Teaching a vocal tract simulation to imitate stop consonants

Mark Huckvale[1] & Ian Howard[2]

[1]Phonetics & Linguistics, [2]Institute of Neurology
University College London, London, U.K.
m.huckvale@ucl.ac.uk, i.howard@ion.ucl.ac.uk

Abstract

The imitation of spoken stop consonants by an articulatory synthesizer using only general learning principles addresses significant issues in speech inversion and speech acquisition. Stop consonants are relatively large, complex acoustic events resulting from discrete articulations, so inversion based on the use of small time windows or based on the minimisation of average articulatory error across multiple places of articulation will not provide a satisfactory solution. This paper explores the effect of variation in inversion window size and the use of smoothing constraints on the quality of imitation of the stops [b], [d] and [g]. However good results are only obtained when inversion is supplemented by a phonetic labelling performed over a large time window. This source of additional phonetic information allows inversion to exploit different discrete gestures for the different places of articulation. The results demonstrate the importance of a phonological layer of perceptual analysis prior to imitation and speech acquisition.

1. Introduction

In the scientific study of speech production, speech imitation is a variant of the speech inversion problem with different goals and applications. Whereas speech inversion seeks to find a set of articulatory parameters of a vocal tract that would have generated a given acoustic signal, speech imitation seeks to find a possible output of a vocal tract that best matches a given acoustic signal. Speech inversion is a tool useful to investigate sound production in the vocal tract, while speech imitation is a way of studying the learning system that underlies spoken language acquisition.

The significant differences between speech imitation and speech inversion include: (i) that exact imitation may not be possible, so that the system must seek to produce the best imitation using the resources available, (ii) that the criterion for success is measured in the acoustic domain, not the articulatory domain, and (iii) that imitation does not rely on any privileged access to the articulation used by the target speaker. For imitation to have relevance to human infant acquisition, it must be built solely on access to a (simulated) vocal tract, a system for auditory analysis and general learning principles.

Solutions to the speech inversion problem have generally fallen into three categories: a search through a large codebook of articulatory to acoustic mappings, e.g. [1]; a constrained mathematical solution to the inversion of the articulatory sound generation function, e.g. [2]; or the use of trainable mappings between measured articulatory and acoustical parameters, e.g. [3]. On the whole these methods are not suitable for imitation, either because they rely on privileged access to correct articulations, or because they use cognitively unrealistic mathematical analysis.

The problems of imitation are also intimately related to issues of phonological development. As a child learns to speak, so there are changes in his/her perceptual system in terms of sensitivity to acoustic differences within and across phonological categories. These categories, in turn, appear to influence the inventory of articulatory gestures available for production. The acquisition models of Guenther [4] and Bailly [5] exploit the link between discrete phonological categories in perception and production. However there is still much to be learned in terms of how the discrete categories are learned from audio signals, what determines which categories are used, and how speaking and listening interact. The speech imitation problem is a convenient framework in which to investigate the problems facing an infant learner. We try to keep our solutions within the bounds of what is logically and cognitively plausible. We aim to use fairly realistic articulatory synthesis and auditory analysis of real sounds. We expect our solutions to be sensitive to the properties and deficiencies in the articulatory and auditory processing systems in an analogous way to human infant learners.

2. The problems of imitating stop consonants

In this study we concentrate on the imitation of the stop consonants [b] [d] and [g] since this simple task highlights issues that are significant to both speech imitation and speech inversion. In particular, the estimation of articulatory synthesizer parameters from an acoustic recording such as [əbədəgə] suffers from the following problems:

a. The stops are relatively large events, extending over 100ms of signal, so any analysis needs to accumulate evidence over a wide time window
b. The central silent region of the stops is the same for each place of articulation, causing a one-to-many acoustic to articulatory mapping.
c. The learning of an "average" stop is not a good solution since an average of the articulations for [b] [d] and [g] may have neither the place nor manner of a stop.
d. The learning of a "partial" stop without a complete closure is not a good solution as this may result in the realisation of an approximant or fricative.

