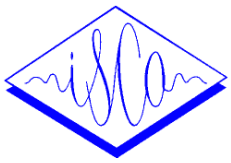

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

QUASI-ARTICULATORY FORMANT SYNTHESIS

Jon Iles and William Edmondson

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

ABSTRACT

This paper describes work leading from discussion of hybrid articulatory/formant synthesis [3, 4]. We briefly outline the problems that we consider are facing both formant and articulatory synthesis techniques, and how we overcome these using a hybrid solution. Construction of a prototype hybrid model is discussed, the intelligibility of the resulting synthetic speech, and the benefits of this type of approach are examined.

1 BACKGROUND

Articulatory synthesis is commonly agreed to be a strategy that will be capable of producing by rule synthetic speech which is almost indistinguishable from natural speech. However, a number of problems are holding back the rapid development of articulatory synthesis strategies. The first problem is obtaining measurements of the physiological parameters of real vocal tracts, for example using X-ray data or Magnetic Resonance Imaging. The second problem is providing control over the articulators in an articulatory model. Work in this area has focussed on recording the trajectories of articulators in a real vocal tract by direct measurement, inverting the articulatory-acoustic mapping, copy-synthesis and parameter optimization.

Formant synthesis techniques have demonstrated their potential quality when used to produce synthetic copies of natural utterances. However, this quality is not apparent in rule-based synthetic speech based on the same formant synthesizer. We believe that this lack of quality can be attributed to the control parameters of formant synthesizers themselves. Formant synthesizers are controlled using parameters that can be measured directly from a speech signal i.e. formant frequencies, amplitudes and bandwidths. Although these parameters are the result of articulatory processes, they may not be the ideal method of manipulating synthetic speech. A combination of a few simple articulatory gestures may result in complex formant patterns. To mimic these and other subtle features of natural speech a number of ad-hoc rules generated by careful observation of speech from several speakers. It would obviously be of benefit if we could specify speech in articulatory terms, but as discussed in the previous paragraph, traditional approaches to articulatory modelling may require considerable effort before they are useful for general work in text-to-speech conversion.

Our 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 (e.g. [6]). 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.

Other approaches have attempted to simplify the control parameters required for speech synthesis. Stevens and Bickley [9] used a hybrid set of articulatory and acoustic parameters to simplify control of a Klatt formant synthesizer 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. Browman and Goldstein (e.g. [1]) use a small set of articulatory gestures as the representation used to drive an articulatory model. The set of gestures chosen encompasses a wide range of more complex manipulations of an underlying articulatory model.

Our approach differs from these as we acknowledge our inability to produce a vocal tract model of the desired accuracy. We have concentrated on producing a very simple approximation to articulatory control, rather than seek direct simplification of the control parameters. We have attempted to produce a linguistically motivated model based on the notion of distinctive features that approximates articulation in a real vocal tract. 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", and the acoustic-phonetic realization of speech familiar to speech synthesis researchers.

2 BUILDING THE MODEL

The construction of the model was a straight-forward process. Two sets of data were required - a specification of a set of phones in terms of synthesizer parameters, and a specification of the same set of phones in terms of distinctive features. The chosen set of phones was derived from the phonetic representation used by the JSRU text-to-speech conversion system [5]. This allowed test data in the form of phonetic transcriptions of English text to be generated automatically and used as input to the model. 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 hope to demonstrate non-segmental synthesis similar in concept to YorkTalk [7], as discussed in [2].

The set of synthesizer parameters required was constructed manually by comparing the spectra of natural speech samples to the spectra of synthetic speech samples, and adjusting the synthesizer parameters by hand until a close match was made. The target synthesizer for this work was one based on the Klatt cascade-parallel formant synthesizer design. For this work only the parallel branch of the synthesizer was used. This decision was taken as a closer match to target vowel spectra could be made using the parallel branch than could be achieved using the cascade branch.

Finally a set of quasi-articulatory features was required. These features were derived from the distinctive feature description used by linguists to characterize the articulation of speech segments. The features chosen are briefly described in Figure 1. This list was produced after consulting a number of phonetics and linguistics text books. Traditionally, distinctive features use binary coefficients, but as the table in Figure 1 illustrates some of the features we use have been given continuous coefficients. This step takes us from a static approximation of articulation to a representation that allows us to specify the dynamics of articulation. The actual values for

these features were taken from a number of linguistics text books.

Correlation of the two sets of data (phones in terms of synthesizer parameters, and phones in terms of articulatory features) was achieved by using multiple regression analysis. For simplicity only data for vowels and glides were used to construct the first model. 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.

Name	Type	Description
High	continuous	Used to represent tongue height. Values represented as a percentage. 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 in its most posterior position.
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.
Fricated	continuous	Specifies the degree of frication, represented as a percentage.
Aspirated	continuous	Specifies the degree of aspiration, represented as a percentage.
Nasality	continuous	Specifies the degree of nasality, represented as a percentage.

Figure 1: Table of quasi-articulatory features

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 experienced in making accurate measurements for these values.

