial	1_PW	last syllable in PW
cor_	preceding coronal	AP_1	first syllable in AP
dor_	preceding dorsal	AP_2	post-initial syllable in AP
glt_	preceding glottal	AP_m	medial syllable in AP
hiV_	preceding high vowel	2_AP	penultimate syllable in AP
mdV_	preceding mid vowel	1_AP	last syllable in AP
loV_	preceding low vowel	IP_1	first syllable in IP
_V	following vowel	IP_2	post-initial syllable in IP
_vce	following voiced segment	IP_m	medial syllable in IP
_nas	following nasal	2_IP	penultimate syllable in IP
_lat	following lateral	1_IP	last syllable in IP
_fla	following flap	UTT_1	first syllable in sentence
_stp	following plosive	UTT_2	post-initial syllable in sentence
_aff	following affricate	UTT_m	medial syllable in sentence
_fri	following fricative	2_UTT	penultimate syllable in sentence
_asp	following aspiration	1_UTT	last syllable in sentence
_tns	following tense consonant		

An example feature string from the "Binary feature set" is shown below.

(4.10)

100 uu 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0
1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1

In the above example, the first number indicates ms duration of the segment (replaced by a z-score value in some experiments), the second character is the identity of the target segment, the rest of the digits are the binary features of the 69 feature vectors which were shown in Table 4-11. The first digit 1 indicates the segment has the positive value for the feature "mono" (monophthong), the next one 0 the negative value for the feature "di" (diphthongs), and the last digit is the value 1 for the feature "1_UTT" (last syllable in sentence). This output is used in the CART analysis of Experiment 2 in chapter 5. One example set of binary features for a complete sentence is given in Appendix 1.

4.10 Distribution of Prosodic Phrases and Segments

The distribution of prosodic phrases was similar across the training and test data sets as shown in Table 4-12.

Table 4-12.
Distribution of prosodic phrases in the data sets.

Data	UTT	IP	AP	PW	SYL	SEG
Training	535	1,028	2,270	6,410	19,071	42,103
Evaluation	135	259	597	1625	4778	10,609
Test	135	261	552	1,646	4,829	10,737

SYL=syllables, SEG=segments

The average number of daughters of each phrase unit type for each data set are shown in Table 4-13.

Table 4-13.
Average number of prosodic unit daughter nodes.

Data	IP/UTT	AP/IP	PW/AP	SYL/PW	SEG/SYL
Training	1.92	2.21	2.82	2.98	2.21
Evaluation	1.91	2.30	2.72	2.94	2.20
Test	1.93	2.11	2.98	2.93	2.22

Table 4-13 provides information about the distribution of segments and prosodic phrases in Korean. From this table, we can see that most Korean syllables have fewer than 3 segments. This means that most syllables in Korean have CV structure. This is mainly due to the syllabification process in connected speech, where every coda consonant in an intervocalic position tends to be syllabified to the onset of the following syllable in a connected speech. The parser used in this experiment reflects this by putting compatible consonants into the onset in this context. The label also shows information related to prosodic phrasing. The typical number of syllables per phonological word is close to three; the typical number of phonological words per accentual phrase is also close to three. The typical number of accentual phrases in an intonational phrase is close to two. It is possible that these facts could be useful in the construction of an algorithm for prosodic phrasing, because they are independent of the length of a sentence. However, this idea is not further investigated in this thesis.

The distribution and mean duration of the 42,103 segments in the training data set are shown in Table 4-14. In this table, segment [a] is the most frequent vowel with 3,786 occurrences (8.99%) and the segment [ɯi] is the least frequent with 49 occurrences (0.12%). Among sonorants, [n] is the most frequent with 4,399 occurrences (10.45%), and [ɾ] is the least frequent with 1,155 occurrences (2.74%). Among obstruents, [k] is the most frequent with 2,839 occurrences (6.74%), and [p'] is the least with 57 occurrences (0.14%). There was a very similar pattern of distribution, mean duration and standard deviation in the test data set. Further analysis of the test data set can be found in Appendix 2.

Table 4-14.
Distribution of segments in the training data set.

Phone	Counts	%	Mean (ms)	sd. (ms)
i	3650	8.67	58	37.40
u	1223	2.90	52	30.41
e	2197	5.22	84	45.39
o	1831	4.35	81	49.70
a	3786	8.99	86	44.27
ʌ	1725	4.10	75	40.29
ɯ	2264	5.38	49	27.41
wa	339	0.81	94	58.54
we	291	0.69	71	30.42
wi	106	0.25	86	42.75
wʌ	150	0.36	83	37.08
ja	84	0.20	101	42.28
je	86	0.20	87	32.23
jo	188	0.45	82	40.95
ju	207	0.49	80	38.04
jʌ	895	2.13	78	33.78
ɯi	49	0.12	111	53.03
m	1779	4.23	56	23.62
n	4399	10.45	62	39.66
ŋ	1572	3.73	69	27.95
l	1363	3.24	67	34.00
ɾ	1155	2.74	30	9.20
p^h	287	0.68	88	29.19
p	1179	2.80	53	25.10
p'	57	0.14	61	21.83
t^h	294	0.70	88	28.08
t	1952	4.64	49	22.10
t'	264	0.63	68	16.97
k^h	247	0.59	93	23.59
k	2839	6.74	57	33.19
k'	314	0.75	70	23.44
ts^h	503	1.19	101	30.35
ts	1458	3.46	68	33.47
ts'	191	0.45	72	16.39
s	1679	3.99	75	29.03
s'	602	1.43	104	20.23
h	898	2.13	45	24.53

sd. = standard deviation

NOTES

[1] The table below shows the comparison between the number of segments used in this experiment and in Lee and Oh (1999) and the comparison between the number of vowels used in this experiment and in van Santen (1992).

	Training data	Test data
	Segment count	Segment count
Vowel (this experiment)	19,071	4,829
Vowel (van Santen, 1992)	18,000	6,000
Consonant (this experiment)	23,032	5,908
Total (this experiment)	42,103	10,737
Total (Lee & Oh, 1999)	15,037	9,494

[2] Letter-to-phoneme conversion was carried out by using Jang's (2000) "Romanize" and "pronounce" scripts.

[3] The author actively participated in this project from 1993 to 1997 as a research assistant. The guidelines were decided after discussions between the author and other researchers. Topics of the projects were "A Study of Phonological and Grammatical Structures of Korean for the Implementation of ATS (Automatic Telephone System)" (1993-1995) and "A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System" (1996-1998).

[4] Zero crossing rate is determined by the frequency of zero line crossing in the waveform. Sibilant sounds such as fricatives and affricates have a higher zero crossing rate than stops.

[5] Clements (1989) proposes that vowels and consonants are characterised by a uniform set of articulator features, comprising the set [labial, coronal, dorsal, radical].

[6] Kim (1990) suggests that the root node of the segment should be divided into two class nodes [MANNER] and [PLACE]. The [PLACE] node dominates place feature [coronal] and [PERIPHERAL] place features of [labial] and [dorsal], which in turn dominate their own relevant features.

[7] This is still controversial, because some researchers believe that Korean vowel has central vowels /ɨ/ and /ə/ instead of back vowels /ɯ/ and /ʌ/.

[8] The script was created by the author in collaboration with Gordon Hunter.

5. ANALYSIS OF CORPUS

In this chapter, CART models, simple additive models, multiplicative models, and additive-multiplicative models are built and evaluated from the Korean corpus. The additive models, multiplicative models, and additive-multiplicative models are described under the headings of sums-of-products models. In the first experiment, the "Compact feature set" is used for modelling. Performance results with independent test data for these models are reported. In order to investigate which features are most important, CART trees are built in a stepwise fashion. Then, using the results of CART modelling, sums-of-products models are explored. In the second experiment, the "Binary feature set" is used for CART modelling only. As described in chapter 4, this data contains more detailed linguistic features. The aim of this experiment is to investigate which of these features are most important. At the end of the chapter, the linguistic implications of the behaviour of the models are summarised. We reflect back on the linguistic issues raised in chapter 2 about the timing of Korean.

5.1 Experiment I: "Compact Feature Set"

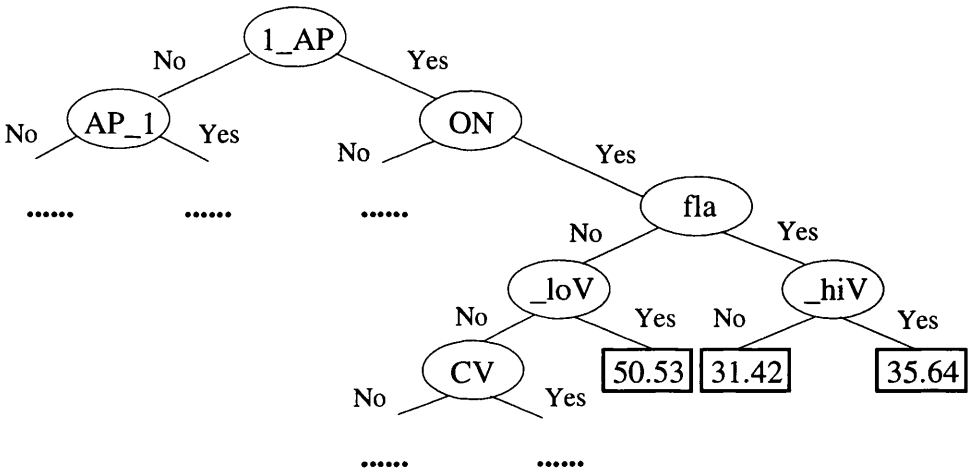
5.1.1 Classification and regression tree (CART) models

In this experiment, trees were built in a stepwise fashion in order to establish which contextual features were most important. In this approach each single feature is taken in turn and a tree consisting of nodes only asking questions of that feature is built. The single best tree is then kept and each remaining feature is taken in turn and tested to find the best tree possible with two features. The procedure is then repeated for a third, fourth, fifth feature and so on. This process continues until no significant gain in accuracy is obtained by adding more features. The *Wagon CART building program* (Black et al., 1999) was used as a tool for running this CART tree building process. An

example of the data format for a single segment record input to this process was illustrated in chapter 4.

An example of a CART decision tree is illustrated in Figure 5-1. The tree below is a part of the actual CART decision tree which is the result of the CART building process described in 5.1.1.3 below.

Figure 5-1.
A simplified example of a CART decision tree.



Taking a look at the question nodes (non-terminal nodes), we can see 1_AP (AP-final position), AP_1 (AP-initial position), and ON (onset) are principally used at the question nodes. In particular, 1_AP is used at the root node. The number on the terminal nodes is the mean linear ms duration of the segments identified by the combination of question nodes. When 1_AP feature combines with ON, fla (flap), and _hiV (following hi vowel) features, it produces a mean duration of 35.64 ms. But when 1_AP feature and ON feature are combined with _loV (following low vowel) feature, it shows a mean duration of 50.53 ms. This part of the tree can be summarised as following three bundles of features.

Table 5-1.
Estimated mean durations (ms)
of various feature bundles
in the CART decision tree.

1_AP	1_AP	1_AP
ON	ON	ON
fla	fla	_loV
_hiV		
35.64 ms	31.42 ms	50.53 ms

An advantage of CART analysis is that is easy to see the relative importance of each feature and the interactions of features in estimating the duration within the decision tree itself. However a disadvantage is that CART trees can grow quite large.

Four approaches to CART modelling with the "Compact feature set" are presented below. The first allows the tree to use questions based on the name and major class feature of the target segment as well as its segmental and prosodic context; this gives good performance but a tree which is less easy to interpret. The second just uses the name of the target segment in constructing a decision tree. The third restricts questions on the tree to the major class features of the target segment but not its name; it is hoped that this will force generalisations across segment types. The fourth replaces the linear ms duration values with durations calculated in z-scores of the log duration value of each segment type, without using the name or the major class feature of the target segment. The idea is to remove from the tree any influences caused by differences in inherent duration and variability of segment type.

5.1.1.1 CART analysis using segment names and class features

This CART modelling investigated the effect of using both the name and the manner features of each target segment on duration prediction. Two separate stepwise CART models for vowels and for consonants were trained. These used 19,071 vowels and 23,032 consonants in the training data set described by the name and major class features of each segment and the segmental and prosodic phrasal features describing the context. Training ended when additional features made no significant improvement in performance. This tree was then 'pruned' by removing questions and pooling leaf nodes so that the performance of the tree on the evaluation data set was maximised. The tree was then tested on 4,829 vowels and 5,908 consonants. In the CART decision tree for vowels, the pruning did not affect the number of different features used by the model, while in the model for consonants, only one feature "syll (syllable structure)" out of 8 was removed from the leaf nodes. Finally the correlation between actual and predicted durations and the mean squared error of prediction was found for the test set. The performance result is described in Table 5-2 and the relative importance of each feature in constructing the decision trees is illustrated in Table 5-3 and Table 5-4. The correlation for vowels was 0.78 and the root mean squared prediction error (RMSE) was 25.51 ms. The correlation for consonants was 0.71 and the RMSE was 24.20 ms. The observed and predicted values by this model are illustrated as a scatter plot in Figure 5-2 and Figure 5-3.

In the CART decision tree for vowels, the syllable (or segment) distance to the right phrase boundary (right_pos) was the most important factor and the manner feature of the target segment was the least important. It is interesting that the name of the target segment was less important than either the "right_pos" parameter or the property of

following segment (foll). In the CART decision tree for consonants, the segment name was the most important factor and the position of the segment in the syllable (syllpo) was the least important feature among those features used in the decision tree. The syllable structure feature was not used at all in the decision tree. An interesting result was that while the syllable distance to the right phrase boundary was more important than the distance to the left phrase boundary for vowels, it is reversed in consonants. However, for both vowels and consonants, the following segment (foll) feature was more important than the preceding segment (prev) feature.

5.1.1.2 CART analysis using segment names

This analysis was tried to find out how the performance is changed if the tree was built without using the manner feature of the target segment. Two separate stepwise CART models were trained using vowels and consonants from the training data set described by the name of each segment and the segmental and prosodic phrasal features describing the context. The constructed and pruned tree was then tested on the test data set. The correlation for vowels was 0.78 and the RMSE was 27.68 ms. The correlation for consonants was 0.70 and the RMSE was 24.56 ms. The rankings of feature importance in the decision trees were not different from those in section 5.1.1.1. This confirms that in the presence of segment names, the manner feature is not important.

5.1.1.3 CART analysis using segment class features

This tree is built to explore whether information of the manner of the target segment alone has any effect on the performance. Two separate stepwise CART models were trained and tested on vowels and consonants described by the class features of the target segment rather than the name of each segment, and contextual features. The correlation

for vowels was 0.76 and the RMSE was 28.56 ms. The correlation for consonants was 0.63 and the RMSE was 26.76 ms. Though there was no difference in the feature ranking for vowels, there were changes in the rankings for consonants. When the name of the target segment was not incorporated in the decision tree, the manner feature (man) became more important. Also syllable structure features were used in the tree for the first time.

5.1.1.4 CART analysis using z-scores of segments

In this CART model, each segment duration was first converted into log ms. Then each log duration was transformed to a z-score using the mean and standard deviation log ms for each phoneme. The log transformation was used to create more normal probability distributions for duration. In the CART model, a positive z-score corresponds to longer than mean duration and a negative z-score is shorter than mean duration. Because z-scores encode the inherent properties of each segment, the names and the major class features of the target segment were not used in this model.

A stepwise CART model was trained on the training data set using the z-score duration of each segment and the segmental and prosodic phrasal features describing the context. The resultant tree was tested on the test data. The correlation for vowels was 0.77 and that for consonants was 0.70. The RMSEs for this analysis are reported in Table 5-2. When only z-scores were used instead of the names and the major class features of the target segment, the preceding segment (prev) feature was more important than that of following segment (foll) in vowels, which was the opposite result from previous analyses. In terms of consonants, the preceding segment features was more important than the syllable distance to the right phrase boundary (right_pos), which was also the opposite of

the previous results. Other than these, there were no further changes in the feature rankings.

Table 5-2.
CART performance results for vowels and consonants in Experiment I using "Compact feature set".

	Vowels		Consonants	
	RMSE	Correlation	RMSE	Correlation
Name & manner	25.51 ms	0.78	24.20 ms	0.71
Name only	27.68 ms	0.78	24.56 ms	0.70
Manner only	28.50 ms	0.76	26.76 ms	0.63
z-score	26.01 ms	0.77	25.21 ms	0.70

Table 5-3.
Rankings of feature importance for vowels in the CART decision tree in Experiment I.

Rank:	1	2	3	4	5	6	7
Name & manner	right_pos	foll	name	prev	left_pos	syll	man
Name only	right_pos	foll	name	prev	left_pos	syll	
Manner only	right_pos	foll	prev	left_pos	syll	man	
z-score	right_pos	prev	foll	left_pos	syll		

Table 5-4.
Rankings of feature importance for consonants in the CART decision tree in Experiment I.