Thus imitation of stop consonants given only an acoustic target is a difficult task for an articulatory synthesizer. But rather than seek *ad hoc* engineering tricks to get the best inversion, we are interested in how well a general purpose learning scheme performs on this problem, and where it

fails. In this way we hope the analysis will provide insights into more general aspects of speech acquisition.

In this paper we use supervised learning to train an inverse model to control an articulatory synthesizer. We show how adjustments to the configuration and training of the model affect its performance on the imitation of the test phrase [əbədəgə]. We investigate:

a. How accuracy of imitation is related to the size of the acoustic and articulatory windows (section 4),
b. Whether "oracle" labelling of place of articulation improves learning (section 5),
c. Whether imitation is improved by a combination of phonetic labelling and inversion (section 6).

Section 3 provides background technical details.

3. Research environment

The articulatory synthesizer was developed from the design by Maeda [6] as distributed in a MATLAB version [7]. The synthesizer is controlled by one jaw parameter, three tongue position/shape parameters, and two lip parameters. A version of the LF model [8] was used as a voice source and was controlled by fundamental frequency and glottal area parameters. For more information see the project web site [9]. The articulatory parameters were low-pass filtered at 20Hz and sampled at 100 frames/s. The output sampling rate was 20,000 samples/s.

The acoustic analysis was performed using a 26-channel auditory filterbank delivering energies in 26 bark-scaled channels every 10ms. This was accompanied by a "voicing degree" track which gave a probability of voicing and by a fundamental frequency track. Analysis was performed by the `voc26`, `vdegree` and `fxrapt` programs of SFS [10]. The acoustic parameters were further pre-processed before presentation to the pattern classifier. The mean slope of the auditory spectrum in each 10ms frame was subtracted from the filterbank energies and added as two further parameters, to make 30 parameters in total. All parameters were then normalised to zero mean and unit variance using a long spoken passage.

The pattern recognition technique used to implement the inverse model was a conventional feed-forward multi-layer perceptron with linear output units trained by back propagation. Training data with matched articulatory and acoustic parameters was generated by babbling, that is by random variation of articulator parameters. The statistics of the babbling were controlled to simulate the steady-state and transitional durations of human speech.

The artificial target signal (Fig 3 top) was generated by the Maeda synthesizer using hand-crafted parameters unrelated to those use to generate the babble training set. The natural target signal (Fig 9 top) was recorded from an adult male speaker.

4. Baseline performance

Using 100s of babble, networks with 50 hidden units were trained to map acoustic parameters back to articulatory parameters for a number of different input and output window sizes. Figure 1 shows RMS articulatory error on the artificial target signal as a function of input acoustic window size for three different output articulatory window sizes.

Results show best performance with input windows around 50-70ms, with performance decreasing as the window size increased above 70ms. Small benefits are gained from the output of 30ms or 50ms of articulatory parameters per input window rather than just a single 10ms frame (output windows are overlapped and averaged across the signal).

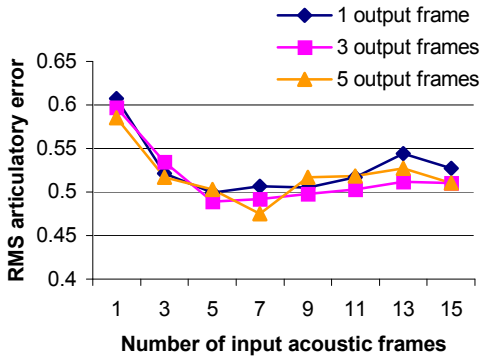

Figure 1. RMS articulatory error on artificial target as a function of network input and output window size.

Small improvements in performance can also be gained if the inverse model is penalised for generating articulatory parameters which jump in value from frame to frame. To implement this the training error E on an output unit becomes

$$E(t) = T(t) - O(t) + s.(O(t-1) - O(t))$$

where T(t) is the target at time t, O(t) is the output at time t, and s is the smoothing coefficient. The effect of the addition of a smoothing constraint is shown in Figure 2.

