Proceedings of the Institute of Acoustics

FEATURE DRIVEN FORMANT SYNTHESIS

Jon Iles and William Edmondson

The University of Birmingham, School of Computer Science, Edgbaston, Birmingham. B15 2TT.

1. INTRODUCTION

The aim of the work described in this paper has been to develop a hybrid speech synthesis strategy that combines the potential quality available from formant synthesis techniques with the advantages offered by articulatory control of the synthesis process. The investigation leading to this paper was prompted by the lack of high quality natural sounding synthetic speech produced by rule. It is our belief that there are two factors responsible for slow progress in this area. The first factor is the framework within which synthetic speech is generated from textual input. Our approach to this problem is further discussed in [4]. The second factor we have identified is the method employed to control the process of converting from a parametric representation of the required speech to its acoustic realization. This is the area covered by this and previous papers [10, 11].

2. BACKGROUND

It is commonly agreed that in the long term articulatory synthesis will provide very high quality natural sounding synthetic speech. If an articulatory model is constructed that accurately copies a real vocal tract, with a control strategy that simulates the natural behaviour of the articulators - then this model will be capable of synthesizing close copies any utterance produced using the original vocal tract. Articulatory control over the synthesis process allows more natural manipulation of the resulting speech, and thus the generation of simple-yet-elegant rules to model coarticulation and other effects seen in natural speech. However, considering the early stages of work in many current investigations into articulatory models this goal will require a considerable investment in both time and effort.

The main problems facing current articulatory control strategies are the gathering of data for estimation of the vocal tract shape, and control of the resulting model. To construct a convincing articulatory model measurements of the physical attributes of a real vocal tract are required. In the past this type of data has been obtained using x-ray photography and cinematography. Data gathered in this manner are two dimensional, typically tracings of the shape of the vocal tract along the midsagittal line [2, 18]. This two dimensional data has to be converted into approximate cross-sectional area measurements. More recent work in this area has utilized Magnetic Resonance Imaging (MRI) to allow accurate cross-sectional measurements of the vocal tract to be taken [5, 7]. MRI scanners do not yet offer near real-time scanning so large amounts of accurate cross-sectional area data are not presently available.

Even when good measurements are available, a good control strategy must be employed to ensure that the trajectories of the articulators follow those of the articulators being modelled. Work has been proceeding in this area based on direct measurement of articulatory movement [19], copy synthesis using codebooks [6], parameter optimization [20] and finally mapping the acoustic speech signal back to articulator trajectories (e.g. [19, 21]). All of the methods above, bar those using direct measurements, rely on approximation and solving the many-to-one mapping problem that arises when considering acoustic-to-articulator conversion.

Now consider the work that has been undertaken using formant synthesizers. Current formant-based synthesis offers some of the highest quality synthetic speech available. Copy synthesis of natural utterances in the past has demonstrated that synthetic speech can be almost indistinguishable from the original natural utterance [8, 9]. This quality of synthetic speech has yet to be produced as a result of formant based synthesis-by-rule. We believe that the reason for this lack of success with synthesis-by-rule is the control parameters themselves. Formant synthesizers are controlled using parameters that can be directly measured from representations of the speech signal. These parameters are seen to be complex when compared to the simple articulatory activity that produced the acoustic signal from which they were measured. They are therefore not necessarily the optimum choice for speech representation. Many subtle features of natural speech modelled using acoustic parameters require the application of sets of ad-hoc rules during the synthesis-by-rule process. These rules have been developed by observation of natural speech spectra; they do not attack the problem of simulating the articulatory activity that produced the original effect. Thus a simple articulatory gesture resulting in a given feature of natural speech may require any number of ad-hoc rules to simulate in the acoustic domain.