Rank:	1	2	3	4	5	6	7
Name & manner	name	left_pos	foll	right_pos	prev	man	syllpo
Name only	name	left_pos	foll	right_pos	prev	syllpo	
Manner only	left_pos	foll	man	right_pos	prev	syllpo	syll
z-score	left_pos	foll	prev	right_pos	syllpo		

Figure 5-2.
Observed vs. predicted duration for all tested vowels using names and manner feature of the target segment in "Compact feature set" (CART model).

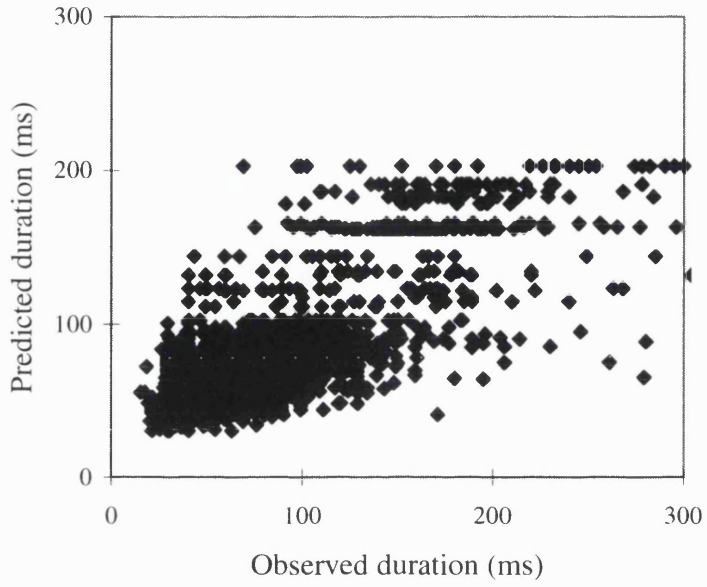
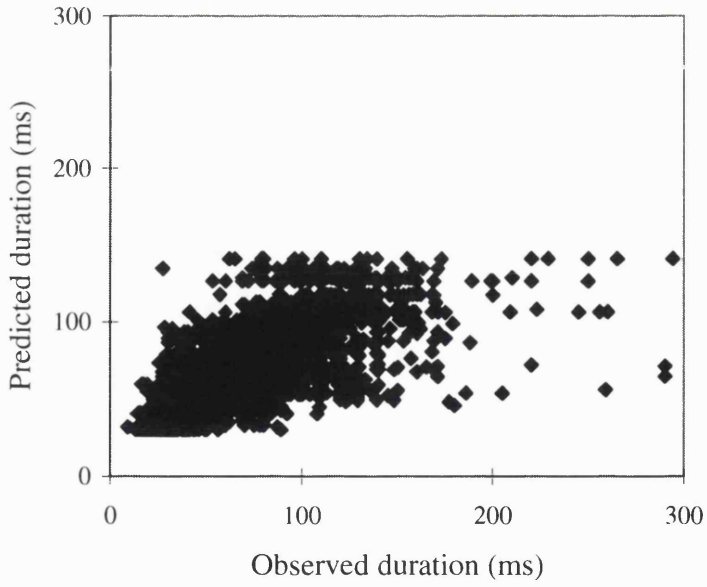


Figure 5-3.
Observed vs. predicted duration for all tested consonants using names and manner feature of the target segment in "Compact feature set" (CART model).



5.1.2 Sums-of-products models

Separate sums-of-products models were built for vowels and consonants because of differences in the feature sets used to describe them. Firstly, the model formula was specified in terms of which factors were to be incorporated and how the parameters associated with each factor were to be combined in the model. Each parameter was then initialised to a value specified by the experimenter. Limits on the allowed range of values were established. A function optimisation strategy was then employed whereby perturbations in the values of the parameters were investigated in terms of their effects on the model performance. An objective measure of sum squared error of prediction was used for this. Two optimisation strategies were used: the downhill simplex method (Press et al., 1992) which seeks to find a single value for the parameters which minimised the objective measure; and the simulated annealing method (Press et al., 1992) which has additional benefits in its ability to avoid sub-optimal solutions[1]. Models were trained on 19,071 vowels and 23,032 consonants in the training data and tested on 4,829 vowels and 5,908 consonants in the test data. The data were exactly the same as used in the CART modelling.

5.1.2.1 Additive models

This is the simplest sums-of-products model where there is one parameter per factor level and all parameters are added together based on the assumption that features operate in an additive manner to predict the duration. The additive model for vowels is shown below.

(5.1)

Model 1 for vowels: "pure additive model"

DUR(id, man, prev, foll, syll, left_pos, right_pos) =

$S_{1,1}(id) + S_{2,1}(man) + S_{3,1}(prev) + S_{4,1}(foll) + S_{5,1}(syll) + S_{6,1}(left\_pos) + S_{7,1}(right\_pos)$

The correlation of the model trained by the downhill simplex method was 0.58 and the RMSE was 39.69 ms. The correlation using the simulated annealing method was 0.61 and the RMSE was 36.89 ms. The parameter values derived for each feature is shown in Table 5-5 for the better simulated annealing method.

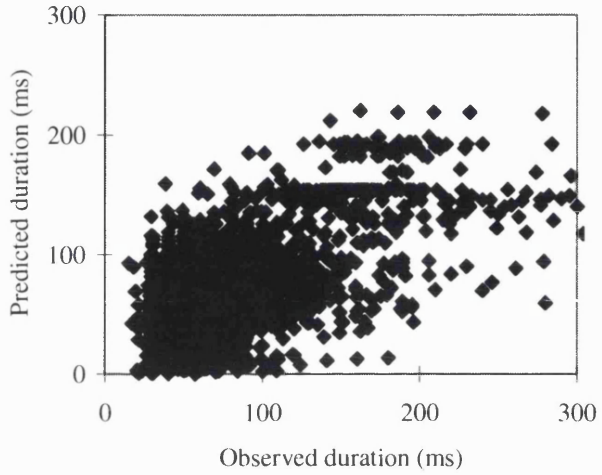
Table 5-5.
Parameters of "pure additive model" for vowels (model 1; simulated annealing method; values marked with '*' are fixed.).

$S_{1,1}$								
id								
aa	ee	ii	oo	uu	vv	wa	we	wi
53	56	40	50	15	64	1	13	1
wv	xi	xx	ya	ye	yo	yu	yv	
36	23	20	53	5	5	67	16	

$S_{2,1}$		$S_{3,1}$		$S_{4,1}$		$S_{5,1}$		$S_{6,1}$		$S_{7,1}$	
man		prev		foll		syll		left_pos		right_pos	
mono*	0	vow*	0	vow*	0	cv	112	utt-init	3	utt-fi	-48
di	44	nas	-74	nas	-9	cvc	98	ip-init	11	ip-fi	-10
		lat	-90	lat	-21	v*	0	ap-init	17	ap-fi	77
		fla	-90	fla	-11	vc	39	pw-init	7	pw-fi	26
		aspstp	-107	aspstp	-22			non-init*	0	non-fi*	0
		aspaff	-101	aspaff	-23						
		tnsstp	-99	tnsstp	-66						
		tnsaff	-146	tnsaff	-123						
		tnsfri	-106	tnsfri	-17						
		laxstp	-94	laxstp	-5						
		laxaff	-103	laxaff	-1						
		laxfri	-104	laxfri	-4						
		pause	-0.40	pause	131						

Figure 5-4 shows a scatter plot of observed and predicted values for all vowels by using this additive model.

Figure 5-4.
Observed vs. predicted duration for test vowels using "pure additive model".



The additive model for consonants was formulated as:

(5.2)

Model 1 for consonants: "pure additive model"

DUR(id, man, prev, foll, syll, syllpo, left_pos, right_pos) =
$S_{1,1}$(id) + $S_{2,1}$(man) + $S_{3,1}$(prev) + $S_{4,1}$(foll) + $S_{5,1}$(syll) + $S_{6,1}$(syllpo) + $S_{7,1}$(left_pos) + $S_{8,1}$(right_pos)

All features described in the data were added together to predict the duration. The correlation by the downhill simplex method was 0.54 and the RMSE was 29.29 ms. The simulated annealing method produced a correlation of 0.51 and an RMSE of 30.08 ms. The parameter values calculated by the downhill simplex method are shown in Table 5-6.

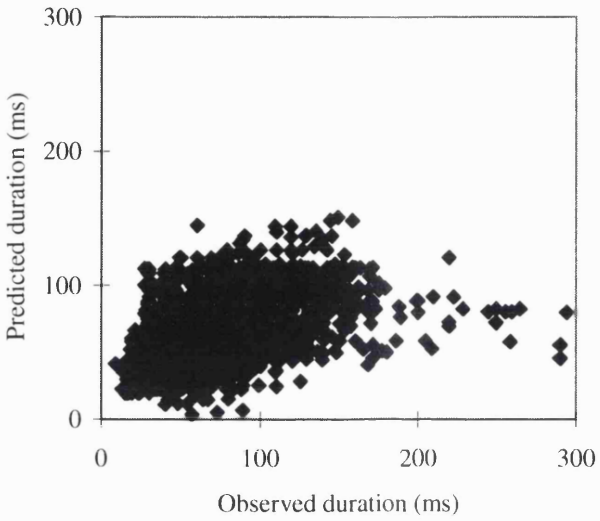
Table 5-6.
Parameters of "pure additive model" for consonants (model 1: downhill simplex method; parameters marked with '*' are fixed).

$S_{1,1}$									
id									
mm	nn	ng	ll	rr	ph	p0	pp	th	t0
51	47	28	69	22	49	11	4	60	17
tt	kh	k0	kk	ch	c0	cc	s0	ss	hh
33	49	21	24	64	41	57	32	67	7

$S_{2,2}$		$S_{3,3}$		$S_{4,4}$		$S_{5,5}$		$S_{6,6}$		$S_{7,7}$		$S_{8,8}$	
man		prev		foll		syll		syllpo		left_pos		right_pos	
Stop *	0	vow*	0	vow*	0	cv	29	on*	0	utt-init	-31	utt-fi	1
Aff	-6	nas	-6	nas	-81	cvc	27	co	86	ip-init	-15	ip-fi	18
Fri	1	lat	8	lat	-83	v*	0			ap-init	35	ap-fi	23
Nas	-31	fla	-180	fla	-148	vc	33			pw-init	8	pw-fi	0
Lat	-44	aspstp	-109	aspstp	-73					non-init*	0	non-fi*	0
Fla	-22	aspaff	-127	aspaff	-71								
		tnsstp	-102	tnsstp	-93								
		tnsaff	-71	tnsaff	-95								
		tnsfri	-89	tnsfri	-102								
		laxstp	-3	laxstp	-62								
		laxaff	-59	laxaff	-59								
		laxfri	-26	laxfri	-73								
		pause	67	pause	-30								

A scatter plot of observed and predicted duration for consonants using this model is shown in Figure 5-5.

Figure 5-5.
Observed vs. predicted duration for all tested consonants using "pure additive model".



The results from "pure additive models" reflect that they were not as good as CART in predicting durations. They failed to predict that longer durations could be found in some situations.

5.1.2.2 Multiplicative models

The next simplest sums-of-products model is one in which all factors combine multiplicatively. This is equivalent to an additive model working in the log domain. The model for vowels is:

(5.3)

Model 2 for vowels: "pure multiplicative model"

DUR(id, man, prev, foll, syll, left_pos, right_pos) =
$S_{1,1}$(id) × $S_{1,2}$(man) × $S_{1,3}$(prev) × $S_{1,4}$(foll) × $S_{1,5}$(syll) × $S_{1,6}$(left_pos) ×
$S_{1,7}$(right_pos)

The multiplicative model for vowels trained by the downhill simplex method gave a correlation of 0.49 and an RMSE of 48.81 ms. The simulated annealing method gave a

correlation of 0.51 and an RMSE of 44.09 ms. These results are significantly worse than the additive model for vowels. The parameter values are shown below:

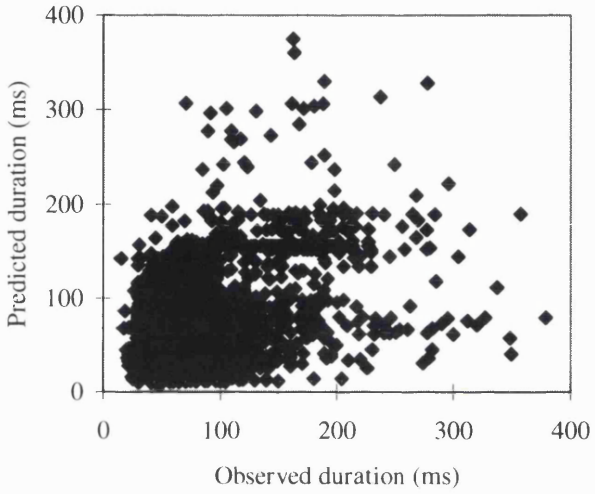
Table 5-7.
Parameters of "pure multiplicative model" for vowels (model 2: simulated annealing method; values marked with '*' are fixed.).

$S_{1,1}$								
id								
aa	ee	ii	Oo	uu	vv	wa	we	wi
35	29	27	28	25	31	87	99	126
wv	xi	xx	Ya	ye	yo	yu	yv	
122	93	25	199	109	115	183	106	

$S_{1,2}$		$S_{1,3}$		$S_{1,4}$		$S_{1,5}$		$S_{1,6}$		$S_{1,7}$	
man		prev		foll		syll		left_pos		right_pos	
mono*	1.00	vow*	1.00	vow*	1.00	cv	0.20	utt-init	0.61	utt-fi	2.93
di	0.31	nas	1.99	nas	1.20	cvc	0.22	ip-init	0.47	ip-fi	3.59
		lat	5.43	lat	0.97	v*	1.00	ap-init	1.02	ap-fi	2.16
		fla	6.33	fla	1.76	vc	0.86	pw-init	1.01	pw-fi	1.23
		aspstp	5.24	aspstp	2.30			non-init*	1.00	non-fi*	1.00
		aspaff	3.78	aspaff	3.83						
		tnsstp	5.40	tnsstp	4.24						
		tnsaff	8.70	tnsaff	5.36						
		tnsfri	5.69	tnsfri	3.57						
		laxstp	5.08	laxstp	2.18						
		laxaff	3.38	laxaff	2.12						
		laxfri	5.13	laxfri	2.41						
		pause	3.05	pause	1.44						

Figure 5-6 is a scatter plot of observed and predicted durations for all vowels using model 2.

Figure 5-6.
Observed vs. predicted duration for all tested vowels using "pure multiplicative model".



The full "purely multiplicative model" for consonants is:

(5.4)

Model 2 for consonants: "pure multiplicative model"

DUR(id, man, prev, foll, syll, syllpo, left_pos, right_pos) =
$S_{1,1}$(id) × $S_{1,2}$(man) × $S_{1,3}$(prev) × $S_{1,4}$(foll) × $S_{1,5}$(syll) × $S_{1,6}$(syllpo) × $S_{1,7}$(left_pos) × $S_{1,8}$(right_pos)

The correlation with fitting by the downhill simplex method was 0.14 and the RMSE was 51.42 ms, which was the poorest result of all models. The simulated annealing method gave a correlation of 0.49 and an RMSE of 31.32 ms.

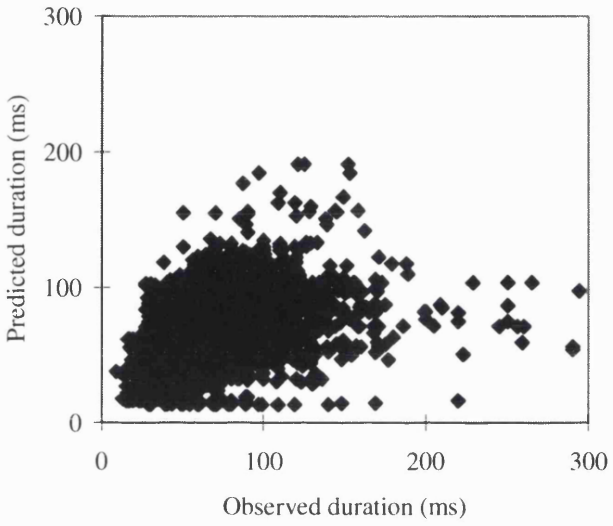
Table 5-8.
Parameters of "pure multiplicative model" for consonants (model 2: downhill simplex method; values marked with '*' are fixed.).