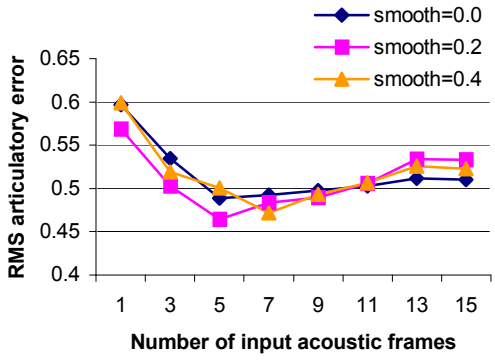

Figure 2. RMS articulatory error on artificial target phrase for a network with 30ms output window as a function of input window size and smoothing coefficient.

The use of a smoothing constraint makes a small but significant improvement in RMS error. The best output was given by a model with 50ms of acoustic input and 30ms of articulatory output using a smoothing coefficient of 0.2. The original test signal and the imitated test signal generated by the best performing inverse model are shown in Figure 3.

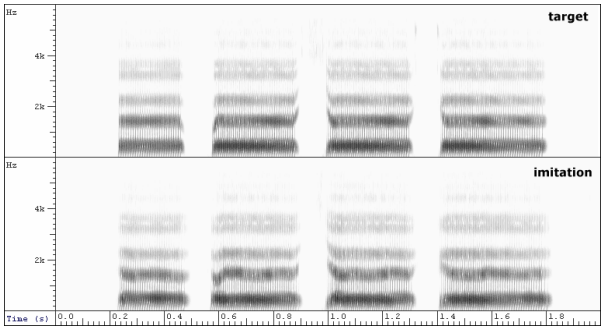

Figure 3. Top – artificial target phrase [əbədəgə], bottom – direct imitation using best inverse model.

Although the spectrographic pattern in Figure 3 looks quite good, the articulation of the stops themselves is far from convincing, as can be seen by a comparison of the articulatory parameters generated by the model and the parameters used to generate the test phrase. See Figure 4.

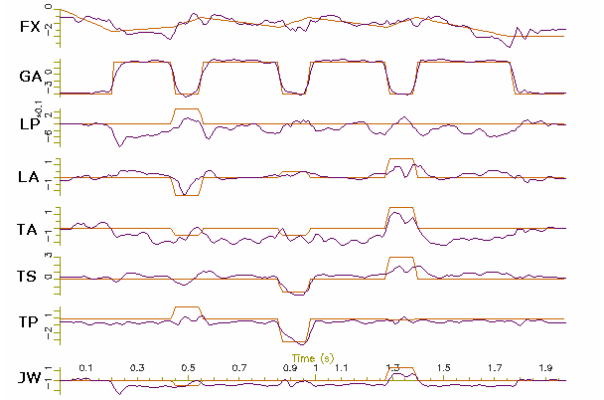

Figure 4. Comparison between correct and inverted articulatory parameters for the artificial test phrase. From bottom: jaw height, tongue position, tongue shape, tongue apex, lip area, lip protrusion, glottal area, fundamental frequency.

Significantly, Figure 4 shows an incomplete lip closure in [b] and an incomplete tongue raising gesture in [g]. Furthermore, similar lip protrusion gestures are seen in both [b] and [g]. On the other hand the tongue tip movement in [d] is handled quite well, possibly because tongue tip approximation causes small amounts of alveolar friction that makes the alveolar stop distinctive. It is important to note that the visible silent gaps in Figure 3 were caused by larynx adjustments rather than by oral closures.

5. Test of ideal performance

To determine how much of the inadequacy of the inverse model was due to limitations of the machine learning techniques (i.e. training regime and network structure) rather than to the problem itself, an "oracle" training method was tested in which each training data vector was increased in size to include binary features representing the presence of a bilabial stop closure, an alveolar stop closure or a velar stop closure. Although such perfect labels could not occur in real imitation, they make the acoustically identical regions distinct to the inverse model so that it should then be able to generate distinct articulatory gestures for each place of articulation

Figure 5 shows that this was indeed the case. With the benefit of place labelling, the correct articulations are almost perfectly recreated by the inverse model. The conclusion is that it is the one-to-many mapping arising from the use of too little context that is the problem, not the inadequacy of the machine learning. The labels allow the inversion to make use of a larger context without requiring a larger input window.

