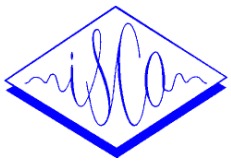

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

FEATURE DRIVEN FORMANT SYNTHESIS

Jon Iles and William Edmondson

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

ABSTRACT

This paper presents a brief summary of work carried out with a hybrid speech synthesis strategy. This synthesis strategy has been developed to allow control of the synthesis process using a representation based on distinctive feature notation. The hybrid approach has been found to offer advantages over other synthesis strategies, including the ability to mimic subtle features of natural speech, provision for variability in the precision of articulation, and the possibility of inverting the articulatory-acoustic mapping.

1 INTRODUCTION

Our investigation began when we considered the poor quality of rule-based synthetic speech. This poor quality is highlighted when rule-generated synthetic speech is compared to speech generated using copy-synthesis techniques. Copy-synthesized speech is often indistinguishable from the natural speech on which it was based. Why has this level of quality not been achieved by rule-based synthetic speech? In an attempt to answer this question we have identified two problems in the field of text-to-speech conversion: the control strategy used to drive the low-level synthesis process, and the architecture within which the whole process of text-to-speech conversion is undertaken. The architectural issue is discussed elsewhere [3, 4, 6], the background and development of the hybrid approach is discussed below.

2 BACKGROUND

Generally speaking formant-based synthesis techniques can be demonstrated producing high quality copy-synthesized speech. This indicates that there is no fundamental flaw in the implementation of a typical formant synthesizer that prevents high quality speech production. We believe that the problem lies with the control parameters used to drive the synthesizer itself. These are typically acoustic domain parameters such as formant frequencies, amplitudes and bandwidths. Although these parameters can be derived from natural speech and are the result of articulation, they are not the ideal method of describing speech. This is illustrated by an example provided by Fant ([5], p.84) in a diagram indicating the trajectories of f_1 through f_5 generated by a 3 parameter vocal tract model in which the tongue constriction centre is moved linearly away from the glottis. This simple action over a short period of time produces wide variation in formant position (and no doubt in formant amplitude and bandwidth also). Reproduction of these acoustic features would required a number of ad-hoc rules operating in the acoustic domain. High-quality synthesis would be easier if the articulation that produced the original acoustic features were to be modelled directly.

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. This is directly attributable to the advantages of using articulatory rather than acoustic domain model of speech: simplification of the rules required to control the synthesis process, and the ability to model the subtle effects of articulation that are difficult to mimic directly in the acoustic domain. Unfortunately, two problems are holding back the rapid development of articulatory synthesis strategies: difficulty in obtaining measurements of the physiological parameters of real vocal tracts, and generation of trajectories of the articulators in an articulatory model.

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. This gives us the advantages of articulatory-style control parameters while building on the potential quality of formant synthesis. Hybrid approaches to synthesis, and higher levels of control parameters superimposed over more complex low-level controls are not new (e.g. [1, 7, 10]). 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. Our work is linguistically motivated: the model is based on the notion of distinctive features that approximate articulation in a real vocal tract.

3 IMPLEMENTATION

The implementation of the hybrid model we have discussed above requires two sets of data - articulatory parameters and the equivalent acoustic parameters. This type of data was available to us at the phonemic level, so it was decided to concentrate on providing a segmental synthesis model to simplify implementation. The work discussed in the following pages of this paper is all based on the segmental model we have developed. This does not however preclude a non-segmental approach to synthesis, as pioneered by Local et al [2, 8] and discussed elsewhere by Iles and Edmondson [6]. An investigation in this area is planned as a continuation of the work described here.

To construct the model, the two sets of data mentioned above (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) were derived. The synthesizer parameters (for a Klatt cascade-parallel formant synthesizer) were measured by hand from recorded examples of natural speech from a target speaker. The articulatory features used were derived from distinctive feature notation as used by linguists, the main difference being the use of continuous coefficients for some of the features rather than the traditional binary coefficients. The features chosen were **HIGH** (tongue height - continuous), **BACK** (tongue back-front position, continuous), **ROUND** (lip rounding, continuous), **TENSE** (tongue tension, continuous), **LABIAL** (binary), **CORONAL** (binary), **STRIDENT** (binary), **ANTERIOR** (binary), **VOICED** (continuous), **FRICATED** (continuous), **ASPIRATED** (continuous), and **NASAL** (continuous).

The two sets of data were correlated using multiple regression analysis. The result of this was a model capable of approximating vowel articulation using the continuous features listed above. To model the articulation of other classes of sounds the acoustic properties of these sounds were superimposed over vowel articulation. That is to say, all speech sounds are treated as if they are vowels with additional acoustic information superimposed when required. A simple example of this is the articulation of fricatives. Suitable articulatory features are used as input to the model which generate a series of coarticulated vowel sounds. The binary features listed above are used to compute suitable amplitude and bandwidth values which replace those originally calculated by the model. When fricated excitation is supplied with these parameters, rather than voiced excitation, the correct fricative sound is produced.

