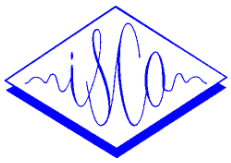


The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

AUTOMATIC DERIVATION OF SEGMENT MODELS FOR SYNTHESIS BY RULE

Wendy J. Holmes and David J. B. Pearce
GEC-Marconi Limited, Hirst Research Centre,
East Lane, Wembley, Middlesex, HA9 7PP, U.K.

ABSTRACT

This study aims to improve synthesis quality using the Holmes, Mattingly and Shearme (1964) phonetic-level synthesis-by-rule (SbR) method, both by increasing the inventory of allophone segments and by *automatically* optimising the values of the segment-table entries. Every occurrence of each phoneme is first optimised using a separate segment model. Initial estimates are iteratively refined using an analysis-by-synthesis procedure based on comparisons between the natural and rule-synthesised speech spectra, so imposing the inherent continuity constraints of the SbR model. The paper describes this automatic process, whose output is a set of individual segment tables for high quality segmental copy synthesis. These individual tables will later be combined to form allophone models for improved synthesis by rule.

1. INTRODUCTION

At the acoustic-phonetic level, speech synthesis by rule (SbR) usually involves applying rules to generate speech from a specification in phonemic units together with some prosodic information. The work described in this paper uses the system of Holmes, Mattingly and Shearme (1964), to generate frame-by-frame control data for the JSRU parallel-formant synthesiser (Holmes, 1985). So far, the quality of speech produced by the JSRU synthesiser when controlled from this level has not been very good. The synthesiser has, however, been used to obtain good quality copy synthesis, for both male and female speech (Holmes, 1973; W. J. Holmes, 1989). In copy synthesis, the frame-by-frame control signals are derived directly by acoustic measurement from natural speech, and it thus appears that the limitation of the SbR is in the modelling at the segmental level. However, using copy synthesis control signals obtained previously (W. J. Holmes, 1989), informal experiments have demonstrated good quality speech using the Holmes-Mattingly-Shearme (HMS) segment structure to model frame-by-frame values. Every occurrence of each phoneme in the utterance was modelled with a different segment specification, and the parameters for each table were carefully estimated by hand. The quality of the speech from this "segmental copy synthesis" sounded only slightly different from, and in some cases actually better than, the corresponding frame-by-frame copy. The problems in obtaining good quality synthetic speech by rule thus seem to be caused by a lack of appropriate values in the segment tables rather than by the nature of the parameter generation algorithm. The segment tables are deficient in two areas: Firstly, some of the allophonic variation cannot be accommodated by the co-articulation modelling ability of a single element table, so more allophone tables are needed. Secondly, appropriate values must be determined for the table entries of each allophone model, to provide acceptable speech quality over all environments in which that model is used.

Most existing sets of segment tables for the JSRU synthesiser have been derived by hand, using an iterative process of refining values and listening to the results. This is both difficult and time-consuming. An automatic process should enable better models to be obtained more easily, and makes it simple to add new voices to a system, as all that is required is sufficient transcribed speech data from a suitable speaker. A set of automatic procedures is being developed, which together are expected to be capable both of determining the inventory of allophone segments and of finding appropriate values for their table entries. The approach taken is to begin by treating every occurrence of each phoneme as a separate segment model when estimating the parameters, and then to apply a second stage of model clustering to combine similar models. This paper describes the first stage, of automatic segmental copy synthesis. The resulting individual segment models will then either be clustered to derive the set of allophone models, or they can be combined using a pre-determined allophone inventory.

One method for performing automatic segmental copy synthesis, adjusting SbR models for the JSRU synthesiser based on a single natural utterance, has been described by Bridle and Ralls (1985). Their procedure adjusted formant frequency and amplitude targets in the segment tables so that the rule-generated parameter tracks were as close a match as possible to parameter tracks obtained from frame-by-frame copy synthesis. However, the values of the original frame-by-frame parameters would not necessarily have been optimum for segmental modelling. To overcome these problems the present study uses a distance metric based on analysis-by-synthesis, whereby element tables are iteratively refined until the spectrum of the speech they produce is as

close as possible to the spectrum of the natural speech. Using analysis-by-synthesis directly at the segmental level has the advantage that the inherent continuity constraints of the SbR model are imposed.