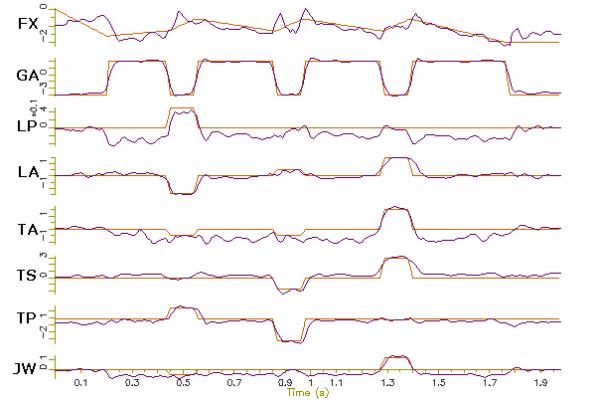

Figure 5. Improved inversion possible from perfect place labelling in the training and test data.

6. Combined labelling and inversion

Since the availability of stop place labels makes a significant improvement to the inverse mapping, this suggests a two phase inversion strategy, where phase one estimates the place labels, while phase two performs the inversion. See Figure 6.

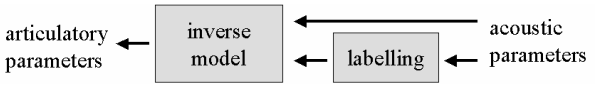

Figure 6. The labelling system augments the acoustic data input to the inverse model.

The first phase place labelling network is trained using labelled babble data to generate the [b], [d], [g] label tracks from the acoustic signal using a wide symmetric window of 150ms. The second phase inversion network takes the acoustic frames and the label tracks with a narrower window of 50ms and generates 30ms overlapping windows of articulatory data as before. The output of the place labelling and the improved imitation is shown in Figure 7 (cf. Figure 3), with the articulatory parameters shown in figure 8 (cf. Figure 4).

Clear improvements in the imitation can be seen in the articulatory tracks and also heard in the synthesizer output. Greater articulator movement in [b] and [g] leads to more stop-like formant transitions, which make them sound more convincing.

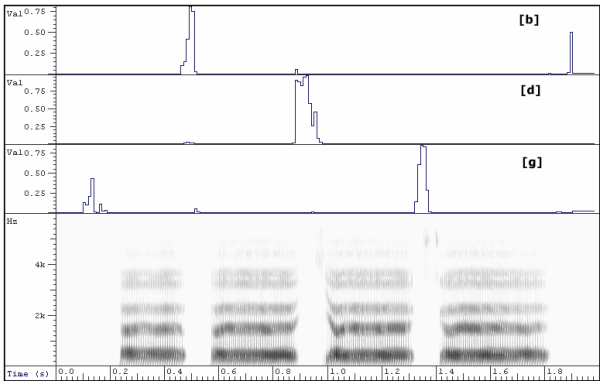

Figure 7. Recognised place labels and improved imitation of [əbədəgə] using recognised labelling.

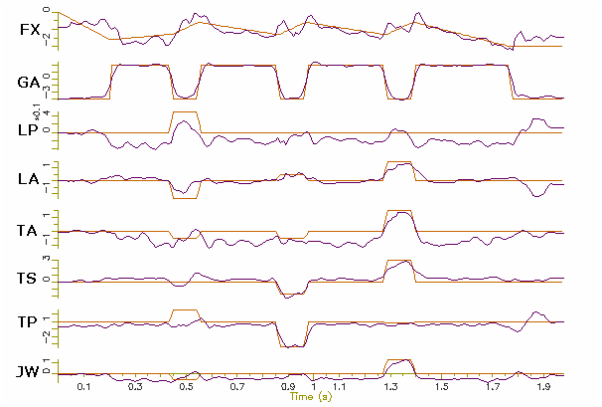

Figure 8. Improved inverted articulatory parameters for the test phrase using recognised place labelling as well as acoustic input.