Our proposed solution to this problem was to develop a hybrid synthesis strategy that allows speech to be specified in articulatory terms, but utilizes formant synthesis to produce the required speech output. Similar approaches to this have been taken in the past. Some articulatory synthesizers have utilized a vocal tract model to allow calculation of speech spectra, followed by a formant synthesizer to realize this acoustically [2, 15]. This has allowed manipulation of the synthetic speech both in the articulatory and the acoustic domain. This approach still relies on having an accurate vocal tract model. Our approach differs from this as we acknowledge our inability to produce a vocal tract model of the desired accuracy. Instead we have concentrated on producing a very simple approximation. Other approaches have utilized a similar idea to simplify the control parameters required for a formant synthesizer. Stevens and Bickley [22] used a hybrid set of articulatory and acoustic parameters to simplify control of a Klatt formant synthesizer [13] from 40 parameters to 10 parameters. The parameters Stevens and Bickley chose are motivated by the desire to simplify Klatt's control parameters by reference to constraints the vocal tract places on permissible configurations. The work we have undertaken takes a different approach to this as we are not strictly adhering to an accurate model of the vocal tract. We have attempted to produce a linguistically motivated model.

The model we describe in the following section is controlled using a set of quasi-articulatory features. These features have been derived from the notion of distinctive features used by linguists and phonologists to describe speech. In some respects this could be considered to be bridging the gap between the phonologists' view of speech as exemplified in "The Sound Pattern of English" [1], and the acoustic-phonetic realization of speech familiar to speech synthesis researchers.

3. BASIC REQUIREMENTS

Work to develop a mapping between quasi-articulatory features and synthesizer control parameters began by defining the nature of the articulatory features required and the type of formant synthesizer to be used. The features chosen are briefly described in figure 1. This list was produced after consulting a number of phonetics and linguistics text books. They represent the minimal set of features required to differentiate all of the phones used in the English language. The features themselves are based on distinctive features, although some of them have been given continuous coefficients rather than binary values. This step takes us from a static approximation of articulation to a representation that allows us to specify the dynamics of articulation.

A formant synthesizer based on the Klatt cascade-parallel model [12, 13] was chosen as the target for this work. In trials the Klatt architecture proved to be flexible enough to produce close synthetic copies of natural utterances. Only the parallel branch was used for the remainder of the investigation as it was found to be easier to closely match natural utterances with the wider range of control the parallel branch offers. Attempts to match the spectra of vowel sounds using the cascade branch was less fruitful.

Name	Type	Description
High	continuous	Used to represent tongue height. Values represented as a percentage between. 0% high indicates the tongue in its lowest position. 100% high represents the tongue in its highest position.
Back	continuous	Used to represent tongue back-front position. Values are represented as a percentage. 0% back indicates that the tongue is in its most forward position, 100% back indicates that the tongue is back as far as it will go in the mouth.
Round	continuous	Used to represent lip rounding. Values are represented as a percentage. 0% round indicates that the lips are spread. 100% round indicates that the lips are fully rounded.
Tense	continuous	Used to represent tongue tension. Values are represented as a percentage. 0% tense indicates that the tongue is lax. 100% tense indicates full tongue tension.
Labial	binary	Indicates that a sound is articulated at the lips.
Coronal	binary	Indicates that a sound is articulated by raising the tongue blade towards the hard palate.
Strident	binary	Indicates that during articulation, frication is caused by the air stream coming against the teeth or hard alveolar ridge.
Anterior	binary	Indicates that articulation takes place at, or in front of the alveolar ridge.
Voiced	continuous	Specifies the degree of voicing, represented as a percentage value. 0% voiced indicates no voiced excitation is present, 100% voiced indicates that full voiced exitation is present.
Fricated	continuous	Specifies the degree of frication, represented as a percentage value. 0% fricated indicates no fricated excitation is present, 100% fricated indicates full fricated excitation is present.
Aspirated	continuous	Specifies the degree of aspiration, represented as a percentage value. 0% aspirated indicates that no aspiration is present, 100% aspirated indicates that full aspiration is present.
Nasality	continuous	Specifies the degree of nasality, represented as a percentage value. 0% nasality indicates that no nasality is presnet, 100% nasality indicates that full nasality is present.

Figure 1: Table of features used by FDFS.

The remaining parameter to be defined was the high-level representation of the speech signal used to drive the synthesis process. A phonetic representation was chosen, closely related to the "lower phonetic" format used by the JSRU text-to-speech system [14]. This representation allowed us to generate phonetic parameters from English text utilizing the JSRU system, after which the parameters could be hand-edited to match duration and pitch values measured from natural speech. Using an input based on phonetic segments implies that this work differs little from traditional segmental approaches to speech synthesis. We would argue that this is not the case. At this point driving the synthesis technique using a segmental representation is an experimental convenience. In further work we would hope to demonstrate non-segmental synthesis similar in concept to YorkTalk [3, 16]. This is discussed further in [4].