$S_{1,1}$									
id									
mm	nn	ng	Ll	rr	ph	p0	pp	th	t0
32	32	65	49	60	105	87	94	100	61
tt	kh	k0	Kk	ch	c0	cc	s0	ss	hh
97	104	84	115	81	51	89	51	85	72

$S_{1,2}$		$S_{1,3}$		$S_{1,4}$		$S_{1,5}$		$S_{1,6}$		$S_{1,7}$		$S_{1,8}$	
man		prev		foll		syll		syllpo		left_pos		right_pos	
stop*	1.00	vow*	1.00	vow*	1.00	cv	0.19	on*	1.00	utt-int	1.04	utt-fi	6.20
aff	1.44	nas	1.96	nas	3.24	cvc	0.24	co	0.05	ip-init	2.03	ip-fi	2.95
fri	1.71	lat	2.94	lat	3.39	v*	1.00			ap-init	2.33	ap-fi	1.53
nas	5.26	fla	5.01	fla	4.36	vc	2.08			pw-init	1.30	pw-fi	0.96
lat	3.58	aspstp	4.61	aspstp	3.45					non-init*	1.00	non-fi*	1.00
fla	1.43	aspaff	4.68	aspaff	6.20								
		tnsstp	4.71	tnsstp	3.33								
		tnsaff	3.74	tnsaff	4.38								
		tnsfri	4.72	tnsfri	3.96								
		laxstp	3.55	laxstp	3.25								
		laxaff	5.37	laxaff	2.84								
		laxfri	4.79	laxfri	4.13								
		pause	2.11	pause	3.43								

The scatter plot for observed and predicted durations for the consonants by the model 2 is shown in Figure 5-7.

Figure 5-7.
Observed vs. predicted duration for all tested consonants using "pure multiplicative model".



The results from "pure multiplicative models" show that they were slightly better at predicting some of the longer durations found in the data, but still they are not very good either in this respect or overall.

5.1.2.3 Additive-multiplicative models

As van Santen (1997) suggested, models need to be elaborate enough to capture systematic variability in the data, yet use few parameters. He claims that additive-multiplicative models capture the "directional invariance" of the segment duration. He said "directional invariance is the property that, holding all else constant, the effects of a factor have always the same direction". We have not attempted to derive a new additive-multiplicative model for Korean but instead we have adapted Venditti and van Santen's (1998) model of Japanese timing. This is justified because Japanese has a similar linguistic structure to Korean. Because feature classes of this model are different to those we have used for Korean, they were adjusted as follows. Venditti and van Santen (1998) suggested the following sums-of-products model for Japanese vowels:

(5.5)

DUR(id, leng, prev, foll, left_pos, right_pos, acc, syll, spec) =
$S_{1,1}$(id) + [$S_{2,1}$(leng) × $S_{2,2}$(prev)] + [$S_{3,1}$(leng) × $S_{3,2}$(foll)] + $S_{4,1}$(left_pos) + [$S_{5,1}$(leng) × $S_{5,2}$(right_pos)] + [$S_{6,1}$(leng) × $S_{6,2}$(acc)] + [$S_{7,1}$(foll) × $S_{7,2}$(syll)] + $S_{8,1}$(spec)]

where "id" is the identity of the vowel, "leng2 is the length, "prev" the manner of the preceding vowel, "foll" the manner of the following vowel, "left_pos" the syllable distance to the left phrase boundary, "right_pos" the syllable distance to the right phrase boundary, "acc" the accent status of the syllable, "syll" the syllable structure, "spec" the special morphological information. Because the "acc" and "spec" features are not available in the Korean data, these features were ignored. The "leng" feature was substituted by "man" (manner) in the modified model, because the phonemic length does not exist in the modern Korean as explained in chapter 2. Other than the phonemic length, the manner feature of the target segment is believed to be the best candidate to describe the property of the segment. The modified model for the Korean language is thus:

(5.6)

Model 3 for vowels: "additive-multiplicative model"

DUR(id, man, prev, foll, syll, left_pos, right_pos) =
$S_{1,1}$(id) + [$S_{2,1}$(man) × $S_{2,2}$(prev)] + [$S_{3,1}$(man) × $S_{3,2}$(foll)] + [$S_{4,1}$(foll) × $S_{4,2}$(syll)] + $S_{5,1}$(left_pos) + $S_{6,1}$(right_pos)

The correlation of this model by the downhill simplex method was 0.69 and the RMSE was 31.80 ms. The correlation by the simulated annealing method was 0.68 and the RMSE was 32.13 ms. These are the best results for sums-of-products model for vowels. The parameter values for this model are shown in Table 5-9. Ideally, each feature value

in each parameter has different value when it is used in different product terms. For example, in the above equation, $S_{2,1}$(man) and $S_{3,1}$(man) should have different values when they are used in a different product term; so should $S_{3,2}$(foll) and $S_{4,1}$(foll). However, in our models, due to limitations of the modelling program, only one value for each parameter was calculated. This situation was the same in model 3 for consonants.

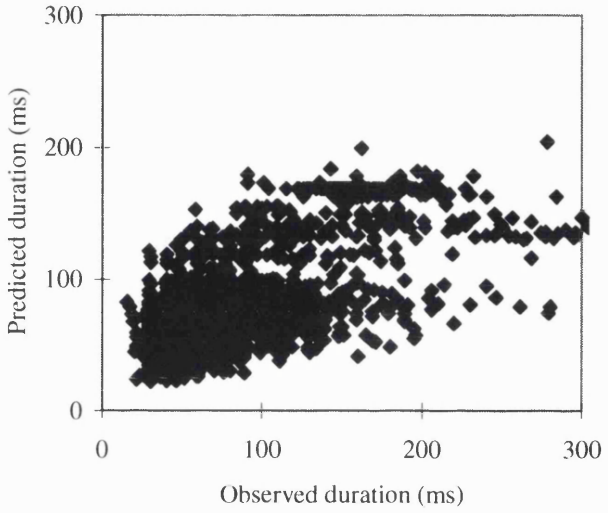
Table 5-9.
Parameters of "additive-multiplicative model" for vowels (model 3; downhill simplex method; values marked with '*' are fixed.).

$S_{1,1}$								
id								
aa	ee	ii	oo	uu	vv	wa	we	wi
61	64	45	60	41	59	71	52	75
wv	xi	xx	ya	ye	yo	yu	yv	
76	104	24	84	29	77	73	62	

$S_{2,1}$;$S_{3,1}$		$S_{2,2}$		$S_{3,2}$;$S_{4,1}$		$S_{4,2}$		$S_{5,1}$		$S_{6,1}$	
man		prev		foll		syll		left_pos		right_pos	
mono*	1.00	vow*	1.00	vow*	1.00	cv	1.15	utt-int	6.15	utt-fi	80.84
di	1.51	nas	3.60	nas	0.07	cvc	0.24	ip-init	1.48	ip-fi	89.90
		lat	4.70	lat	0.83	v*	1.00	ap-init	5.62	ap-fi	69.29
		fla	1.76	fla	0.66	vc	2.99	pw-init	-2.01	pw-fi	11.72
		aspstp	4.63	aspstp	2.45			non-init*	0.00	non-fi*	0.00
		aspaff	1.69	aspaff	6.66						
		tnsstp	3.41	tnsstp	3.56						
		tnsaff	3.22	tnsaff	3.29						
		tnsfri	1.47	tnsfri	1.64						
		laxstp	3.82	laxstp	3.17						
		laxaff	3.07	laxaff	7.13						
		laxfri	1.73	laxfri	3.06						
		pause	3.11	pause	6.54						

The scatter plot of observed and predicted durations by this model is shown in Figure 5-8.

Figure 5-8.
Observed vs. predicted duration for all tested vowels using "additive-multiplicative model".



To obtain an additive-multiplicative model for consonants, we adapted the model used for vowels, having no specific information which might guide an alternative design. The model is:

(5.7)

Model 3 for consonants: "additive-multiplicative model"

$$DUR(id, man, prev, foll, syll, syllpo, left\_pos, right\_pos) = S_{1,1}(id) + [S_{2,1}(man) \times S_{2,2}(prev)] + [S_{3,1}(man) \times S_{3,2}(foll)] + S_{4,1}(syll) + S_{5,1}(syllpo) + [S_{6,1}(man) \times S_{6,2}(left\_pos)] + [S_{7,1}(man) \times S_{7,2}(right\_pos)]$$

where 'syllpo' is the segment position in the syllable, i.e. onset or coda. The correlation was 0.54 both by the downhill simplex method and simulated annealing method. The RMSE was 29.02 ms by the downhill simplex method and 28.86 ms by the simulated annealing method. The parameter values of this model were as follows.

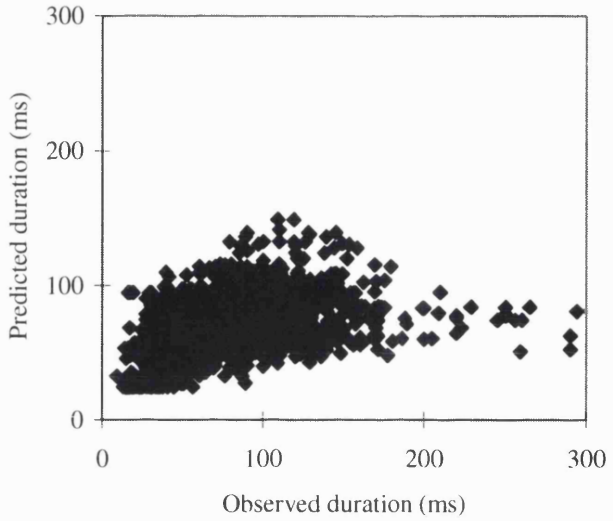
Table 5-10.
Parameters of "additive-multiplicative model" for consonants (model 3; simulated annealing method; values marked with '*' are fixed.)

$S_{1,1}$									
id									
mm	nn	ng	ll	rr	ph	p0	pp	th	t0
47	57	62	64	37	105	58	46	89	58
tt	kh	k0	kk	ch	c0	cc	s0	ss	hh
68	92	64	78	100	63	64	70	107	38

$S_{2,1}$;$S_{3,1}$;$S_{6,1}$;$S_{7,1}$		$S_{2,2}$		$S_{3,2}$		$S_{4,1}$		$S_{5,1}$		$S_{6,2}$		$S_{7,2}$	
man		prev		foll		syll		syllpo		left_pos		right_pos	
vow*	1.00	vow*	1.00	vow*	1.00	cv	-12.91	on*	0.00	utt-int	4.14	utt-fi	3.59
aff	4.18	nas	0.59	nas	2.80	cvc	-15.96	co	1.49	ip-init	4.06	ip-fi	5.81
fri	2.87	lat	3.30	lat	4.64	v*	0.00			ap-init	9.45	ap-fi	6.05
nas	3.18	fla	2.47	fla	2.60	vc	-8.96			pw-init	2.12	pw-fi	1.45
lat	2.20	aspstp	1.47	aspstp	1.50					non-init*	1.00	non-fi*	1.00
fla	0.77	aspaff	3.24	aspaff	0.90								
		tnsstp	2.37	tnsstp	1.25								
		tnsaff	4.67	tnsaff	2.37								
		tnsfri	2.47	tnsfri	6.05								
		laxstp	2.22	laxstp	4.60								
		laxaff	3.88	laxaff	3.04								
		laxfri	1.04	laxfri	1.82								
		pause	4.40	pause	5.21								

These results were the best among the sums-of-products models for consonants. The observed and predicted values by this model are illustrated as a scatter plot in Figure 5-9.

Figure 5-9
Observed vs. predicted duration for all tested consonants using "additive-multiplicative model".



The additive-multiplicative models still failed to predict longer durations and the performance was worse than that obtained using CART modelling.

A summary of the performance results of all of the above models are illustrated as follows.

Table 5-11.
Performance results summary for vowels using sums-of-products models.

	Downhill simplex		Simulated annealing	
	RMSE	Correlation	RMSE	Correlation
Model 1	39.69 ms	0.58	36.89 ms	0.61
Model 2	48.81 ms	0.49	44.09 ms	0.51
Model 3	31.80 ms	0.69	32.13 ms	0.68

Table 5-12.
Performance results summary for consonants using sums-of-products models.

	Downhill simplex		Simulated annealing	
	RMSE	Correlation	RMSE	Correlation
Model 1	29.29 ms	0.54	30.08 ms	0.51
Model 2	51.42 ms	0.14	31.32 ms	0.49
Model 3	29.02 ms	0.54	28.86 ms	0.54

5.1.3 Summary of Experiment I

The results of Experiment I with the Compact feature set showed that the CART decision tree models had overall better performance than the sums-of-products models. Peformance was best when the segment names and manner features were used in a CART model. However, once names were used, the manner of the target segment barely contributed to the performance, though the greatest effect of segment name came from identifying the manner of segment. CART models showed that the syllable distance to the right phrase boundary (right_pos) was most important feature in predicting vowel durations. In consonant duration prediction, the syllable distance to the left phrase boundary (left_pos) was most important. In both vowels and consonants, the following segments (foll) were more influential than preceding segments (prev) except when z-scores were used in vowel duration modelling. Though this result agrees with many experimental results on English, it contradicts findings in Korean, where researchers have argued that preceding segments are more influential than following segments. The dominance of preceding segments was only found when z-scores were used in vowel duration modelling. This will be investigated further in the next experiment where the "Binary feature set" is used. Among sums-of-products models, the additive-multiplicative models were better than the "pure additive models" or the "pure multiplicative models". This might be evidence for presence of interactions between

factors that could not be modelled by uniform addition or multiplication. Further investigation of these interactions is clearly required.

5.2 Experiment II: "Binary Feature Set"

The results of experiment I told us which main factors are important for predicting durations of vowels and consonants. However these models do not give us much information about the relative importance of individual feature settings. Though we can see the coefficients allocated to different feature levels, we cannot see the relative importance of the levels, for example, the relative importance of the individual values of the "left_pos" or "right_pos" factors. Thus a different kind of investigation is necessary to explore the importance of individual levels of each factor. In order to achieve this goal, the feature set was extended, so that each n-ary feature was replaced with a number of binary features. The format of this feature analysis was illustrated in chapter 4. A total of 69 features were available in the data set. A stepwise CART decision tree model was used for the investigation. The procedure for this experiment is similar to that of 5.1.1, except for the feature set.

5.2.1 CART analysis using segment names and class features

A stepwise CART model was trained on 19,071 vowel tokens and 23,032 consonant tokens in the training data and pruned on 4,785 vowels and 5,824 consonants in the evaluation data before being tested on 4,829 vowels and 5,908 consonants in the test data. The feature set described the name and major class features of each target segment and the contextual environment of surrounding segments along with phrase structure and syllable structure. The stepwise procedure ended when additional features made no significant improvement in performance. The model was then 'pruned' by removing

questions and pooling leaf nodes so that the performance of the tree on the evaluation data set was maximised. The tree was pruned back to 35 features in this process. The resulting tree was then tested on the test data. The correlation coefficient between the observed and predicted durations and the root mean squared prediction error (RMSE) were calculated. This performance result is illustrated in Table 5-14. The correlation of this tree was 0.77 and its RMSE was 25.11 ms. The observed and predicted values by this model are illustrated as a scatter plot in Figure 5-10. The tree also showed which individual features have the most important role in predicting duration. The ranking of the ten most important features is shown in Table 5-15. In this model where both the name and the major class features of each target segment are available, the single most important feature in the model was unsurprisingly the name of the target segment. Other important features were the prosodic phrase features and the syllable structure features. The second most important feature, AP-final position feature (1_AP) had a large effect, followed by the AP-initial position feature (AP_1), onset position feature (ON), CVC syllable structure feature (CVC), and preceding voicing feature. Subsequent features had much less effect. The change in correlation coefficient resulting from the stepwise addition of each of the 35 features can be found in Appendix 3.

5.2.2 CART analysis using segment names

To determine whether the manner of the target segment has influence on the duration when used in combination with the name of the target segment, a model was built without the manner feature. A stepwise CART model was trained and tested with just the name of each target segment and 61 segmental and prosodic phrasal features describing the context. The fitted tree had 34 features. The correlation was 0.76 (compared 0.77 in section 5.2.1) and the RMSE was 25.25 ms (compared 25.11 ms in

section 5.2.1). The 10 most influential factors in this tree were exactly same as they were in 5.2.1 (See Table 5-15). This shows that the manner features are redundant in the presence of segment names—they provided no useful additional grouping in the building of the decision tree. The feature rankings and the growth in the correlation for each stepwise refinement of this CART model was very similar to the model in section 5.2.1 (See Appendix 3).

5.2.3 CART analysis using segment class features

In this stepwise CART model, the target segment name was dropped leaving only the class features of the target segment to identify the segment. The idea was to force the model to make generalisations across segment types. In total, 69 segmental and prosodic phrase features were used to describe the contextual information. The fitted tree had 42 features. As shown in Table 5-14, the correlation coefficient was 0.72 and the RMSE was 27.12, which were a little worse than the two previous models when target segment names were incorporated in the decision tree. Though it could be thought that major class features of the target segment should have had a significant role in predicting durations when the names of the segments were not used, only the "fla (flap)" feature occupied a place in the 10 most important factor in this decision tree as shown in Table 5-15. The prosodic phrase features such as "1_AP", "AP_1", and "PW_1 (PW-initial position feature)" and the syllable structure feature such as "ON", "CVC", and "V" still played the most important roles in the duration prediction.