The model generated by the steps described above was then set in a rule framework which converted from a phonetic specification of a piece of speech (phone name, duration and f_0), to a series of frames of articulatory data, and finally to synthesizer parameters. The first step in this process is to look up the "idealized" articulatory feature values for each phone in the input specification. These idealized values are then used to generate a set of identical frames of articulatory feature data. These frames each represent 5ms of speech, and the required number of frames are generated to match the specified duration of the input phone. These 5ms feature frames are combined into a list. This list can still be considered as a segmental representation of the required speech: there are distinct boundaries between each of the segments in this list and the values across the duration of each segment are constant. A set of rules is then applied to this list of frames to generate smooth transitions across phone boundaries in each of the continuous articulatory features. A rule exists to specify the transition between each "class" of phone, for example vowel to voiced fricative, vowel to vowel and so on. A cosine interpolation function is used to provide the basis for these smooth transitions. Finally the list of articulatory feature frames containing the smooth transitions are mapped to synthesizer parameters using the procedure described above.

4 ASSESSMENT OF THE MODEL

A rough assessment of mapping between articulatory features and synthesizer parameters is provided by the R^2 measure. This statistic gives an indication of the percentage of the observed data that can be explained by the model we have produced. The average R^2 value for formants f_1 through f_6 is 73.2%. R^2 values for the lower formants are in the high 90% region. The average value is pulled down by poor results for the higher, less perceptually relevant formants. This may reflect problems encountered in making accurate measurements for these values. This pattern of R^2 values is repeated for the other main groups of synthesizer parameters, (i.e. formant amplitudes and bandwidths).

Modified Rhyme Tests (MRT) have also been carried out to provide a rough guide to the intelligibility of the model. A similar methodology was used to that described by Logan et al [9], allowing direct comparison of the results obtained with other synthesis systems. Our hybrid system scored 18.0% error in a closed response format

test, and 53.8% error in an open response format test. We believe these are good results, indicating that our hybrid approach is comparable in intelligibility to many of the systems tested by Logan et al, especially when considering that this work has only been proceeding for a period of about a year.

5 ADVANTAGES OF A HYBRID APPROACH

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.

5.1 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 1 illustrates how this is achieved. Parts (a) and (b) of Figure 1 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 1, we see these targets illustrated as 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 1 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.

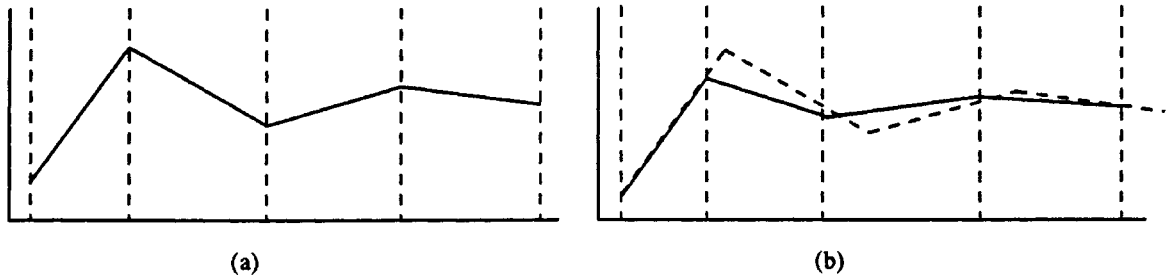

Figure 1: Precise and imprecise articulation

We have implemented precision of articulation control as a part of our hybrid model. This has allowed us to demonstrate that effects seen in rapid speech can be modelled by varying the precision of articulation of synthetic speech. We have also noted that there seems to be a strong correlation between precision, the perception of stressed syllables and vowel reduction. This area will require further research before anything more than tentative conclusions may be drawn.

5.2 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?

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.

This procedure proved to be successful for voiced samples of natural speech. Not only was good copy synthesis achieved, but the effect of poor formant tracking was considerably reduced by the use of heuristics governing permissible rates of articulator motion. The copy synthesized speech also contained 'synthetic' high frequency detail created from the articulatory specification. This adds to the naturalness of the copy-synthesized speech when compared to a straight-forward re-synthesis of the acoustic parameters originally measured from the natural speech. The "compressed" effect heard in speech sounds restricted to the first three formants is reduced by the additional high-frequency content.

6 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.

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.

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] 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.

[3] W.H. Edmondson and J.P. Iles. A non-linear architecture for speech and natural language processing. To appear in the Proceedings of the 1994 International Conference on Spoken Language Processing.

[4] W.H. Edmondson and J.P. Iles. Pantome: an architecture for speech and natural language processing. To appear in the Proceedings of the Institute of Acoustics 1994 Autumn Conference on Speech and Hearing.

[5] G. Fant. *Acoustic theory of speech production.* Mouton, The Hague, The Netherlands, second edition, 1970.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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