4. CONSTRUCTING THE MODEL

The process by which we produced a mapping between articulatory features and synthesizer control parameters was simply by correlating two sets of data. We constructed a table of all of the phones used by the JSRU system, and specified them in terms of articulatory features.

These data were gathered from standard text books. For the continuous variables, approximations were made for the required values by reference to articulation diagrams and descriptions from a number of sources. For the same set of phones, a specification in terms of synthesizer control parameters was also constructed. The relevant values were initially measured by hand from spectrograms of natural speech, and the results improved incrementally by copy synthesis and comparison with the original utterance. At a later stage comparison of spectra produced by linear prediction for both natural and copy-synthesized phones allowed a closer match to be made.

Correlation of the two sets of data was achieved by using multiple regression analysis. For simplicity only data for vowels and glides were used to construct the first model. An assumption regarding the nature of glides was made to expand the available data set. It was assumed that at the mid-point between the start and end point of a glide, all values (synthesizer parameters and articulatory features) are at the mid point between their respective start and end points. Observation indicated that this assumption was true in all but a handful of cases. In these exceptions f3 tended to show a marked drop towards f2, and away from its final end point. These data were included in the analysis and a more accurate model resulted.

Formant	Frequency R^2	Amplitude R^2	Bandwidth R^2
1	97.8	78.3	–
2	94.9	89.2	91.2
3	74.6	77.4	62.8
4	80.0	47.8	50.6
5	34.4	82.8	72.2
6	57.9	45.3	90.9

Figure 2: R^2 values as an indication of model accuracy

In quantative terms, the accuracy of the model can be measured using the R^2 statistic which indicates how much of the observed data is explained by the model. The table in figure 2 gives R^2 values for formant frequencies, amplitudes and bandwidths calculated by the model. The value for b1 is left blank as a constant value is used for this parameter. Note that the accuracy of the model is lower for the higher, less perceptually important formants. This may be a reflection of the difficulty of making accurate measurements for these values.

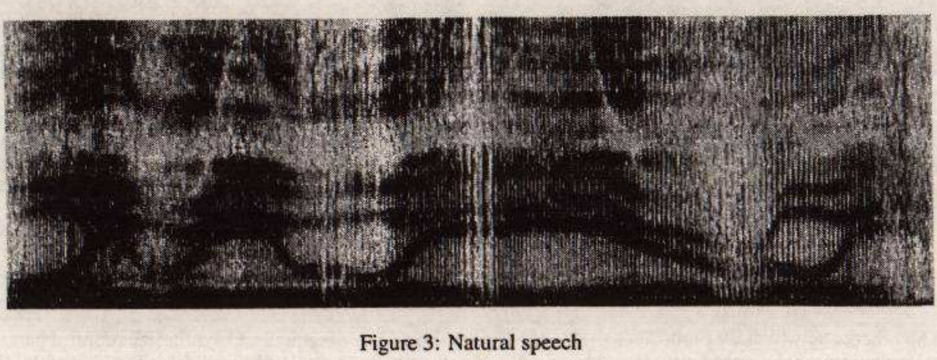

Figure 3: Natural speech

The result of the analysis described in the previous paragraphs was a model capable of producing vowel sounds from articulatory feature specifications. To construct a complete synthesis-by-rule system from this model three further steps were taken. The first step was to define how this model could be used to synthesize other classes of sounds, such as stops and fricatives. This was achieved by making the assumption that all phones are articulated as vowels, but with some phones having a structure "overlaid" on top of the vowel articulation. For example, in the case of fricatives, suitable amplitude and bandwidth parameters are substituted for those generated by the model so that when fricated excitation is used rather than voiced excitation, the correct spectra results. The binary place and manner features described in figure 1 are used to look up the appropriate values to be "overlaid" on the results from the vowel model.