5.2.4 CART analysis using z-scores of segments

In this model, the linear ms duration of each segment was substituted by z-scores as was described in 5.1.1.4. Since z-scores encode the inherent properties of each target

segment type, the names and major class features of the segment were not incorporated into the decision tree model. This left 61 segmental context and prosodic phrase features in the data set. After fitting, the tree contained 40 features. The correlation between observed and predicted durations was 0.74 and the RMSE of prediction was 26.44 ms as in Table 5-14. Once again, the prosodic phrase features dominated the 10 most influential features. In this analysis, such features as "1_AP", "AP_1", "PW_1", and "1_PW" were among the 10 most important factors in predicting duration as in Table 5-15. The growth in the correlation coefficients for each stepwise refinement of this CART model is shown in Appendix 3.

Based on this CART decision tree, the mean z-score changes arising from each selected feature acting on its own were calculated. We call this "mean feature effect" analysis. The objective of this analysis is to obtain from the CART tree information about the relative size of the effect of each feature on the segment duration. We know from the stepwise building of the tree which features were most important and in which order they were applied in the tree from root towards the leaf nodes. We can use this information to re-analyse the training data to establish the mean effect of each feature. The procedure is as follows: firstly the data is partitioned into two groups according to the value of the most important feature (here, 1_AP) and the means of each partition are calculated (1_AP=0 : -0.12, and 1_AP=1 : 0.87, values in z-scores). The difference between these means (0.99) is called the mean effect of feature 1_AP. Next the mean duration value of each partition is then subtracted from the individual segment durations in that partition. In effect this "takes into account" the mean operation of feature 1_AP. The data can then be partitioned according to the value of the second most important feature in the CART analysis (here, ON). This gives us two further means (ON=0 :

0.035, ON=1 : -0.055) and the mean effect of feature ON (-0.09). The mean values from the two partitions can be subtracted as before to take into account feature ON, and the process repeated for the third most important feature and so on. The top 10 changes are given in Table 5-13 and the complete list is shown in Appendix 4.

Table 5-13.
Mean feature effect caused by selected features in the training data.

Ranking	Feature	Partition ∅		Partition 1		Diff
		Mean	Size	Mean	Size	
1	1_AP	-0.12	37170	0.87	4933	0.99
2	ON	0.04	25704	-0.06	16399	-0.09
3	AP_1	-0.06	37041	0.47	5062	0.53
4	nas_	0.03	34478	-0.12	7625	-0.15
5	_nas	0.07	34467	-0.30	7636	-0.36
6	PW_1	-0.03	28198	0.06	13905	0.09
7	vce_	-0.12	13783	0.06	28320	0.18
8	1_PW	-0.06	27915	0.11	14188	0.16
9	CVC	0.05	24978	-0.08	17125	-0.13
10	cor_	-0.03	22082	0.03	20021	0.06

Partition ∅ = mean and size of partition when feature is ∅.
Partition 1 = means and size of partition when feature is 1.

Mean effect analysis gives an overall picture of the effect of the most important features, but it doesn't accurately reflect the actual operation of the tree, since it ignores interactions between features. Thus it could be that feature ON has a very different effect in AP_1 positions than elsewhere. However we have found no evidence of strong interactions in the top 10 most important features. In this table, a positive mean feature effect in z-score corresponds to a lengthening effect of duration and a negative z-score is a shortening effect of duration. When the segment is in AP-final position (1_AP), the segment has the positive mean feature effect of 0.99, so it has a large lengthening effect. Also in this table, the AP-initial position feature (AP_1), the PW-initial position feature (PW_1), the PW-final position feature, the preceding voicing feature (vce_), and the

preceding coronal feature (cor_) had lengthening effects. On the other hand, the onset position feature (ON), the preceding nasal feature (nas_), the following nasal feature (_nas), and the CVC syllable structure feature (CVC) had shortening effects. The full table in Appendix 4 also shows that the 1_IP features does not have a large effect once the 1_AP feature has been applied.

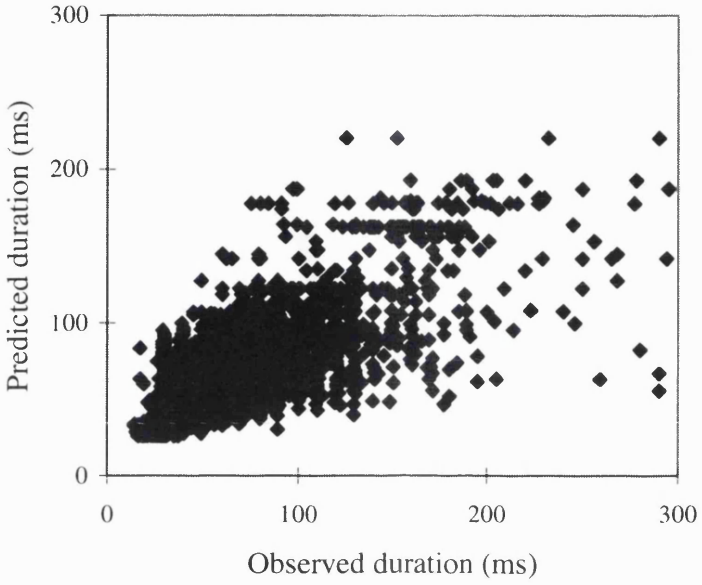
Table 5-14.
CART performance results summary from Experiment II.

	RMSE	Correlation
Name & manner	25.11 ms	0.77
Name only	25.25 ms	0.76
Manner only	27.12 ms	0.72
z-score	26.44 ms	0.74

Table 5-15.
Rankings of 10 most important factors in the CART decision trees from Experiment II.

	1	2	3	4	5	6	7	8	9	10
Name & manner	name	1_AP	AP_1	ON	CVC	vce_	_V	1_PW	nas_	VC
Name only	name	1_AP	AP_1	ON	CVC	vce_	_V	1_PW	nas_	VC
Manner only	1_AP	ON	AP_1	CVC	V_	nas_	1_PW	fla	V	_hiV
z-score	1_AP	ON	AP_1	nas_	_nas	PW_1	vce_	1_PW	CVC	cor_

Figure 5-10.
Observed vs. predicted duration for all tested segments using names and manner of the target segment in the "Binary feature set" (CART model).



5.2.5 Summary of Experiment II

As expected from the result of Experiment I, the performance result was best when names and manner of the target segment was used. However, once names were used, the manner of the target segment barely contributed to the performance. Though the model was still not good at predicting longer durations, the performance was better than Experiment I where "Compact feature set" was used. The results also showed that the prosodic phrase boundaries had major effect in duration modelling, followed by the effect of surrounding segments. We look at the results in more detail in the next section.

5.3 Analysis of Models

5.3.1 Performance

The prediction error and correlation coefficients of the CART models were comparable with the best published results in Korean (Lee and Oh, 1999) as in Table 5-16. In their

CART modelling of spoken Korean on segmental duration, Lee and Oh (1999) trained on 240 sentences (15,037 segments) and tested on 160 sentences (9,494 segments). Their RMSE was 22 ms, and the correlation coefficient was 0.82. They used the segment names of surrounding segments and of the observed segment in question, the part-of-speech features of the word, the position features of the segment in the prosodic phrase, and the length of the prosodic phrases in syllables. In another regression tree modelling of spoken Korean using 15 sentences by three male and four female speakers in three different tempos, Lee (1996) showed correlations between 0.74 and 0.69 and an RMSE of less than 25 ms. Though all results were based on regression tree models, the details of the statistical analyses were slightly different across experiments.

Table 5-16.
Comparisons between best CART model from this experiment and other results.

Experiment	Correlation
This experiment	0.77
Lee & Oh (1999)	0.82
Lee (1996)	0.74

5.3.2 Linguistic interpretation

In section 2.2.6, the following issues were highlighted as unsolved or controversial in the research of the timing of spoken Korean.

(5.8)

a. Which type of phrase boundary has the most influence? UTT, IP, AP, PW boundaries were all claimed to have lengthening effects. However, more information was needed over which boundary is more important, the relative size of initial and final boundary effects, and whether syllables in post-initial or penultimate positions are also lengthened.

b. How does the structure of a syllable affect its constituents? In Korean, CVC, VC, V, and CV syllable structures can be observed. These structures are believed to have an influence on the segment duration. More information is required about how each syllable structure affects segment duration. The different behaviours of onset consonants and coda consonants also needs further study.

c. Which segmental features show a systematic effect on duration? In English, following segments have more influence than preceding segments. In Korean, it is claimed that preceding segments are more important than following segments.

Experiment II investigated the importance of each phonological feature in duration prediction and gave some answers about the above issues. Among phrase boundary features, the AP boundary had the most influence either to AP-initial or to AP-final syllables. The AP boundary significantly lengthened final segment duration. Though both the PW-initial position and the PW-final position were important in duration prediction, the PW-final position feature had more lengthening effect. UTT boundaries and IP boundaries did not contribute much to the duration once the AP boundary had been taken into account. This is believed to be partly because each UTT boundary and IP boundary is also an AP boundary and a PW boundary in the phonetic transcriptions. It is interesting that shortening effects were seen in all post-initial positions and in penultimate positions from boundaries except in post-initial position of AP. It shows that in Korean, the lengthening effect of the phrase boundary does not penetrate into the syllables in these positions.

These results can be compared to previous analyses of the timing pattern of spoken Korean. Han (1964) and Kim (1974) found that a vowel in sentence-final position is longer than in other positions. Lee and Koo (1997) found that the syllable before a sentence boundary was longest, and at normal speed, the syllables at IP, AP, and PW boundaries had similar duration. On the other hand, Chung et al. (1997), Jun (1993), and Lee (1990) argued that when an IP was followed by a pause, the IP-final position had a greater lengthening effect than did the AP-final position. It is possible that because their data was restricted to constrained carrier phrase sentences, they failed to find a generalisation in the duration pattern.

The CART analysis did not find that syllable structure had a general effect on vowel duration except in the case of CVC. Though certain types of adjacent consonants affected vowel duration, there was no general effect on the duration that could be attributed to the structure of the syllable without consideration of consonant type. In the CART models, CVC syllables had a small shortening effect with a mean change in z-score, -0.13. In contrast, Han (1964) and Koo (1998) found that vowels in a CVC syllable structure were much shorter than those in CV or V syllable structures.

In terms of the effects by preceding segments and following segments, the preceding nasal feature (nas_), the following nasal feature (_nas), and the preceding voicing feature (vce_) had the most influence in the CART models in Experiment II. Although sonorants are generally thought to have a lengthening effect, this is instead evidence of segment shortening both before and after nasal consonants. This is in partial agreement with Lee (1996) for Korean, where consonant shortening after nasals was observed; and also with Lehiste (1970) for English, where shorter vowels before nasals were seen. A

convincing explanation for this effect is yet to be proposed. Although other influences of surrounding segments did not contribute as much to the performance of the CART decision tree, the preceding aspiration feature (-0.50) and the fricative feature (-0.25) significantly shortened the following segment. Because both features share the [spread glottis] feature, the opening of the glottis could be the major cause of shortened following vowels as suggested by Kim (1974). Because the [spread glottis] feature cannot be observed in the onset of stops, the lengthening effect of aspirated consonants is much bigger in preceding aspirated stops than in following aspirated stops. This effect can be also observed in the significant shortening by the "following glottal feature (-0.41)". The glottal [h] in Korean also involves a wide opening of the glottis. Preceding tense stops have [constricted glottis] feature and also significantly shortened the following vowels (-0.35). Based on these observations, it can be concluded glottal opening is a major controller in segment durations. In agreement with previous studies of English and Korean, the models showed little effect caused by the place feature of surrounding segments. However, in contrast to other studies, the models did not show any significant effect of voicing either.

5.4 Summary

This chapter showed that CART analysis and sums-of-products models can be used to address linguistic issues such as the boundary effect, the syllable structure effect, and the effect of surrounding segments on the segment duration. In addition, these models also provide duration predictions which can be used in the prosody component of TTS systems. The best performance results from CART models were similar to other studies on Korean segment duration. Future studies might examine the internals of the CART decision trees to find interactions among features, and this might feed into better sums-

of-products models. The results from sums-of-products models for Korean in this thesis were worse than those for English or Japanese found in other studies. The calculation of the values for the same feature in different product terms should be pursued to improve the performance of sums-of-products models.

NOTES

[1] Mark Huckvale incorporated these mathematical methods into a computer program for the sums-of-products modelling.

6. PERCEPTUAL EVALUATION

Perceptual evaluation is essential to decide the quality of synthesised speech. It is not always the case that improved statistical modelling leads to improved speech quality. This chapter is divided into two parts. The first part explains the details of a Korean speech signal generation system. The second part describes a perceptual evaluation that was carried out using this system in combination with duration models developed in the experiments. Perceptual evaluation investigates the clarity and the listener preference for durations calculated by the best CART and sums-of-products models and durations calculated by a commercial Korean TTS system.

6.1 Hanmal Korean Language Diphone Database (HN 1.0)

In order to create synthetic speech manipulated by a temporal model and to evaluate its perceptual quality, a new Korean language diphone database "Hanmal (HN 1.0)"[1] was developed based on the MBROLA synthesis system (Dutoit et al., 1996) in collaboration with Professor Gyeongseog Gim. This diphone database has been publicly available since September 17, 1999 from the MBROLA web site so that other researchers could synthesise Korean speech and investigate the relationships between prosody variation and naturalness[2].

MBROLA is a speech synthesis system based on the concatenation of diphones. It takes a list of phones as input, together with prosodic information (duration of phones and a piecewise linear description of pitch), and produces speech signals, at the sampling frequency of the diphone database used. Dutoit et al. (1996) point out that the ability of concatenative synthesisers to produce high quality speech is dependent on the type of

segments chosen and the model of speech signal to which the analysis and synthesis algorithms refer. The design should be able to account for as many co-articulatory effects between segments as possible. Given the restricted smoothing capabilities of the concatenation technique, they should be easily connectable.

6.1.1 Creating a text corpus

Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. Their main usefulness in synthesis is that they minimise concatenation problems, since they contain most of the transitions and co-articulations between phones. They also require relatively small amounts of memory, as their number remains small (compared to synthesis units such as half-syllables or triphones).

The first step in building a diphone database is to generate a list of all the phones of the language. Notice that phones are acoustic instances of phonemes. To obtain a list of phones from a list of phonemes requires the investigation of which acoustic versions of phonemes differ significantly due to co-articulation. Although it is not necessary to account for all allophonic variations to build an intelligible synthesiser, the naturalness of synthetic speech may be affected if too few allophones are considered. When a complete list of phones has emerged, a corresponding list of diphones is readily obtained, and a list of words can be constructed such that each diphone appears at least once.

To prepare a diphone database capable of satisfying these requirements for Korean, 1,986 nonsense words were created to cover a catalogue of 1,986 diphones. In order to make the database acceptable to the general public, the MBROLA project team asked us

to use the SAMPA (Speech Assessment Methods Phonetic Alphabet) transcription convention[3]. Table 6-1 and Table 6-2 list the consonants used in the diphone database in IPA and SAMPA notation.

Table 6-1.
SAMPA notation and descriptions of onset consonants.

IPA	SAMPA	Description
k	k	velar lax plosive, voiceless
k'	k_>	tense velar plosive
n	n	alveolar nasal
t	t	alveolar lax plosive, voiceless
t'	t_>	tense alveolar plosive
ɾ	4	alveolar tap
m	m	bilabial nasal
p	p	bilabial lax plosive, voiceless
p'	p_>	tense bilabial plosive
s	s	alveolar fricative
s'	s_>	tense alveolar fricative
ŋ		not assigned
ts	ts\	alveolo-palatal lax affricate, voiceless
ts'	ts_>	postalveolar tense affricate
ts^h	ts_h	postalveolar aspirated affricate
k^h	k_h	velar aspirated plosive, voiceless
t^h	t_h	alveolar aspirated plosive
p^h	p_h	bilabial aspirated plosive
h	h	glottal fricative, voiceless
g	g	velar plosive, voiced
d	d	alveolar plosive, voiced
b	b	bilabial plosive, voiced
dz	dz\	postalveolar affricate, voiced
ɕ	s\	alveolo-palatal fricative
l	l	alveolar lateral

Table 6-2.
SAMPA notation and descriptions of coda consonants.