Finally the whole process was repeated on the natural target speech signal, shown in Figure 9. The imitation of the natural target is worse than for the artificial target, but retains some attributes of the target stop consonants. The formant transitions for [d] are actually exaggerated and more like the transitions used on the artificial data.

7. Conclusions

The imitation of stop consonant articulations from an acoustic signal is a simple task that is hard to perform well and which highlights important issues in speech inversion and speech imitation. Large time windows are required to identify each consonant so that categorically distinct articulatory gestures can be performed. However large windows are difficult to train because of the large number of degrees of freedom in the model, the amount of training data required for good estimation of the model, and the interference caused by irrelevant data in the input pattern.

The results here show that a two phase strategy that first labels and then performs articulatory inversion using both the acoustic data and the labels as input allows the exploitation of a larger context. Since similar networks are used to implement both pattern recognition tasks, it is possible to foresee a training strategy in which both networks are trained together, to seek to minimise the acoustic error of the imitation. This approach is related to "distal" learning [11] and to methods for the inference of underlying phonetic parameters [12].

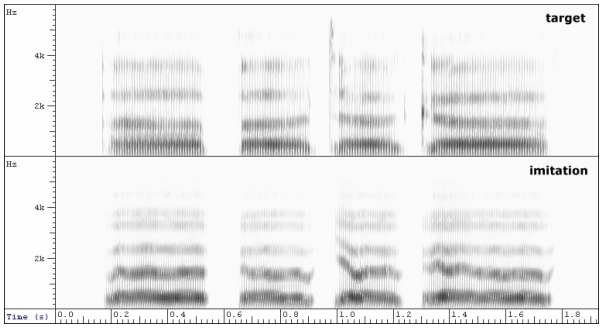

Figure 9. Natural target phrase and imitation using the combined method

Infant learners could discover the utility of discrete categories in two ways: either because they help explain the statistics of the auditory signal, or because they help guide discrete articulatory gestures. Speech imitation can make a contribution to settling this long-standing debate about the interaction of perception and production.

8. References

[1] Schroeter, J., Sondhi, M.M., "Techniques for estimating vocal-tract shapes from the speech signal", *IEEE Trans. Speech and Audio Processing 2 (1994), p133-150.*

[2] Sorokin, V.N., Leonov, A.S., Trushkin, A.V, "Estimation of stability and accuracy of inverse problem solution for the vocal tract", *Speech Communication 30 (2000), p55-74.*

[3] Rahim, M.G., Goodyear, C.C., Kleijn, W.B., Schroeter, J., Sondhi, M.M., "On the use of neural networks in articulatory speech synthesis", *J. Acoust. Soc. Am. 93 (1993), p1109-1121.*

[4] Guenther, F.H., "A neural network model of speech acquisition and motor equivalent speech production". *Biological Cybernetics 72 (1994), p43-53.*

[5] Bailly, G., "Learning to speak. Sensori-motor control of speech movements", *Speech Communication, 22 (1998), p251-267.*

[6] Maeda, S., "Compensatory articulation during speech", in *Speech production and speech modelling* (W. J. Hardcastle and A. Marchal, eds.), *Kluwer Academic Publishers, Boston, (1990), p131-149.*

[7] Ghosh, S.S., *VTCALCS for MATLAB,* http://speechlab.bu.edu/VTCalcs.php

[8] Fant G., "Glottal flow: models and interaction", *Journal of Phonetics, 14 (1986), p393-399.*

[9] Huckvale, M.A. *Speech Synthesis Research*, http://www.phon.ucl.ac.uk/home/mark/synthesis/

[10] Huckvale, M.A., *Speech Filing System*, http://www.phon.ucl.ac.uk/resource/sfs/

[11] Howard, I., Huckvale, M., "Training a vocal tract synthesizer to imitate speech using distal learning", *Proc. InterSpeech 2005, Lisbon*.

[12] Richards, H., Bridle, J., "The HDM: a segmental hidden dynamic model of coarticulation", *Proc. ICASSP'99 (1999), p357-360.*