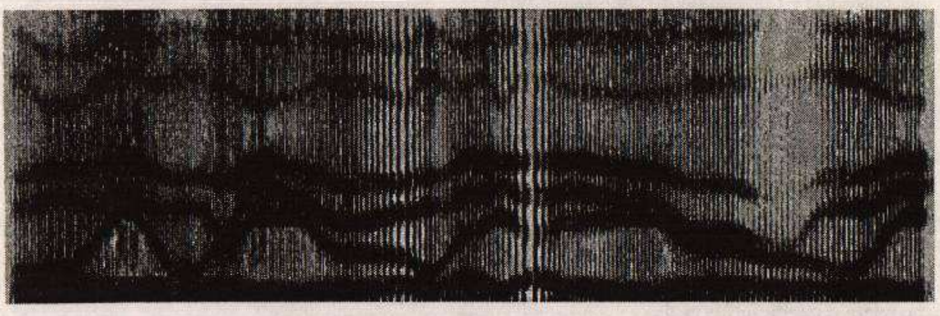

Figure 4: Synthetic speech

The second step taken was to determine how to interpolate between targets for the continuous-valued articulatory features. After investigating different methods of parameter interpolation, a cosine based interpolation function was chosen. This gave smooth transitions between idealized targets, and modelled the acceleration and deceleration of articulators during their transition between targets.

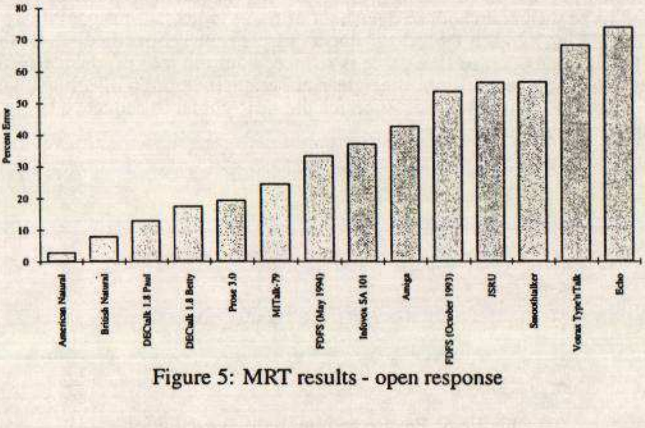

Figure 5: MRT results - open response

The final step was to define a set of rules governing the trajectories of the articulators between phones, their relative timings and the timing of fricated and voiced excitation. Effectively these rules implement segment level coarticulation. Segment level coarticulation is being modelled at this stage for demonstration purposes. Given a suitable architecture within which FDFS can operate, contextual effects at any given level of abstraction may be modelled. One hundred and ten rules were constructed, each specifying the transition between two segments of given "classes". For example there are rules dealing with fricative to vowel transitions, stop to vowel transitions, vowel to fricative transitions and so on.

The rules described operate on a list of frames, generated from the original phonetic input. Each frame represents 5ms of speech, and each contains a set of articulatory features describing the state of articulation at that point. The rules produce smooth trajectories between phone targets by applying the interpolation function as described. Finally these frames are mapped onto 5ms frames of synthesizer parameters which are used to drive the formant synthesizer. Two spectrograms are given as a comparison of a natural utterance (figure 3), and a synthetic equivalent (figure 4). The phrase is "Why were you away a year Roy?". The phonetic transcription used as input to FDFS was produced using the JSRU system, then hand edited to match the duration and pitch contour of the natural utterance.

Two modified rhyme tests (MRT) have been carried out to provide a rough guide to the intelligibility of FDFS. A similar methodology was used to that described in [17]. This allowed direct comparison of the results obtained with other synthesis systems. The MRT can also be used to provide diagnostic information and hence aid the development of the synthetic speech. The graph in figure 5 is a summary of the error rate of the systems tested using the MRT in open response format. FDFS appears twice on this graph, indicating the intelligibility of the system at two points during its development. The reader should be aware that there are a number of caveats attached to the results of MRT's, and that they only provide an approximate guide to the relative intelligibility of the speech stimuli listed.