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

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 consonants. This was achieved by making the assumption that all phones are articulated as vowels, with consonants having a structure superimposed over 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 spectrum is produced. The binary place and manner features described in Figure 1 are used to look up the appropriate values to be superimposed over the results from the vowel model.

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 this movement.

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 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 the model 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 "We were away a year". 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.

Modified Rhyme Tests (MRT) have been carried out to provide a rough guide to the intelligibility of the model. A similar methodology was used to that described in [8]. This allowed direct comparison of the results obtained with other synthesis systems. The MRT can also be used to provide some diagnostic information and was used to aid the development of the model. The graph in Figure 5 is a summary of the error rate of the systems tested using the MRT in open response format. The model described in this paper (now named "Feature Driven Formant Synthesis") is labelled FDFS. The reader should be aware that there are a number of caveats attached to the results of MRTs, and that they only provide an approximate guide to the relative intelligibility of the speech stimuli listed.

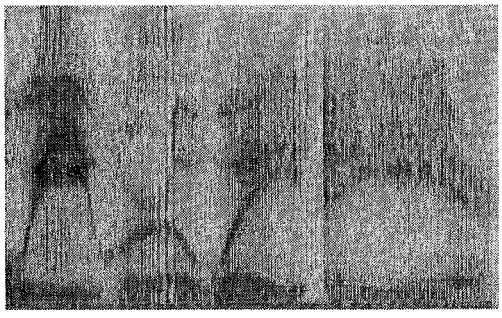

Figure 3: Natural speech

Figure 4: Synthetic speech

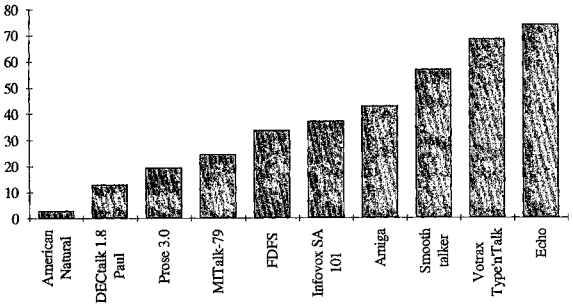

Figure 5: MRT results - open response

3 ADVANTAGES OF QUASI-ARTICULATORY SYNTHESIS

In this section we discuss two advantages that our quasi-articulatory synthesis strategy offers: variable precision of articulation and inversion of the articulatory-acoustic mapping.

3.1 Variable precision of articulation

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 illustrates how this is achieved. Parts (a) and (b) of Figure 6 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 part (a) of Figure 6, we see these targets illustrated by the point reached by the articulator at each of the segment centres. Imprecision in articulation could be considered as the undershoot 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 part (b) of 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.

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. Figure 3 is a phrase spoken at a normal rate of articulation. The synthetic version of this phrase is seen in Figure 4. The same phrase was then spoken at approximately twice the rate of the original phrase. The resulting spectrogram is shown in Figure 7. Generating a synthetic version of this phrase simply by halving the duration of the original phrase does not produce the desired result as illustrated in Figure 8.