IPA	SAMPA	Description
k̚	k_}	velar plosive, voiceless, no audible release
n	n_}	alveolar nasal
t̚	t_}	alveolar plosive, voiceless, no audible release
l	l_}	alveolar lateral
m	m_}	bilabial nasal
p̚	p_}	bilabial plosive, voiceless, no audible release
ŋ	N	velar nasal

The consonants were grouped into 19 onset consonants and 7 coda consonants, because the developers believed that Korean listeners were likely to be sensitive to unreleased consonants occurring in coda position. The position of consonants in the syllable is determined based on the phonetic form of the utterance. So when the syllable final consonant is not resyllabified, then it is in the coda position; when it is resyllabified, it should be in the onset position. In order to distinguish coda consonants from syllable onset consonants, the unreleased stop diacritic "_}" was appended to coda consonants "k", "n", "t", "m", "l" and "p". Allophonic variants of consonants were then established as a function of their segmental context. For instance, every lax obstruent stop and affricate was matched with its voiced counterpart. The lax velar stop has two allophones in the onset position: voiceless "k" and voiced "g". If the segment follows a voiced segment, it becomes voiced. In the coda position, it becomes "k_}". The alveolar stop has "t" and "d" in the onset position, "t_}" in the coda position. The bilabial stop has "p", "b" and "p_}". The lax alveopalatal affricate also has two allophones: "ts\" and "dz\" in the onset position, but in the coda position they are neutralised to "t_}". The lax alveolar fricative has two allophones in onset position: "s\" before a high vowel and "s" otherwise. Among obstruents, tense unaspirated stops, tense aspirated stops and

fricatives are all neutralised in the coda position. Alveolar/palatal obstruents "ts_h", "ts_>", "t", "t_>", "s_>", and "s" are neutralised to "t_}"; velar obstruents "k" and "k_}" are neutralised to "k_}"; bilabial obstruents "p" and "p_>" are neutralised to "p_}"; and the glottal fricative "h" is neutralised to "t_}". None of these obstruents have voiced equivalents. Among sonorants, "n", "l", and "m" appear in syllable initial position. "l" has an allophone "4" when it appears in intervocalic position. Though phonologically, "N" can appear in the syllable initial position, it rarely appears in that position. So "N" was put in the coda position. In the coda position, sonorants can be "n_}", "l_}", "m_}" and "N".

Table 6-3.
SAMPA notation and descriptions of vowels.

IPA	SAMPA	Description
a	a	open front unrounded, Cardinal 4
ɛ	E	open-mid front unrounded, Cardinal 3
ja	ja	palatal approximant + open front unrounded
jɛ	jE	palatal approximant + open-mid front unrounded
ʌ	V	open-mid back unrounded
e	e	close-mid front unrounded, Cardinal 2
jʌ	jV	palatal approximant + open-mid back unrounded
je	je	palatal approximant + close-mid front unrounded
o	o	close-mid back rounded, Cardinal 7
wa	wa	voiced labial-velar approximant + open front unrounded
wɛ	wE	voiced labial-velar approximant + open-mid front unrounded
ø	2	close-mid front rounded
jo	jo	palatal approximant + close-mid back rounded
u	u	close back rounded, Cardinal 8
wʌ	wV	voiced labial-velar approximant + open-mid back unrounded
we	we	voiced labial-velar approximant + close-mid front unrounded
wi	wi	voiced labial-velar approximant + close front unrounded
ju	ju	palatal approximant + close back rounded
ɯ	M	close back unrounded
ɯi	M\i	velar approximant + close front unrounded
i	i	close front unrounded, Cardinal 1

Table 6-3 lists the vowels used in the diphone database in IPA and SAMPA notation. Korean vowels consist of 9 monophthongs and 12 diphthongs. Each diphthong was treated as a unitary segment in the diphone database. Because there are no significant variations of vowel realisation in context, no allophonic variants of vowels were considered.

From this list of segments, 12 groups of nonsense words were constructed to define all the available diphone contexts. Group 1 covers all the voiced syllable onset consonants in combination with following vowels. Group 2 covers all vowel to vowel combinations, while Group 3 all vowel and coda consonant combinations, and Group 4 all vowel and pause combinations. Other groups covered coda consonant and onset consonant combinations, vowel and onset consonant combinations, syllable coda consonant and pause combinations, pause and onset consonant combinations, pause and vowel combinations, voiceless onset consonant and vowel combinations, coda and vowel combinations, and pause alone. A list of the groups and their sizes is shown in Table 6-4.

Table 6-4.
Diphone groups in contexts.
Onset in Group 1 is "the voiced onset + nucleus" combination.
Onset in Group 10 is "the voiceless onset + nucleus" combination.

Group	Combination	Number
Group 1	onset + nucleus	378
Group 2	nucleus + nucleus	441
Group 3	nucleus + coda	147
Group 4	nucleus + pause	21
Group 5	coda + onset	133
Group 6	nucleus + onset	399
Group 7	coda + pause	7
Group 8	pause + onset	18
Group 9	pause + nucleus	21
Group 10	onset + nucleus	399
Group 11	coda + nucleus	21
Group 12	pause + pause	1
Total number of diphones used		1,986

6.1.2 Recording the corpus

The speaker was a speaker of standard Korean who had lived in Seoul for 32 years before coming to the UK. He had been away from Korea, studying in the UK, for 3 years before the recording. The recordings were made four times in an anechoic chamber on digital tape using 2 channels at 44,100 samples/sec/channel. Channel 1 was the speech signal from microphone, channel 2 was a Laryngograph signal. They were resampled to 16 kHz and transferred to disk. In order for the MBROLA resynthesis operation to achieve best results, the corpus was read with a monotonous intonation. The speaker was also requested to keep the pitch and rhythm consistent across phrases. This consistency aids in the production of smooth segment concatenation. However, in order to improve the naturalness of speech made from the diphone database, the speaker was requested to read each nonsense phrase rapidly and fluently. In order to avoid any

vocal fry in the diphone database, a neutral vowel /ʌ/ was inserted before the target words except for those starting with a pause or a voiceless consonant.

6.1.3 Segmenting the corpus

The Speech Filing System (SFS)[4] was used to analyse and annotate the speech data. The segmentation was decided with reference to three signals: waveform, spectrogram, and Laryngograph signal (Lx). Three boundary points were identified: the mid-point of each target segment and the boundary between the two target segments. Annotations were stored as sample numbers in a database and then exported in a text file for diphone processing. They look like the following.

(6.1)

a.	a-a.d16	a	a	4526	7374	5844
b.	a-ae.d16	a	E	5148	7757	6306
c.	a-b.d16	a	b	3741	5334	4868
d.	a-bb.d16	a	p'	2874	4971	3619
e.	a-bc.d16	a	p_}	4274	6918	5346
f.	a-ch.d16	a	ts_h	2342	4443	3062

In (6.1), *.d16 refers to the speech signal data filename. Segments in the second and third columns are the target diphones. The fourth column is the starting sample number of the diphone and the next column is the end point of the diphone. The last column indicates the mid point of the diphone, that is, the boundary between two target segments.

6.1.4 MBROLA program

The diphone recordings were processed by the MBROLA team in Belgium to produce the Hanmal diphone database. Applications based on this database are supported on a

wide range of computing platforms using the MBROLA signal generation engine. Diphone concatenation and prosody manipulation can be performed using the MBR-PSOLA algorithm (Dutoit et al., 1996). This method is an interesting alternative to purely time-domain PSOLA, in the context of a multi-lingual TTS system, for which the ability to derive segment databases automatically, to store them in a compact way, and to synthesise high quality speech with a minimum number of operations per sample is of considerable interest. The format of the control data input to the MBROLA application is as follows. The target word is "kan_}da (to go)".

(6.2)

```
_       100
k        35
a        79 20 140 50 135 80 135
n_}     120
d        70
a       150 20 135 50 140 80 135
_       100
```

In (6.2), "_" stands for the pause. The second column of each row represents the duration of the target segment in milliseconds. The other columns describe the pitch contour for the segment in pairs of numbers: the first value in the pair is the percentage position through the segment, the second value is the fundamental frequency in hertz. Pitch values are linearly interpolated inside and across segments. The input transcription needs to be fully specified for allophonic variants. For example, for the input /halapʌtsi/ (grandfather)" the file contains "_ h a 4 a b V dz\ i _" not "_ h a l a p V ts\ i _".

6.2 Perceptual Evaluation

6.2.1 Test sentences

Nine sentences with various length were selected from broadcast news scripts, which were different from the data set used in the experiment. The list of sentences was as follows:

(6.3)

a. /ulinɯn mintsoktsuŋhɯŋɰi jʌksatsʌk samjʌŋɯl t'iko it'aŋe t^{h}ɛʌnas'ta/
"We were born in this country with a duty to promote national prosperity."

b. /palamkwa hɛsnimi sʌlo himi tʌ setako tat^{h}uko is'ʌs'sɯpnita/
"The wind and the rain were competing with each other to test their power."

c. /onɯl halumanto tsʌnnamkwa kjʌŋnam nɛljuktsipaŋɯlonɯn pɛkmillimit^{h}ʌka nʌmnɯn manhɯn pika nɛljʌs'sɯpnita/
"Today also there was heavy rain – over 100 mm in the inner area of Jeonnamand Gyeongnam counties."

d. /pʌls'ʌ sahɯlts'ɛ nampu tsipaŋɰi hounɯn kjesoktweko is'sɯpnita/
"It has already been three days since heavy rain started in the southern area."

e. /tsikɯmto jʌtsʌnhi kjʌŋnamkwa tsʌnnam namhɛan tsipaŋɯlonɯn houkjʌŋpoka nɛljʌtsj is'ko tsʌnpukkwa tsʌnnam nɛljuktsipaŋɯlonɯn houtsuɰipoka nɛljʌtsin kaunte onɯlpam saieto ts^{h}ʌntuŋpʌnkɛka ts^{h}imjʌnsʌ tsiptsuŋhouka naɛlil kanɯŋsʌŋi nopsɯpnita/
"There are still warnings of severe rain in the southern coastal areas of Gyeongnam and Jeonnam counties and forecasts of moderately heavy rain in the inner area of Jeonbuk and Jeonnam counties, and there is a high chance of thunder and local heavy rain tonight."

f. /sʌul sikatsiɰi kjot^{h}oŋtsʌŋpo ipnita/
"This is the traffic bulletin for central Seoul."

g. /nampu sunhwantolonɯn sataŋsakʌlɰi koŋsa t'ɛmune jaŋpaŋhjaŋɰi kjot^{h}oŋhɯlɯmi motu ʌljʌpsɯpnita/
"There is heavy congestion for both outbound and inbound traffic on the southern circular road due to road works at the Sadang intersection."

h. /namt^{h}ɛljʌŋ kokɛesʌ isu kjots^{h}alo paŋhjaŋɯlonɯn hwamults^{h}aɰi tsɯŋkalo sisok isipkilloɰi soktolɯl nɛl su is'ko jesulɰi tsʌntaŋesʌ poŋts^{h}ʌn sakʌli paŋhjaŋɯlonɯn tsits^{h}eka kjesoktweko is'sɯpnita/

"The speed of traffic in the area from the Namtaereong Hill to the Isu intersection is just 20 km per hour and delays are continuing in the area from the Festival Hall to the Bongcheon intersection."

i. /ollimp^{h}ik tɛlonɯn jʌɰitoesʌ kimp^{h}okoŋhaŋ paŋhjaŋi sisok sipkillo tsʌŋtoɰi kʌpuki kʌlʌmɯl kjesokhako is'sɯpnita/
"In the Olympic Grand Road, traffic is only moving at 10 km per hour in the direction from Yeoido to Kimpo Airport."

These sentences were chosen because TTS demos were available for comparison from an ETRI (Korea Electronics and Telecommunications Research Institute) TTS system web site. Currently only three passages of TTS demo are available in the web site. Durations were calculated by using the best CART model and sums-of-products model. For CART modelling, both the name and manner features of segments in the "binary feature set" were used. For the sums-of-products model, the additive-multiplicative model by simulated annealing was adopted both for vowels and consonants. To compare the quality of the duration modelling with a commercial Korean TTS systems, durations were also extracted from the ETRI TTS demonstration system. The ETRI TTS system is known to be one of the best Korean language TTS systems in Korea. In the experiments, the CART model was named model 1 and the sums-of-products model 2. The model by ETRI was named model 3. F_0 contours for the sentences were copied from natural read versions. The same F_0 contours were used for all systems. The duration and F_0 contour information of these models were then applied to the MBROLA Korean diphone data "Hanmal". The synthesised speech by the three models were played to subjects for perceptual evaluation.

(6.4)

a. Model 1: CART decision tree model using names and major class features of the target segment
b. Model 2: Sums-of-products models using simulated annealing method for vowels and consonants

c. Model 3: Durations from ETRI TTS demos

6.2.2 Test procedure

This perceptual study was only a brief and rather informal investigation. Ten subjects participated in the perceptual evaluation, all of whom are native Korean speakers. They had all been studying in London for two years. They did not have any known hearing impairment and were not especially familiar to listening to synthetic speech. They were given the selected nine sentences produced using each of the three different models in two judgement tasks. Thus each subject listened to nine triplets of sentences twice. In order to avoid any judgement bias derived from the order in which the models were used, the order of the models within the triplets was randomised. The "random" ordering was done in a way so that, overall, each model had the same distribution with respect to position in the triplets. After each triplet, the subjects were given 10 seconds to make a ranking decision on the quality of the synthesised speech. Two aspects of the quality were evaluated: clarity and preference. Subjects were asked to make a judgement on the clarity of the sentences the first time they listened to the triplet and to make a judgement on their general preference in a second listening. The best one was graded 3, the next graded 2, and the worst 1.

6.2.3 Results

90 judgements (9 sentences × 10 subjects) for each of the three models were obtained. Each was converted to 3 pairwise preferences and are summaried as shown in Table 6-5 and Table 6-6. These summaries are plotted as a preference matrix as in Table 6-7 and Table 6-8. After summing up individual rows, we obtain the total number of preferences for each model. "Sign test"[5] was used to check whether any differences between these

models were simply the result of chance. Subjects were also encouraged to discuss their subjective impression of the synthetic speech.

The 90 preference judgments for the 3 models for clarity or for general preference are given in Appendix 5.

Table 6-5.
Pairwise preference summaries for clarity level.

Model 1	Model 2
65	25

Model 1	Model 3
29	61

Model 2	Model 3
25	65

Table 6-6.
Pairwise preference summaries for general preference.

Model 1	Model 2
57	33

Model 1	Model 3
53	37

Model 2	Model 3
28	62

These summaries are plotted in preference matrix form as follows.

Table 6-7.
Preference matrix for clarity level.

	Model 1	Model 2	Model 3	Total
Model 1		65	29	94
Model 2	25		25	50
Model 3	61	65		126

Table 6-8.
Preference matrix in general preference.

	Model 1	Model 2	Model 3	Total
Model 1		57	53	110
Model 2	33		28	61
Model 3	37	62		99

In terms of clarity, subjects preferred model 3, where ETRI durations were used, to the other models. The CART model followed the ETRI model and the sums-of-products model was the least preferred. All differences were statistically significant at $p < 0.01$, see Table 6-9. In terms of general preference, subjects' preferences were more balanced, though sums-of-products model was still the least preferred. The CART model was most preferred by subjects, though the difference is not statistically significant, see Table 6-9.

Table 6-9.
Likelihood of results occurring by chance for each pair of models (sign test).

Clarity level		
Model 1 vs. Model 2	Model 1 vs. Model 3	Model 2 vs. Model 3
$p < 0.01$	$p < 0.01$	$p < 0.01$
General preference		
Model 1 vs. Model 2	Model 1 vs. Model 3	Model 2 vs. Model 3
$p < 0.05$	$p = 0.1$	$p < 0.01$

During a final discussion, the sentences were played again to obtain their impressions. A number of subjects suggested that the speech produced by ETRI model was too slow to consider it natural-sounding. One of the subjects' complaints was that nasal consonants were too long in many instances from the CART model and sums-of-products model. Most subjects agreed that in most cases vowel durations were satisfactory in all models. Some indicated that in the cases of the CART model and the sums-of-products model, bilabial stops sounded tense in some cases. Overall, subjects seemed more sensitive to consonant duration than vowel duration. Comments about discontinuities in the synthesised speech were made for all models.

The result suggests that the difference in tempo is significant perceptually, because it could explain why durations obtained by ETRI were preferred for clarity. The fact that the perceptual preference of the CART and ETRI durations were so similar means that these models almost certainly *do* produce realistic segment durations. The poor perceptual performance of the sums-of-products model shows that the objective performance measures—correlation and RMSE—were not completely useless.

NOTES

[1] "Hanmal (HN 1.0)" can be downloaded from http://tcts.fpms.be.ac/synthesis/.