5. ADVANTAGES OF FDFS

One of the advantages gained by using articulatory controls to drive the synthesis process is the ability to vary the precision with which the synthetic speech is articulated. Figure 6 attempts to illustrate this. The two graphs detail the trajectory over time of a given articulator (e.g. tongue height). In a segmental view of speech the movement of this articulator could be envisaged as being governed by a number of idealized targets, one per segment. In the left hand graph in figure 6, we see these targets illustrated by the point reached by the articulator at each of the segment centres. Imprecision in articulation (e.g. in rapid speech), could be considered to be the undershoot or overshoot of these targets. To model this we assume that the articulator is still attempting to reach the "ideal" targets as previously specified. In this instance however, the time allowed for this transition to take place is reduced, but the rate of change of the position of the articulator is not modified to reflect this. As the right-hand graph in figure 6 illustrates, a different trajectory for the articulator results. Effectively, articulation for the next phone in sequence begins before the target for the previous phone is reached.

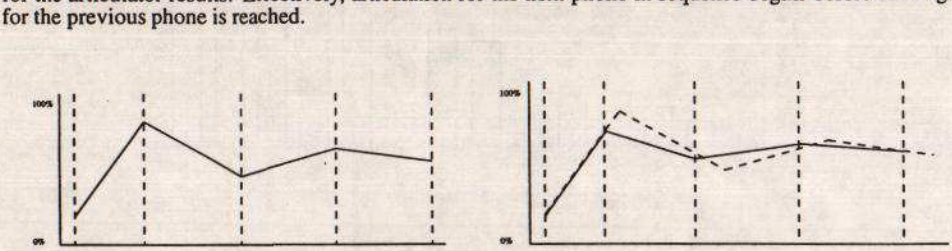

Figure 6: Precise and imprecise articulation

The work is at an early stage. However a number of interesting results have been obtained. First it is possible to use precision of articulation to model rapid speech. A phrase spoken at a normal rate of articulation compared to the same phrase spoken at twice the speed has characteristically different spectral properties. Simply doubling the rate of a synthetic version of the phrase produces an un-natural "over precise" effect. By varying the precision of articulation across the synthetic phrase, its spectral properties can be made to match quite closely those observed in the natural rapid speech.

After observing the effect of modifying precision of articulation, we have come to the preliminary conclusion that precision of articulation may play a large part in the perception of stress and vowel reduction. By increasing the precision of articulation in stressed syllables, and reducing the precision of unstressed syllables we have noted that the expected vowel reduction results and the stressed syllable appears correctly stressed. It must be noted at this point that this work is at a very early stage; further experimental work is required before we can fully demonstrate the precision-stress relationship.

An example of varying precision of articulation is given in figure 7 below. The precision of articulation has been adjusted as follows: at the start of the utterance high precision is used, which reduces rapidly after the first syllables. Further stressed syllables have high precision, the remaining syllables have lower precision. If the spectrogram in figure 7 is compared with those of natural speech (figure 3) and the original synthetic speech (figure 4) it will be noted that the formant transitions in the precision-adjusted synthetic speech are smoother than the original synthetic speech and reflect more accurately the original speech being modelled. Informal listening tests indicate that for rapid synthetic speech, adjusting the precision gives more natural sounding results (and can be taken to extremes resulting in "slurred" almost drunken sounding speech).

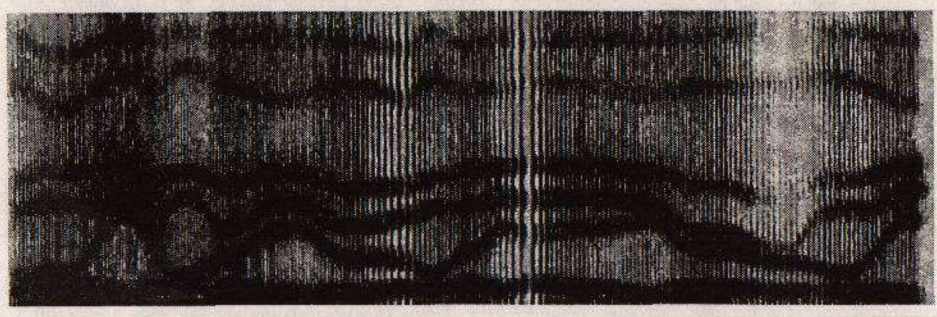

Figure 7: Synthetic speech with precision adjustment

6. SUMMARY

In summary, we have presented a description of the derivation of a quasi-articulatory synthesis strategy. We have demonstrated that it is possible to produce intelligible speech using a crude approximation of articulatory control. We have also demonstrated one of the advantages that this articulatory based approach to synthesis offers when compared to tradition acoustic-domain synthesis: the ability to dynamically modify the precision of articulation of an utterance to enhance its perceived naturalness.