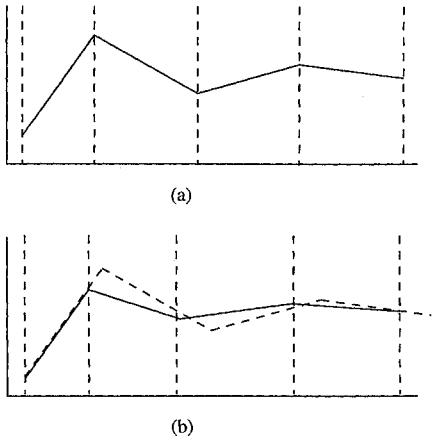

Figure 6: Precise and imprecise articulation

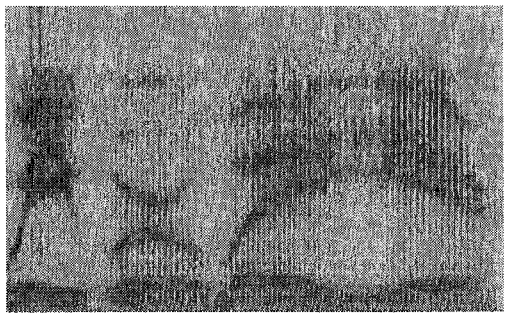

Figure 7: Rapid natural speech

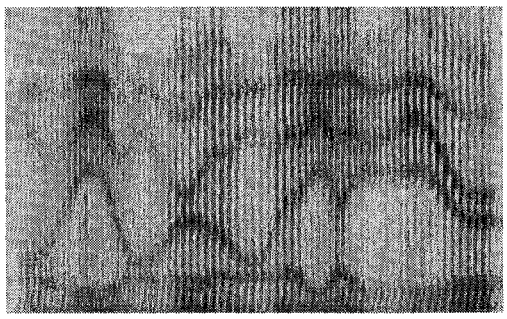

Figure 8: Rapid synthetic speech with no precision adjustment

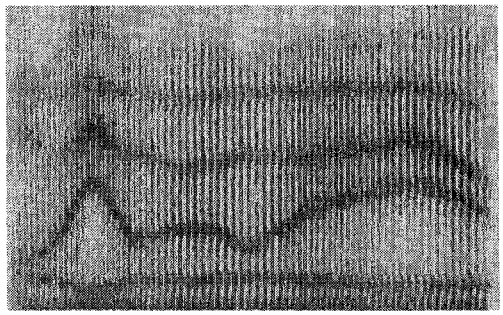

Figure 9: Rapid synthetic speech with precision adjustment

By adjusting the precision of articulation across the synthetic utterance we can immediately see an improvement (Figure 9). This is particularly obvious in the second half of the phrase which matches the natural version quite closely. After some preliminary work with precision of articulation we have noted that there seems to be a strong correlation between precision, the perception of stressed syllables and also vowel reduction. This area will require further research before anything more than tentative conclusions may be drawn.

3.2 Inversion of the articulatory-acoustic mapping

One area of interest for researchers in the field of articulatory models is the inversion of the articulatory-acoustic mapping. A successful inversion of this mapping would allow study of articulatory dynamics in natural speech simply by analysis of the speech signal. A number of problems restrict the reliability and utility of such an approach. Chief among these problems is the "ventriloquist effect": the articulatory-acoustic mapping is non unique - many articulatory configurations many potentially lead to the same acoustic signal. This presents problems when inverting the mapping: which articulatory configuration should be chosen?

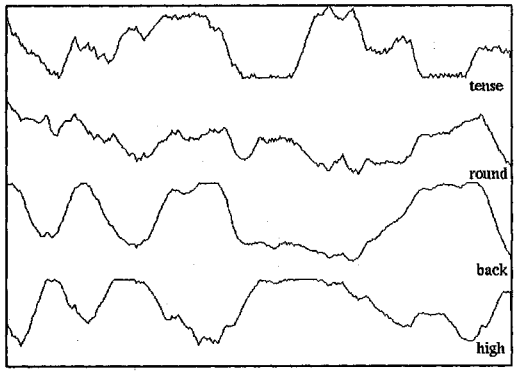

Figure 10: Articulatory features extracted from natural speech

Examination of the model on which our hybrid synthesis approach is based indicated that due to the simplified view of articulation that the model embodies, such many-to-one mappings seemed to be rare. A number of experiments were undertaken to discover whether the inverse mapping could be achieved successfully. This work began by using synthetic speech as the input to the inverse mapping process. We were successfully able to re-create the articulatory trajectories used to synthesize phrases of voiced speech from synthesizer parameters using an inverse-mapping procedure. This procedure involved a coarse and a fine grained search of the articulatory-acoustic solution space using a weighting to emphasize the lower, more perceptually important formants, and heuristic rules to limit the valid range of articulatory movement. With the success of this experiment we moved on to try extracting articulatory parameters from natural speech. The procedure followed was to extract the first three formant positions from a sample of natural speech, and use these as input to the inverse mapping process. Illustrated in Figure 10 are the articulatory feature trajectories for the phrase "Why were you away a year Roy?", as extracted from natural speech. In Figure 11 we illustrate the original formant positions, overlaid with the formant positions generated by copy synthesis from the extracted articulatory trajectories. This figure illustrates that there is only minor variation between the original natural speech and the synthetic copy. These initial results from limited experiments are promising.

4 CONCLUSIONS

We have presented a description of a quasi-articulatory synthesis strategy and have demonstrated that it is possible to produce intelligible speech using a simple approximation of articulatory control driving a formant synthesizer. We have also demonstrated two of the advantages that this articulatory based approach to synthesis offers, namely: the ability to dynamically modify the precision of articulation of an utterance to enhance its perceived naturalness, and the provision of a simple technique for inversion of the articulatory-acoustic mapping.

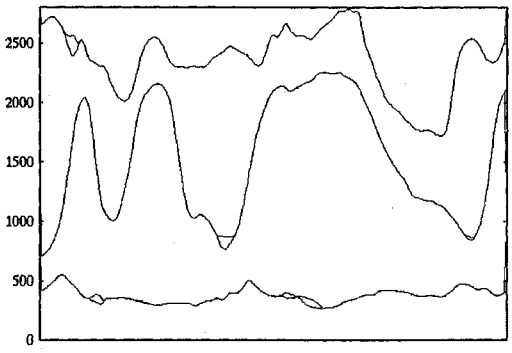

Figure 11: Original and copy-synthesized formants

5 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.

REFERENCES

[1] C. Browman and L. Goldstein. Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. Beckman, editors, *Papers in laboratory phonology I: Between the grammar and the physics of speech*, pages 341–376. Cambridge University Press, 1990.

[2] W.H. Edmondson and J.P. Iles. A non-linear architecture for speech and natural language processing. Appears in these proceedings.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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