[2] After agreement between Gim and Chung, the owners of the diphone database and the author of MBROLA, the database was processed by MBROLA team in Belgium and adapted to the MBROLA format, for free. The resulting MBROLA diphone database is a copyright of T. Dutoit at Faculté Polytechnique de Mons in Belgium. Non-commercial use of the database in the framework of the MBROLA project will be automatically granted to Internet users. The commercial rights on the database was transferred to Gim and Chung under a license agreement.

[3] The details of SAMPA can be found at http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[4] SFS software can be downloaded from http//www.phon.ucl.ac.uk/resource/sfs/.

[5] "Sign test" program was obtained from http://fonsg3.let.uva.nl/Service/Statistics.html.

7. CONCLUSION

This thesis has investigated the linguistic factors affecting segment duration in spoken Korean. It has produced some duration models which could be used in Korean language text-to-speech (TTS) systems.

In chapter 2, based on the previous research on the timing of spoken Korean, it was found that more information was needed on the relative importance of different phrase boundaries, the relative size of initial and final boundary effects, and whether syllables in post-initial or penultimate positions are also lengthened. It was also found that more information is required about how syllable structure affects segment duration, how consonants have different behaviours in onset and coda positions, and on which segmental features have a systematic effect.

Chapter 3 showed that recent research on timing has used sequential rule systems, classification and regression tree (CART) decision tree models and sums-of-products models. It was suggested that rule systems could be ignored as they are a special kind of sums-of-products model and it is difficult to develop and maintain them. CART models were suggested to be simple to build and use, with good performance. It was found that sums-of-products models can show excellent performance but are rather tricky to build, since they require complex data analysis to unravel interactions. It was claimed that formant and diphone-style synthesis requires a numerical model to predict durations in context while corpus-based unit selection synthesis needs to know which factors are most important for unit selection. We suggested that Korean language duration

modelling requires further research, because it is rather undeveloped compared to English and Japanese.

Chapter 4 described the design of the corpus which was used in the analysis of the timing in spoken Korean. For the main corpus, it was shown how the training data set and the test data set were prepared, recorded, and processed for statistical analysis. Guidelines for the annotation of each major class feature based on acoustic information were proposed. Chapter 4 presented a list of phonological rules which were used in the pronunciation of sentences and in the building of a pronunciation dictionary. In the description of database processing, it was shown that how phonological features and prosodic phrase features could be processed to produce a feature analysis of segment duration. Two feature sets were prepared for analysis. The first "Compact feature set" used seven n-ary parameters for vowels and eight n-ary parameters for consonants and the other "Binary feature set" used 69 binary features for all segments.

Chapter 5 described the training of CART decision tree models and sums-of-products models on the training data and their testing on the test data. The best performance was obtained from a CART decision tree model applied to the "binary feature set" where vowels and consonants were modelled together and where names and general class features of the target segment were incorporated into the decision tree. The correlation between the observed and the predicted durations was 0.77 and the mean squared error of prediction was 25.11 ms. The best sums-of-products model gave a correlation of 0.69 for vowels and 0.54 for consonants. By using a CART decision tree model with segment durations as z-scores, the linguistic implications of the model were investigated. The

CART modelling showed that prosodic phrase features have the greatest influence on the segment duration, among them, the AP-final position feature. Syllable structure information such as onset position feature and the syllable structure feature was also influential effects, but on a smaller scale. While phrase boundaries had lengthening effects, these syllable structure features had shortening effects. The contextual effect of surrounding segments were not so consistent or large except for the influence of adjacent nasals. Both preceding and following nasal contexts caused shortening effects.

Chapter 6 described a small-scale subjective evaluation of the quality of the durations produced by these duration models. Durations calculated by the best CART model and the best sums-of-products model were applied to the Korean diphone database "Hanmal (HN 1.0)". This MBROLA format Korean diphone database was also built as part of this study. A reference set of durations were produced by the ETRI TTS system. Clarity and preference were used to evaluate the "naturalness" of the synthesised speech from the three models. It was found that the subjects were more sensitive to consonant durations than to vowel durations. The CART model was preferred in the preference test by a small margin, the synthetic speech from ETRI was superior in the clarity test.

The work described in this thesis has only started to address the important issues in Korean prosody. The limitations of the study are also opportunities for further work in this area. Firstly, this analysis was solely based on one reading of one text by one speaker in one style. It was not able to investigate changes in speaking rate, variations such as a position within a paragraph, or shifts in emphasis and focus (Klatt, 1976). This analysis was only concerned with the timing variation caused by prosodic phrase

structure, word boundary, syllable structure, and phonological and phonetic segmental effects and not with the variation caused by stress, segment and syllable numbers. Secondly, the sums-of-products models were not based on detailed analysis of the interactions between factors and we were not able to reproduce the success of van Santen's (1994) modelling of English.

Despite these limitations, the analyses of this thesis are believed to contribute to the study of spoken Korean in the following aspects. Firstly, it showed how much prosodic phrase features influenced duration and which of these was more important. Secondly, it showed how phonological distinctive features could be used for modelling in such a way as to allow a linguistic interpretation of the model. Thirdly, these observations allowed us to determine which factors and which structures are most important in Korean prosody.

In the course of preparing the experiments, a labelled database of spoken Korean was constructed. As a result of the experiments, a trained CART model for synthesis was obtained. Durations of segments in a new text can be rapidly predicted from this model. The Hanmal diphone database for Korean speech synthesis was also developed as a by-product of the perceptual testing. This database is now publicly available and currently in use by other researchers.

BIBLIOGRAPHY

Allen, J. S. Hunnicut and D. H. Klatt. 1987. *From Text to Speech: The MITalk System*, Cambridge: U.K., Cambridge University Press.

Beckman, Mary and Janet Pierrehumbert. 1986. Intonational structure in English and Japanese, *Phonology 3*, 255-309.

Black, Alan W. and Paul A. Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis, *Proceedings of Eurospeech '97*. 601-604.

Black, Alan W., Paul A. Taylor and R. Caley. 1999. The Festival Speech Synthesis System: system documentation, edition 1.4, for Festival version 1.4.0, *CSTR Web Page*, University of Edinburgh.

Breen, Andrew P. 1995. A simple method for predicting the duration of syllables. *Proceedings of Eurospeech '95*, Madrid, 595-598.

Breen, Andrew P. and P. Jackson. 1998. A phonologically motivated method of selecting non-uniform units. *Proceedings of International Conference on Speech and Language Processing '98*, Sydney, Australia.

Breiman, L., J. Friedman, R. Olshen and C. Stone. 1984. *Classification and Regression Trees*, Monterey, Chapman & Hall/CRC.

Campbell, W. Nick and AlanW. Black. 1997. Prosody and the selection of source units for concatenative synthesis, in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (eds.), New York: Springer, 279-292.

Campbell, W. Nick. 2000. Timing in speech: a multi-level process, in *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce,* Merle Horne (ed.), Dordrecht, the Netherlands: Kluwer Academic Publisher, 281-334.

Cho, S. B. 1967. *A Phonological Study of Korean with a Historical Analysis.* Uppsala: Universitetet.

Choi, Hyun Bae. 1983. *Uli Malpon.* Seoul: Jeongeumsa.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*, New York: Harper & Row, Publishers.

Chung, Hyunsong and Mark A. Huckvale. 2001a. Linguistic factors affecting timing in Korean with application to speech synthesis, *Proceedings of Eurospeech 2001*, vol. 2, 815-819.

Chung, Hyunsong and Mark A. Huckvale. 2001b. Analysis of the timing of spoken Korean using a Classification and Regression Tree (CART) Model, *The Korean Journal of Speech Sciences*, The Korean Association of Speech Sciences, vol. 8 (1), 77-91.

Chung, Hyunsong, and Mark A. Huckvale. 1999. Modelling of temporal compression in Korean, in *Harvard Studies in Korean Linguistics VIII*, Susumu Kuno et al. (eds.), Seoul: Hanshin, 102-116.

Chung, Hyunsong, Gyeongseog Gim and Mark A. Huckvale. 1999. Consonantal and prosodic influences on Korean vowel duration, *Proceedings of Eurospeech '99*, vol. 2, 707-710.

Chung, Hyunsong, Mark A. Huckvale and Gyeongseog Gim. 1999. A new Korean speech synthesis system and temporal model. *Proceedings of 16th ICSP*, Seoul, vol. 1, 203-208.

Chung, Kook et al. 1993. *A Study of the Theoretical Analysis of the Prosodic Structure and Statistical Investigation on Phonotactics*, Korea Electronics and Telecommunications Research Institute (ETRI) Technical Report.

Chung, Kook et al. 1995. *A Study of Phonological and Grammatical Structures of Korean for the Implementation of Automatic Telephone System.* Korea Telecom (KT) Research & Development Group Technical Report.

Chung, Kook et al. 1997. *A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System.* Korea Telecom (KT) Research & Development Group Technical Report.

Chung, Kook. 1994. *Understanding of Generative Phonology*, Seoul: Hanshin Munhwasa.

Chung, So Woo et al. 1994. A study on the phonological structure and the rule application, *Proceedings of '94 HCI Workshop*, Information Society of Korea, 37-48.

Clements, Nick. 1989. A unified set of features for consonants and vowels, *Ms.*, Cornell University.

Coker, C. H., Noriko Umeda and C. P. Browman. 1973. Automatic synthesis from ordinary English text, *IEEE Transactions Audio Electroacoustics*, vol. AU-21 (3), 293-298.

Conkie, A. 1999. A robust unit selection system for speech synthesis, In *137th Meeting of the Acoustical Society of America.*

Crystal, Thomas H. and Arthur S. House. 1982. Segmental durations in connected speech signals: Preliminary results, *Journal of Acoustical Society of America*, vol. 72 (3), 705-716.

Crystal, Thomas H. and Arthur S. House. 1988a. Segmental durations in connected-speech signals: current results, *Journal of Acoustical Society of America*, vol. 83 (4), 1553-1573.

Crystal, Thomas H. and Arthur S. House. 1988b. Segmental durations in connected-speech signals: syllabic stress, *Journal of Acoustical Society of America*, vol. 83 (4), 1574-1585.

Crystal, Thomas H. and Arthur S. House. 1988c. A note on the durations of fricatives in American English, *Journal of Acoustical Society of America*, vol. 84 (5), 1932-1935.

Crystal, Thomas H. and Arthur S. House. 1990. Articulation rate and the duration of syllables and stress groups in connected speech, *Journal of Acoustical Society of America*, vol. 88 (1), 101-112.

De Jong, Kenneth. 1994. Initial tones and prominence in Seoul Korean, *Ohio State University Working Papers in Linguistics 43*, 1-14.

Deans, Paul, Andrew P. Breen, and Peter Jackson. 1999. Cart-based duration modeling using a novel method of extracting prosodic features, *Proceedings of Eurospeech '99*, vol. 4, 1823-1826.

Dirksen, A. and J. S. Coleman. 1997. All-prosodic speech synthesis, in *Progress in Speech Synthesis*, Jan P.H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (eds.), New York: Springer, 91-108.

Dutoit, Thierry, V. Pagel, N. Pierret, F. Bataille, and O. van der Vreken. 1996. The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proceedings of 4th ICSLP*, Philadelphia. vol. 3, 1393-1396.

Dutoit, Thierry. 1997. High-quality text-to-speech synthesis: an overview, *Journal of Electrical and Electronics Engineerings, Austria: Special Issue on Speech Recognition and Synthesis.*

Dutoit, Thierry. and H. Leich. 1993. An analysis of the performances of the MBE model when used in the context of a Text-to-Speech system, *Proceedings of Eurospeech '93.* 531-534.

Febrer, Albert, J. Padrell, and A. Bonafonte. 1998. Modelling phone duration, *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis.*

Fischer-Jørgensen, E. 1964. Sound duration and place of articulation, *Z. Phonet. Sprachwiss. Kommunikationsforsch*, vol. 17, 175-207.

Gay, Thomas. 1968. Effect of speaking rate on diphthong formant movements, *Journal of Acoustical Society of America*, vol. 44 (6), 1570-1573.

Goldsmith, John A. 1976. *Autosegmental Phonology*, Indiana University Linguistics Club.

Gu, Wentao Chilin Shih, Jan P. H. van Santen. 1999. An efficient speaker adaptation method for TTS duration model, *Proceedings of Eurospeech '99*, vol. 4, 1839-1842.

Halle, M. and K. N. Stevens. 1971. A note on laryngeal features, in *Quarterly Progress Reports*, Research Lab of Electronics, MIT, 101, 198-213.

Han, Mieko S. 1963. *Acoustic Phonetics of Korean: Korean Vowels*, Technical Report, University of California, Los Angeles.

Han, Mieko S. 1964. *Duration of Korean Vowels, Studies in the Phonology of Asian Languages II*, Acoustic Phonetics Research Laboratory, University of Southern California, Los Angeles.

Han, Mieko S. and R. Weitzman. 1965. *Acoustic Characteristics of Korean Stop Consonants, Studies in the Phonology of Asian Languages III*, Acoustic Phonetics Research Laboratory, University of Southern California, Los Angeles.

Han, Mieko S. and R. Weitzman. 1967. *Acoustic Features in the Manner-differentiation of Korean Stop Consonants, Studies in the Phonology of Asian Languages* V. Acoustic Phonetics Research Laboratory, University of Southern California, Los Angeles.

Han, Mieko S. and S. Ross. 1968, Korean affricates. *Studies in the Phonology of Asian Languages VII*, Acoustic Phonetics Research Laboratory, University of Southern California, Los Angeles.

Han, Sun-Hee and Mira Oh. 1999. The boundary tone in Korean intonational phrases, *The Korean Journal of Speech Sciences*, vol. 5 (2), 109-129.

Harris, M. O. and Noriko Umeda. 1974. Effect of speaking mode on temporal factors in speech: vowel duration, *Journal of Acoustical Society of America*, vol. 56 (3), 1016-1018.

Hawkins, S., Jill House, Mark Huckvale, J. Local and R. Ogden. 1998. ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Proceedings of 5th ICSLP*, Sydney, 1707-1710.

Hayes, B. 1989. The prosodic hierarchy in meter, Kiparsky P. & G. Youmans (eds.), *Rhythm and Meter*, Orlando: Academic Press, 201-260.

Heo, Woong. 1985. *Gugeo Eumunhak*, Seoul: Saem Munhwasa.

Hogg, R. and C. B. McCully. 1987. *Metrical Phonology*. Cambridge: Cambridge University Press.

Holmes, J. N. 1973. The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer, *IEEE Transactions Audio Electroacoustics.* AU-21, 298-305.

Horne, Merle (ed.). 2000. *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce*, Dordrecht, the Netherlands: Kluwer Academic Publishers.

House, Arthur S. 1961. On vowel duration in English, *Journal of Acoustical Society of America*, vol. 33 (9), 1174-1178.

House, Arthur S. and G. Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels, *Journal of Acoustical Society of America*, vol. 25, 105-113.

Huckvale, Mark A. 1999. Representation and processing of linguistic structures for an all-prosodic synthesis system using XML, *Proceedings of Eurospeech '99*. vol. 4, 1847-1850

Hunt, Andrew J. and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech, and Signal Processing*, IEEE.

Jang, Tae-yeoub. 2000. *Phonetics of Segmental F_0 and Machine Recognition of Korean Speech*, Ph.D. thesis, University of Edinburgh.

Jones, Daniel. 1976. *The Phoneme, its Nature and Use*, Cambridge: Heffer.

Jun, Sun-Ah. 1993. *The Phonetics and Phonology of Korean Prosody*, Ph.D. dissertation, Ohio State University.

Jun, Sun-Ah. 1998. The accentual phrase in the Korean prosodic hierarchy, *Phonology 15 (2)*, 189-226.

Kang, Kyung-Shim. 2000. On Korean fricatives, *The Korean Journal of Speech Sciences*, The Korean Association of Speech Sciences, vol. 7 (3), 53-68.

Kim, D. W. 1994. An acoustic study of word-timing with reference to Korean, *SCAS 11 (1)*, 323-327.

Kim, Kee-Ho. 1990. Revisiting distinctive feature approach in speech recognition, presented at *Seoul International Conference on Natural Language Processing.*

Kim, Kong-On. 1974. *Temporal Structure of Spoken Korean: an Acoustic Phonetic Study*, Ph.D. dissertation, University of Southern California.

Kim, Young Seok. 1987. *English Phonology*, Seoul: Hanshin Munhwasa.

Klatt, D. H. 1972. Acoustic theory of terminal analog speech synthesis, *Proceedings of International Conference of Acoustic Speech Signal Process*, ICASSP-72, 131-135.

Klatt, D. H. 1973. Interaction between two factors that influence vowel duration, *Journal of Acoustical Society of America*, vol. 54 (4), 1102-1104.

Klatt, D. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *Journal of Acoustical Society of America*, vol 59 (5), 1208-1221.