7. ACKNOWLEDGEMENTS

The work described in this paper has been supported in part by Apricot Computers Limited, a subsidiary of Mitsubishi Electric UK Limited, the Science and Engineering Research Council, and GPT Limited.

8. REFERENCES

[1] N. Chomsky and M. Halle. *The Sound Pattern of English.* Harper and Row, 1968.

[2] C.H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, 1976.

[3] J. Coleman. "Synthesis-by-rule" without segments or rewrite rules. In G. Bailly and C. Benoit, editors, *Talking Machines, theories, models and designs.*, pages 43–60. North-Holland, 1992.

[4] W.H. Edmondson and J.P. Iles. Pantome: an architecture for speech and natural language processing. Appears in these proceedings, November 1994.

[5] A.K. Foldvik, U. Kristiansen, and J. Kværness. A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI). In *Proceedings of the European Conference on Speech Technology - EUROSPEECH*, volume 1, pages 557–558. ESCA, 1993.

[6] A.R. Greenwood and C.C. Goodyear. Articulatory copy synthesis using multiple codebooks. In *Proceedings of the Institute of Acoustics*, volume 14, 1992.

[7] A.R. Greenwood, C.C. Goodyear, and P.A. Martin. Measurements of vocal-tract shapes using magnetic-resonance-imaging. *IEE Proceedings-I Communications Speech and Vision*, 139(6):553–560, 1992.

[8] J.N. Holmes. Synthesis of natural-sounding speech using a formant synthesizer. In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication Research*, pages 275–285. Academic, New York, 1979.

[9] W.J. Holmes. Copy synthesis of female speech using the JSRU parallel formant synthesizer. In *Proceedings of the European Conference on Speech Technology - EUROSPEECH*, pages 513–516, 1989.

[10] J.P. Iles and W.H. Edmondson. Control of speech synthesis using phonetic features. In R. Lawrence, editor, *Proceedings of the Institute of Acoustics Autumn Conference on Speech and Hearing*, volume 14, pages 369–373. Institute of Acoustics, December 1992.

[11] J.P. Iles and W.H. Edmondson. The use of a non-linear model for text-to-speech conversion. In *Proceedings of the European Conference on Speech Technology - EUROSPEECH*, volume 2, pages 1467–1470. ESCA, September 1993.

[12] D.H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America*, 67(3):971–995, March 1980.

[13] D.H. Klatt and L.C. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, February 1990.

[14] E. Lewis. A C implementation of the JSRU text-to-speech system. Technical Report TR-89-15, University of Bristol, Department of Computer Science, 1991.

[15] Q. Lin and G. Fant. An articulatory speech synthesizer based on a frequency-domain simulation of the vocal tract. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 57–60, 1992.

[16] J.K. Local. Modelling assimilation in a non-segmental, rule-free phonology. In G.J Docherty and D.R. Ladd, editors, *Papers in Laboratory Phonology II*, pages 190–223. Cambridge University Press, 1992.

[17] J.S. Logan, B.G. Greene, and D.B. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86(2):566–581, 1989.

[18] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1972.

[19] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America*, 92(2):688–700, 1992.

[20] S. Parthasarathy and C.H. Coker. On automatic estimation of articulatory parameters in a text to speech system. *Computer Speech and Language*, 6:37–75, 1992.

[21] M.G. Rahim, C.C. Goodyear, W.B. Kleijn, J. Schroeter, and M.M. Sondhi. On the use of neural networks in articulatory speech synthesis. *Journal of the Acoustical Society of America*, 93(2):1109–1121, 1993.

[22] K.N. Stevens and C.A. Bickley. Constraints among parameters simplify control of Klatt formant synthesizer. *Journal of Phonetics*, 19:161–174, 1991.