Klatt, D. H., 1979. Synthesis by rule of segmental durations in English sentences, in *Frontiers of Speech Communication Research*, B. Lindblom and S. Öhman (eds.), New York: Academic Press, 287-300.

Klatt, D. H., 1980. Software for a cascade/parallel formant synthesizer, *Journal of Acoustical Society of America*, vol. 67 (3), 971-995.

Klatt, D. H., 1987. Review of text-to-speech conversion for English, *Journal of Acoustical Society of America*, vol. 82 (3), 737-793.

Koo, Hee San. 1986. *An Experimental Acoustic Study of the Phonetics of Intonation in Standard Korean.* Ph.D. dissertation, University of Texas at Austin.

Koo, Hee San. 1998. The influence of consonant environment upon the vowel duration. *Korean Journal of Speech Sciences*, The Korean Association of Speech Sciences, 7-18.

Korean Ministry of Education (eds.). 1992. *Gukeo Eomun Gyujeongjip.* Seoul: Daehan Gyogwaseo Jusikhoesa.

Kwack, Dong-Kee. 1992. *A Phonological Study on the Korean Prosodic Structures.* Ph.D. dissertation, Seoul National University.

Lee, Hi Seung. 1956. *Introduction to Korean Language*, Minjung Seogwan, Seoul, Korea.

Lee, Ho-Young. 1990. *The Structure of Korean Prosody*, Ph.D. thesis, University of London.

Lee, Ho-Young. 1996a. Phonological rules with the domain of the rhythm unit, *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee.* Seoul National University Press, 146-153.

Lee, Ho-Young. 1996b. *Korean Phonetics*, Seoul: Taehaksa.

Lee, Hyun Bok. 1973. A phonetic study of the accent in Korean. *Munli Daehak-Bo.* Seoul National University. 113-128.

Lee, Hyun Bok. 1982. Phonetic variation of Korean speech sounds as conditioned by tempo and rhythm, *Language Research*, vol. 18 (1), 115-120.

Lee, Sangho and Young-Hwan Oh. 1999a. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems, *Speech Communication 28*, 283-300.

Lee, Sangho and Young-Hwan Oh. 1999b. CART-based modeling of Korean segmental duration, *Proceedings of the Second International Workshop on East-Asian Language Resources and Evaluation*, Oriental COCOSDA '99, 109-112.

Lee, Sangho and Young-Hwan Oh. 1999c. Generating Korean F0 contour using CART, *Proceedings of 16th International Conference on Speech Processing,* Seoul, 177-182.

Lee, Soon-Hyang and Hee San Koo. 1997. The effects of the speaking rate on the duration of syllable before boundary, *The Korean Journal of Speech Sciences*, The Korean Association of Speech Sciences, vol. 1, 103-111.

Lee, Sung Nyong. 1955. *Korean Phonology*, Seoul, Korea: Minjung Seogwan.

Lee, Yang Hee. 1996. Modelling of segment duration in Korean speech synthesis, *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee.* Seoul National University Press, 249-274.

Lehiste, Ilse. 1970. *Suprasegmentals,* Cambridge: The MIT Press.

Lehiste, Ilse. 1973. Rhythmic units and syntactic units in production and perception, *Journal of Acoustical Society of America*, vol. 54 (5), 1228-1234.

Lehiste, Ilse. 1976. Segmental features of speech, in *Contemporary Issues in Experimental Phonetics*, N. J. Lass (ed.), New York: Academic, 225-239.

Lehiste, Ilse. 1979. Perception of sentence and paragraph boundaries, in *Frontiers of Speech* Research, Lindblom and S. Öhman (eds.), New York: Academic, 191-201.

Liberman, M. and A. Prince. 1977. On stress and linguistic rhythm, *Linguistic Inquiry 8*, 249-336.

Lindblom, D. and K. Rapp. 1973. Some temporal properties of spoken Swedish, *PILUS*, vol. 21, 1-59.

Local, John and Richard Ogden. 1997. A model of timing for nonsegmental phonological structure, in *Progress in Speech Synthesis*, Jan P. H. van Santen, R W. Sproat, J. P. Olive & J. Hirschberg (eds.), New York: Springer, 109-122.

Luce, Paul A. and Jan Charles-Luce. 1985. Contextual effects on vowel duration , closure duration, and the consonant/vowel ratio in speech production, *Journal of Acoustical Society of America*, vol. 78 (6), 1949-1957.

Maack, A. 1953. Die Beeinflussung der Sonantendauer durch die Nachbarkonsonanten, *Z. Phonet. Sprachwiss. Kommunikationsforsch*, vol. 7, 104-128.

Martin, Samuel E. 1951. Korean phonemics, *Language 27*, 519-533.

Moon, Yang Soo. 1996. Syllable-based phonology and Korean syllable structure, *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee.* Seoul National University Press, 26-49.

Nelder, J. A. & R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, The British Computer Society, vol. 7, 308-313.

Nespor, M. and I. Vogel. 1986. *Prosodic Phonology*, Dordrecht, the Netherlands: Fortis.

Ogden, Richard and John Local. 1992. *Yearly Report 1992.*

Ogden, Richard and John Local. 1996. *YorkTalk Annual Report 1991*, York Research Papers in Linguistics, YRPL 96-01.

Ogden, Richard, John Local and Paul Carter. 1999. Temporal interpretation in ProSynth, a prosodic speech synthesis system, *Proceedings of ICPhS99*, 1059-1062.

Oh, Mira. 1989. The phonological word in Korean, *Papers in Honor of Professor Hye Sook Lee*, Seoul: Hanshin Munhwasa.

Parmenter, C. E., and S. N. Treviño. 1935. The length of the sounds of a middle westerner, *American Speech*, vol. 10, 129-133.

Peterson, G. E. and Ilse Lehiste. 1960. Duration of syllable nuclei in English, *Journal of Acoustical Society of America*, vol. 32 (6), 693-703.

Pickett, J. M. 1980. *The Sounds of Speech Communication, A Primer of Acoustic Phonetics and Speech Perception*, Baltimore: University Park Press.

Pierrehumbert, Janet. 1981. Synthesizing intonation, *Journal of Acoustical Society of America*, vol. 70 (4), 945-995.

Port, Robert F. 1981. Linguistic timing factors in combination, *Journal of Acoustical Society of America*, vol. 69 (1), 262-274.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C*, 2nd edition, Cambridge: Cambridge University Press.

Riley, M. 1992. Tree-based modelling of segmental durations, in *Talking Machines. Theories, Models and Designs*, G. Bailly, C. Benôit (eds.), Amsterdam, the Netherlands: North-Holland, 265-273.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: the Relation between Sound and Structure*, Cambridge, Mass.: MIT Press.

Selkirk, Elisabeth. 1986. On derived domains in sentence phonology, *Phonology 3*, 371-405.

Selkirk, Elisabeth. and Tong Shen. 1990. Prosodic domains in Shanghai Chinese, in *The Phonology-Syntax Connection*, in S. Inkelas and D. Zec (eds.), Chicago: University of Chicago Press, 313-338.

Shih, Chilin and Ao, B. 1997. Duration study for the Bell Laboratories Mandarin text-to-speech system, in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (eds.), New York: Springer, 383-400.

Shih, Chilin, Wentao Gu, and Jan P. H. van Santen. 1998. Efficient adaptation of TTS duration model to new speakers, *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis.*

Song, Min-Suck et al. 1995. The acoustic properties of flap [ɾ] in natural sentence utterance, *Proceedings of '95 Spring Meeting*, The Korean Society of Cognitive Sciences, 11-17.

Sproat, Richard (ed.). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Stetson, R. H. 1951. *Motor Phonetics*, 2nd edition, Amsterdam, the Netherlands: North-Holland.

Stevens, K. N. 1972. The quantal nature of speech: evidence from articulatory-acoustic data, in *Human Communication: A Unified View*, E. E. David and P. B. Denes (eds.), New York: McGraw-Hill, 51-66.

Takeda, Kazuya, Yoshinori Sagisaka and Hisao Kuwabara. 1989. On sentence-level factors governing segmental duration in Japanese, *Journal of Acoustical Society of America*, vol. 86 (6), 2081-2087.

Tanaka, Kimihito et al. 1999. A Japanese Text-to-Speech system based on multi-form units with consideration of frequency distribution in Japanese, *Proceedings of Eurospeech '99*, vol. 2, 839-842.

Taylor, Paul A. and Alan W. Black. 1999. Speech synthesis by phonological structure matching, *Proceedings of Eurospeech '99*, 623-626.

Taylor, Paul. A. 1999. Concept-to-speech synthesis by phonological structure matching, *Proceedings of Royal Society Workshop.*

Umeda, Noriko 1977. Consonant duration in American English, *Journal of Acoustical Society of America*, vol. 61 (3), 846-858.

Umeda, Noriko. 1975. Vowel duration in American English, *Journal of Acoustical Society of America*, vol. 58 (2), 434-445.

Van Santen, J. P. H. 1992. Contextual effects on vowel duration, *Speech Communication 11*, 513-546.

Van Santen, J. P. H. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language 8*, 95-128.

Van Santen, J. P. H. 1995. Timing in Text-to-Speech system, *Proceedings of Eurospeech '95*, 1397-1404.

Van Santen, J. P. H. 1997. Prosodic modeling in Text-to-Speech synthesis, *Proceedings of Eurospeech '97.*

Venditti, Jennifer J. and J. P. H. van Santen. 1998. Modeling segmental durations for Japanese text-to-speech synthesis, *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 31-36.

Young, Steve, J. Jansen, J. Ollason, and P. Woodland. 1996. *HTK Book*, Entropic.

Young, Steve. 1996 Large vocabulary continuous speech recognition: a review, *IEEE Signal Processing Magazine, Fall.*

Yun, Ilsung. 1998. *A Study of Timing in Korean Speech*, Ph.D. thesis, Reading University.

APPENDICES

Appendix 1: Feature Strings for an Example sentence

Sentence:

/onɯl halumanto tsʌnnamkwa kjʌŋnam nɛljuktsipaŋɯlonɯn pɛkmillimit^{h}ʌka nʌmnɯn manhɯn pika nɛljʌs'sɯpnita/

"Today also there was heavy rain – over 100 mm in the inner area of Jeonnam and Gyeongnam counties."

Compact feature set: Feature identity can be found in Table 4-9 and Table 4-10.

Feature	1	2	3	4	5	6	7
oo	mono	pause	nas	v	utt-init	non-fi	
nn	nas	vow	vow	cvc	on	non-init	pw-fi
xx	mono	nas	lat	cvc	non-init	pw-fi	
ll	lat	vow	laxfri	cvc	co	non-init	pw-fi
hh	fri	lat	vow	cv	on	pw-init	non-fi
aa	mono	laxfri	fla	cv	pw-init	non-fi	
rr	fla	vow	vow	cv	on	non-init	non-fi
uu	mono	fla	nas	cv	non-init	non-fi	
mm	nas	vow	vow	cvc	on	non-init	non-fi
aa	mono	nas	nas	cvc	non-init	non-fi	
nn	nas	vow	laxstp	cvc	co	non-init	non-fi
t0	stop	nas	vow	cv	on	non-init	ap-fi
oo	mono	laxstp	laxaff	cv	non-init	ap-fi	
c0	aff	vow	vow	cvc	on	ap-init	non-fi
vv	mono	laxaff	nas	cvc	ap-init	non-fi	
nn	nas	vow	nas	cvc	co	ap-init	non-fi
nn	nas	nas	vow	cvc	on	non-init	non-fi
aa	mono	nas	nas	cvc	non-init	non-fi	
mm	nas	vow	laxstp	cvc	co	non-init	non-fi
k0	stop	nas	vow	cv	on	non-init	pw-fi
wa	di	laxstp	laxstp	cv	non-init	pw-fi	
k0	stop	vow	vow	cvc	on	pw-init	non-fi
yv	di	laxstp	nas	cvc	pw-init	non-fi	
ng	nas	vow	nas	cvc	co	pw-init	non-fi
nn	nas	nas	vow	cvc	on	non-init	ap-fi
aa	mono	nas	nas	cvc	non-init	ap-fi	
mm	nas	vow	nas	cvc	co	non-init	ap-fi
nn	nas	nas	vow	cv	on	ap-init	non-fi
ee	mono	nas	fla	cv	ap-init	non-fi	
rr	fla	vow	vow	cvc	on	non-init	non-fi
yu	di	fla	laxstp	cvc	non-init	non-fi	
k0	stop	vow	tnsaff	cvc	co	non-init	non-fi
cc	aff	laxstp	vow	cv	on	non-init	non-fi
ii	mono	tnsaff	laxstp	cv	non-init	non-fi	
p0	stop	vow	vow	cvc	on	non-init	non-fi
aa	mono	laxstp	nas	cvc	non-init	non-fi	
ng	nas	vow	vow	cvc	co	non-init	non-fi

Feature	1	2	3	4	5	6	7
rr	fla	vow	vow	cv	on	non-init	non-fi
oo	mono	fla	nas	cv	non-init	non-fi	
nn	nas	vow	vow	cvc	on	non-init	ip-fi
xx	mono	nas	nas	cvc	non-init	ip-fi	
nn	nas	vow	pause	cvc	co	non-init	ip-fi
p0	stop	pause	vow	cvc	on	ip-init	non-fi
ee	mono	laxstp	nas	cvc	ip-init	non-fi	
ng	nas	vow	nas	cvc	co	ip-init	non-fi
mm	nas	nas	vow	cvc	on	non-init	non-fi
ii	mono	nas	lat	cvc	non-init	non-fi	
ll	lat	vow	lat	cvc	co	non-init	non-fi
ll	lat	lat	vow	cv	on	non-init	non-fi
ii	mono	lat	nas	cv	non-init	non-fi	
mm	nas	vow	vow	cv	on	non-init	non-fi
ii	mono	nas	aspstp	cv	non-init	non-fi	
th	stop	vow	vow	cv	on	non-init	non-fi
vv	mono	aspstp	laxstp	cv	non-init	non-fi	
k0	stop	vow	vow	cv	on	non-init	pw-fi
aa	mono	laxstp	nas	cv	non-init	pw-fi	
nn	nas	vow	vow	cvc	on	pw-init	non-fi
vv	mono	nas	nas	cvc	pw-init	non-fi	
mm	nas	vow	nas	cvc	co	pw-init	non-fi
nn	nas	nas	vow	cvc	on	non-init	ap-fi
xx	mono	nas	nas	cvc	non-init	ap-fi	
nn	nas	vow	nas	cvc	co	non-init	ap-fi
mm	nas	nas	vow	cv	on	ap-init	non-fi
aa	mono	nas	nas	cv	ap-init	non-fi	
nn	nas	vow	vow	cvc	on	non-init	pw-fi
xx	mono	nas	nas	cvc	non-init	pw-fi	
nn	nas	vow	laxstp	cvc	co	non-init	pw-fi
p0	stop	nas	vow	cv	on	pw-init	non-fi
ii	mono	laxstp	laxstp	cv	pw-init	non-fi	
k0	stop	vow	vow	cv	on	non-init	pw-fi
aa	mono	laxstp	nas	cv	non-init	pw-fi	
nn	nas	vow	vow	cv	on	pw-init	non-fi
ee	mono	nas	fla	cv	pw-init	non-fi	
rr	fla	vow	vow	cv	on	non-init	non-fi
yv	di	fla	tnsfri	cv	non-init	non-fi	
ss	fri	vow	vow	cvc	on	non-init	non-fi
xx	mono	tnsfri	nas	cvc	non-init	non-fi	
mm	nas	vow	nas	cvc	co	non-init	non-fi
nn	nas	nas	vow	cv	on	non-init	non-fi
ii	mono	nas	laxstp	cv	non-init	non-fi	
t0	stop	vow	vow	cv	on	non-init	utt-fi
aa	mono	laxstp	pause	cv	non-init	utt-fi	

Binary feature set: Feature identity can be found in Table 4-11.

Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
oo	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0
xx	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ll	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
hh	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
aa	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
rr	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
uu	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
aa	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
t0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
oo	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
c0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0
vv	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
nn	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
aa	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
k0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
wa	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
k0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
yv	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
ng	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
nn	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
aa	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
nn	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ee	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
rr	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
yu	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
k0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
cc	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
ii	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
p0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
aa	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ng	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
xx	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
rr	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
oo	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0
xx	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
p0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ee	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ng	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
mm	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ii	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ll	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
ll	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ii	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
ii	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
th	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0

Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
vv	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0
k0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
aa	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
vv	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
nn	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
xx	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
mm	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
aa	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
xx	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
p0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ii	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
k0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
aa	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
nn	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ee	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
rr	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
yv	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ss	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
xx	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
mm	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
nn	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ii	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
t0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
aa	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0

Feature	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
oo	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
nn	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
xx	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
ll	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
hh	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1
aa	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
rr	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0	0	0
uu	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
mm	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
nn	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
t0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	0	0
oo	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
c0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
vv	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
nn	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
nn	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
mm	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0
k0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0	0	0
wa	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
k0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1
yv	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1
ng	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
nn	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
mm	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
nn	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1
ee	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
rr	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0
yu	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
k0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
cc	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0
ii	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
p0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
ng	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0
xx	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
rr	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	0	0
oo	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
xx	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
nn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
p0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1
ee	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1
ng	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1
mm	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0
ii	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
ll	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
ll	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0
ii	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
mm	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0
ii	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
th	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0

Feature	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
vv	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
k0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
vv	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1
mm	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
nn	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
xx	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
nn	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0
mm	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1
aa	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
nn	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
xx	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
nn	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0
p0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
ii	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1
k0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
aa	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1
ee	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
rr	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0
yv	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
ss	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
xx	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
mm	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
nn	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0
ii	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
t0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
aa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0

Feature	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
oo	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
nn	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
xx	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
ll	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
hh	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
rr	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
uu	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
mm	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
t0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
oo	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
c0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
vv	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
mm	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
k0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
wa	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
k0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
yv	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
ng	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
aa	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
mm	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
ee	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
rr	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
yu	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
k0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
cc	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ii	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
p0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ng	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
xx	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
rr	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
oo	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
nn	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
xx	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
nn	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
p0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
ee	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
ng	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
mm	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
ii	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
ll	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
ll	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ii	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
mm	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ii	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
th	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0

Feature	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
vv	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
k0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
vv	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
mm	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
xx	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
mm	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
xx	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
p0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ii	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
k0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
aa	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ee	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
rr	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
yv	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
ss	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
xx	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
mm	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
nn	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
ii	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
t0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
aa	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1

Appendix 2

Distribution of segments in the test data set.

Phone	Counts	%	Mean (ms)	sd. (ms)
i	899	8.37	57.74	38.88
u	291	2.71	50.91	25.82
e	551	5.13	80.56	39.41
o	481	4.48	84.49	60.13
a	933	8.69	84.18	46.35
ʌ	527	4.91	72.01	36.24
ɯ	601	5.60	46.84	20.31
wa	57	0.53	103.51	71.14
we	67	0.62	66.87	32.06
wi	25	0.23	96.72	39.92
wʌ	31	0.29	94.58	48.76
ja	28	0.26	104.89	51.62
je	25	0.23	88.40	37.23
jo	61	0.57	74.75	41.59
ju	37	0.34	75.65	50.86
jʌ	204	1.90	79.27	42.66
ɯi	11	0.10	107.64	38.19
m	442	4.12	56.60	25.71
n	1108	10.32	67.54	44.54
ŋ	352	3.28	74.59	32.27
l	355	3.31	66.43	27.51
ɾ	301	2.80	28.76	9.23
p^h	96	0.89	85.31	33.81
p	296	2.76	56.93	26.83
p'	26	0.24	72.88	23.47
t^h	94	0.88	91.03	28.75
t	518	4.82	49.58	22.34
t'	71	0.66	68.08	19.39
k^h	71	0.66	90.44	27.01
k	656	6.11	54.76	28.83
k'	89	0.83	71.42	22.30
ts^h	160	1.49	98.34	28.00
ts	344	3.20	70.37	34.77
ts'	48	0.45	69.67	19.76
s	424	3.95	72.22	26.43
s'	168	1.56	101.31	21.42
h	289	2.69	43.19	24.10

sd. = standard deviation

Appendix 3

The growth in the correlation coefficients for each stepwise refinement of the binary feature set for CART model.

	Name & manner		Name only		Manner only		z-score	
1	name	0.389	name	0.389	1_AG	0.324	1_AG	0.313
2	1_AG	0.604	1_AG	0.604	ON	0.474	ON	0.389
3	AG_1	0.645	AG_1	0.645	AG_1	0.538	AG_1	0.476
4	ON	0.677	ON	0.677	CVC	0.583	nas_	0.503
5	CVC	0.697	CVC	0.697	V_	0.600	_nas	0.526
6	vce_	0.710	vce_	0.710	nas_	0.616	PW_1	0.542
7	_V	0.719	_V	0.719	1_PW	0.631	vce_	0.557
8	1_PW	0.726	1_PW	0.726	fla	0.643	1_PW	0.569
9	nas_	0.732	nas_	0.732	V	0.654	CVC	0.581
10	VC	0.740	VC	0.740	_hiV	0.662	cor_	0.589
11	cor_	0.745	cor_	0.745	PW_1	0.669	stp_	0.597
12	PW_1	0.748	PW_1	0.748	_dor	0.675	_hiV	0.605
13	_stp	0.751	_stp	0.751	_nas	0.680	_glt	0.611
14	_cor	0.755	_cor	0.755	stp	0.686	_vce	0.616
15	_tns	0.756	_tns	0.756	di	0.691	_cor	0.621
16	stp_	0.757	stp_	0.757	aff	0.695	asp_	0.625
17	2_UTT	0.758	2_UTT	0.758	VC	0.699	VC	0.628
18	nas	0.760	_lab	0.760	_cor	0.703	_tns	0.630
19	_lab	0.761	asp_	0.761	_stp	0.706	_fri	0.631
20	asp_	0.762	glt_	0.761	AG_m	0.708	fri_	0.633
21	glt_	0.763	_lat	0.762	_V	0.710	_dor	0.634
22	_lat	0.764	_loV	0.763	fri_	0.711	hiV_	0.636
23	_loV	0.764	lat_	0.763	_aff	0.713	NUC	0.637
24	_aff	0.765	1_UTT	0.763	_tns	0.715	tns_	0.638
25	lat_	0.765	_mdV	0.764	mdV_	0.716	AG_2	0.639
26	1_UTT	0.765	AG_2	0.764	_lat	0.716	PW_2	0.640
27	_mdV	0.766	_asp	0.764	2_IP	0.717	2_UTT	0.641
28	AG_2	0.766	_aff	0.764	asp_	0.718	lat_	0.642
29	_asp	0.766	_hiV	0.764	CODA	0.719	_V	0.642
30	_hiV	0.766	_fri	0.765	stp_	0.720	_asp	0.643
31	V_	0.767	_fla	0.765	CV	0.720	PW_m	0.643
32	aff	0.768	2_AG	0.765	lat	0.721	1_IP	0.644
33	loV_	0.768	IP_2	0.765	_asp	0.721	_lab	0.645
34	_fla	0.768	PW_m	0.765	fla_	0.721	_stp	0.645
35	tns_	0.768			2_AG	0.721	fla_	0.645
36					loV_	0.722	IP_2	0.645
37					IP_2	0.722	V_	0.646
38					lat_	0.722	1_UTT	0.646
39					_fla	0.722	CODA	0.646
40					aff_	0.722	UTT_2	0.646
41					_loV	0.722		
42					glt_	0.722		

APPENDIX 4

Mean feature effects in z-score of 40 selected features in the binary feature set for CART model. The ordering follows the ranking of the growth in the correlation coefficients for each stepwise refinement.

Ranking	Feature	Partition ∅		Partition 1		Diff
		Mean	Size	Mean	Size	
1	1_AP	-0.12	37170	0.87	4933	0.99
2	ON	0.04	25704	-0.06	16399	-0.09
3	AP_1	-0.06	37041	0.47	5062	0.53
4	nas_	0.03	34478	-0.12	7625	-0.15
5	_nas	0.07	34467	-0.30	7636	-0.36
6	PW_1	-0.03	28198	0.06	13905	0.09
7	vce_	-0.12	13783	0.06	28320	0.18
8	1_PW	-0.06	27915	0.11	14188	0.16
9	CVC	0.05	24978	-0.08	17125	-0.13
10	cor_	-0.03	22082	0.03	20021	0.06
11	stp_	-0.01	34679	0.05	7424	0.07
12	_hiV	0.00	32731	0.00	9372	0.00
13	_glt	0.01	41258	-0.40	845	-0.41
14	_vce	0.15	13138	-0.07	28965	-0.22
15	_cor	0.01	21479	-0.01	20624	-0.02
16	asp_	0.02	40772	-0.48	1331	-0.50
17	VC	0.00	40255	0.08	1848	0.08
18	_tns	0.01	40708	-0.34	1395	-0.35
19	_fri	0.01	39080	-0.11	3023	-0.12
20	fri_	0.02	38924	-0.23	3179	-0.25

Partition ∅ = mean and size of partition when feature is ∅.
Partition 1 = means and size of partition when feature is 1.

Ranking	Feature	Partition ∅		Partition 1		Diff
		Mean	Size	Mean	Size	
21	_dor	-0.01	29433	0.03	12670	0.04
22	hiV_	0.02	34643	-0.09	7460	-0.11
23	NUC	-0.05	23032	0.05	19071	0.10
24	tns_	0.00	40675	-0.13	1428	-0.14
25	AP_2	-0.01	37450	0.04	4653	0.05
26	PW_2	0.00	38134	-0.03	3969	-0.04
27	2_UTT	0.02	41028	-0.71	1075	-0.73
28	lat_	-0.01	40761	0.21	1342	0.22
29	_V	0.03	23285	-0.04	18818	-0.07
30	_asp	0.00	40847	-0.10	1256	-0.10
31	PW_m	0.00	40519	-0.04	1584	-0.04
32	1_IP	0.00	39956	-0.08	2147	-0.08
33	_lab	-0.01	35048	0.03	7055	0.04
34	_stp	0.01	35039	-0.04	7064	-0.05
35	fla_	0.00	40949	0.06	1154	0.06
36	IP_2	0.00	39766	-0.01	2337	-0.01
37	V_	-0.03	23904	0.04	18199	0.07
38	1_UTT	0.00	41041	-0.19	1062	-0.20
39	CODA	0.00	35470	0.02	6633	0.02
40	UTT_2	0.00	40908	-0.05	1195	-0.05

Partition ∅ = mean and size of partition when feature is ∅.
Partition 1 = means and size of partition when feature is 1.

APPENDIX 5

Preference judgements for the 3 models.

M 1 = CART model; M 2 = Sums-of-products model; M3 = ETRI model

P-value: Probability value

Clarity level						
	M 1	M 2	M 1	M 3	M 2	M 3
1	1	0	1	0	0	1
2	0	1	0	1	0	1
3	1	0	1	0	1	0
4	1	0	0	1	0	1
5	1	0	1	0	0	1
6	0	1	0	1	0	1
7	1	0	1	0	0	1
8	1	0	0	1	0	1
9	0	1	1	0	1	0
10	1	0	0	1	0	1
11	1	0	0	1	0	1
12	0	1	0	1	0	1
13	1	0	0	1	0	1
14	1	0	1	0	0	1
15	1	0	0	1	0	1
16	0	1	1	0	1	0
17	1	0	0	1	0	1
18	0	1	0	1	1	0
19	1	0	1	0	1	0
20	1	0	0	1	0	1
21	1	0	0	1	0	1
22	0	1	1	0	1	0
23	1	0	0	1	0	1
24	1	0	1	0	1	0
25	1	0	0	1	0	1
26	1	0	0	1	0	1
27	1	0	0	1	0	1
28	1	0	0	1	0	1
29	1	0	1	0	1	0
30	1	0	0	1	0	1
31	1	0	0	1	0	1
32	0	1	0	1	0	1
33	1	0	0	1	0	1
34	1	0	0	1	0	1
35	1	0	0	1	0	1
36	1	0	0	1	0	1

Clarity level						
	M 1	M 2	M 1	M 3	M 2	M 3
37	1	0	1	0	1	0
38	1	0	0	1	0	1
39	1	0	0	1	0	1
40	1	0	1	0	1	0
41	1	0	0	1	0	1
42	1	0	1	0	1	0
43	1	0	0	1	0	1
44	1	0	0	1	0	1
45	0	1	0	1	0	1
46	1	0	0	1	0	1
47	0	1	1	0	1	0
48	1	0	0	1	0	1
49	1	0	0	1	0	1
50	1	0	1	0	1	0
51	0	1	0	1	0	1
52	0	1	1	0	1	0
53	0	1	1	0	1	0
54	0	1	1	0	1	0
55	1	0	0	1	0	1
56	1	0	0	1	0	1
57	1	0	0	1	0	1
58	1	0	1	0	1	0
59	1	0	0	1	0	1
60	1	0	0	1	0	1
61	1	0	0	1	0	1
62	1	0	0	1	0	1
63	0	1	1	0	1	0
64	1	0	0	1	0	1
65	0	1	0	1	0	1
66	1	0	0	1	0	1
67	0	1	1	0	1	0
68	1	0	0	1	0	1
69	1	0	1	0	1	0
70	1	0	0	1	0	1
71	0	1	0	1	0	1
72	0	1	0	1	0	1

Clarity level						
	M 1	M 2	M 1	M 3	M 2	M 3
73	1	0	1	0	1	0
74	0	1	0	1	0	1
75	1	0	0	1	0	1
76	0	1	1	0	1	0
77	1	0	0	1	0	1
78	1	0	1	0	1	0
79	1	0	0	1	0	1
80	1	0	1	0	0	1
81	0	1	1	0	1	0
82	1	0	0	1	0	1
83	1	0	0	1	0	1
84	1	0	0	1	0	1
85	0	1	0	1	0	1
86	1	0	0	1	0	1
87	1	0	0	1	0	1
88	0	1	1	0	1	0
89	0	1	0	1	0	1
90	1	0	0	1	0	1
Total	65	25	29	61	25	65
P-value	p <= 2.97e-05		p <= 0.000973		p <= 2.97e-05	

General preference						
	M 1	M 2	M 1	M 3	M 2	M 3
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	0	1	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	1	0	1	0	0	1
8	1	0	0	1	0	1
9	1	0	0	1	0	1
10	0	1	0	1	1	0
11	1	0	1	0	0	1
12	1	0	1	0	0	1
13	0	1	0	1	1	0
14	1	0	1	0	0	1
15	1	0	1	0	0	1
16	1	0	1	0	0	1
17	1	0	1	0	0	1
18	1	0	1	0	0	1
19	1	0	1	0	0	1
20	1	0	1	0	0	1
21	1	0	1	0	0	1
22	0	1	0	1	1	0
23	1	0	1	0	0	1
24	1	0	1	0	0	1
25	0	1	0	1	1	0
26	1	0	1	0	0	1
27	0	1	0	1	1	0
28	1	0	1	0	0	1
29	1	0	1	0	0	1
30	0	1	0	1	1	0
31	0	1	0	1	1	0
32	0	1	1	0	1	0
33	1	0	1	0	0	1
34	1	0	1	0	0	1
35	1	0	0	1	0	1
36	1	0	1	0	0	1

General preference						
	M 1	M 2	M 1	M 3	M 2	M 3
37	0	1	0	1	1	0
38	0	1	0	1	1	0
39	1	0	1	0	0	1
40	1	0	1	0	0	1
41	0	1	0	1	1	0
42	1	0	1	0	0	1
43	1	0	1	0	0	1
44	1	0	1	0	0	1
45	1	0	1	0	0	1
46	0	1	0	1	0	1
47	1	0	1	0	0	1
48	1	0	1	0	0	1
49	0	1	0	1	0	1
50	0	1	0	1	1	0
51	1	0	1	0	0	1
52	1	0	1	0	0	1
53	0	1	0	1	1	0
54	1	0	1	0	0	1
55	0	1	0	1	1	0
56	1	0	1	0	0	1
57	1	0	1	0	0	1
58	1	0	1	0	0	1
59	1	0	1	0	0	1
60	1	0	0	1	0	1
61	1	0	1	0	0	1
62	1	0	1	0	0	1
63	0	1	0	1	0	1
64	0	1	0	1	1	0
65	0	1	0	1	0	1
66	1	0	1	0	0	1
67	0	1	0	1	1	0
68	1	0	1	0	0	1
69	0	1	0	1	1	0
70	1	0	1	0	0	1
71	0	1	0	1	1	0
72	0	1	0	1	1	0

General preference						
	M 1	M 2	M 1	M 3	M 2	M 3
73	1	0	1	0	0	1
74	1	0	1	0	0	1
75	0	1	0	1	0	1
76	1	0	1	0	0	1
77	0	1	0	1	1	0
78	1	0	1	0	0	1
79	0	1	0	1	1	0
80	1	0	1	0	0	1
81	1	0	1	0	0	1
82	0	1	0	1	1	0
83	1	0	1	0	0	1
84	0	1	0	1	1	0
85	1	0	1	0	0	1
86	1	0	1	0	0	1
87	0	1	0	1	1	0
88	1	0	1	0	0	1
89	0	1	0	1	1	0
90	1	0	1	0	0	1
Total	57	33	53	37	28	62
P-value	p <= 0.0149		p <= 0.113		p <= 0.000438	