2. THE JSRU SYNTHESISER AND SYNTHESIS BY RULE METHOD

The JSRU synthesiser (Holmes, 1985) uses a parallel-formant network with five branches containing resonators to model the frequency region up to that of the fourth formant. Besides the excitation specification, values for the following parameters need to be decided for every 10 ms frame:

- frequencies of the first three formants (F1, F2, F3)
- amplitudes of these formants (A1, A2, A3)
- intensity of the frequency region below F1 (ALF)
- intensity of the high-frequency region (AHF).

The fixed frequency band controlled by AHF is intended to cover the F4 range. In the original form of the synthesiser F4 is at 3.5 kHz, which is a suitable value for male speech. When synthesising female speech, this frequency has been raised to 4 kHz (W. J. Holmes, 1989).

The SbR algorithm used in the current work is basically the HMS system, adapted for the latest JSRU synthesiser and with some minor improvements to the algorithm (Holmes, 1988). This system represents each speech sound by a table, and is based on the idea that most speech sounds can be modelled by a target acoustic specification and transition rules for moving between targets. Sounds such as diphthongs and stop consonants, which have a sequence of acoustic properties, are represented by two or more component parts. The term "phonetic element" is used to describe the section of sound specified by one table, which may thus either be a complete phone or part of a phone.

In normal operation, the HMS system uses phonetic context to vary element boundary values, but this facility is not required for the current stage of this study, as each element table is only used in one context. For the frequency and amplitude parameters of the formants of each element, it is therefore necessary to determine :

- target
- parameter value at nominal boundary with preceding element ("initial boundary value")
- initial transition duration
- parameter value at nominal boundary with following element ("final boundary value")
- final transition duration.

The frame-by-frame parameter tracks for one segment are obtained for each parameter by interpolation from the boundary values towards the targets over the specified transition regions. The target value is maintained in any steady-state region between the transitions.

3. THE AUTOMATIC SEGMENT MODEL DERIVATION TECHNIQUE

Segment models are derived using an iterative re-estimation procedure applied to labelled and segmented speech data. Currently the segmentation is performed by hand, although an automatic method will be implemented before processing much larger quantities of speech data. To derive segment tables, it is necessary to obtain values for the relevant table entries for each parameter (see Section 2). In the current implementation the transition durations are set using phonetic knowledge, but in the next version they too will be re-estimated using the iterative optimisation procedure.

3.1. The segment model re-estimation algorithm

Fairly accurate initial estimates for the segment tables were obtained by simply measuring target and boundary values from appropriate regions of copy synthesis parameter tracks (W. J. Holmes, 1989). Phonetic knowledge was used to guide the re-estimation, such that values for each formant were always limited to be within a frequency range considered to be reasonable for that speech segment and type of speaker, and were constrained to be at least 150 Hz apart.

Within any restrictions imposed on allowed values, the target and boundary table entries were re-estimated in an ordered procedure using a grid-based search. The most important aspects of the tables are the parameter targets, so all targets were re-estimated before boundaries were optimised. Further iterations of target and boundary optimisations were then performed. In any one iteration the parameters for the first three formants were re-estimated one at a time, provided they were not close enough to be classed as significantly interacting. To avoid possible problems if the initial values for one formant frequency were a long way from the optimum, the re-estimation was performed in an order determined by the measured sensitivity of the distance score to

changes in frequency. When the starting frequencies for adjacent formants were close together (less than 500 Hz), they were optimised jointly. For each of the three variable-frequency formants, the frequency table entries were re-estimated with the formant amplitude optimimum for each frame, and then the amplitude entries were optimised for the chosen frequency track. At each iteration, the low-frequency amplitude control (ALF) was re-estimated after optimising the first formant, as its optimum value depends on F1.

3.2. The distance measure used for re-estimation

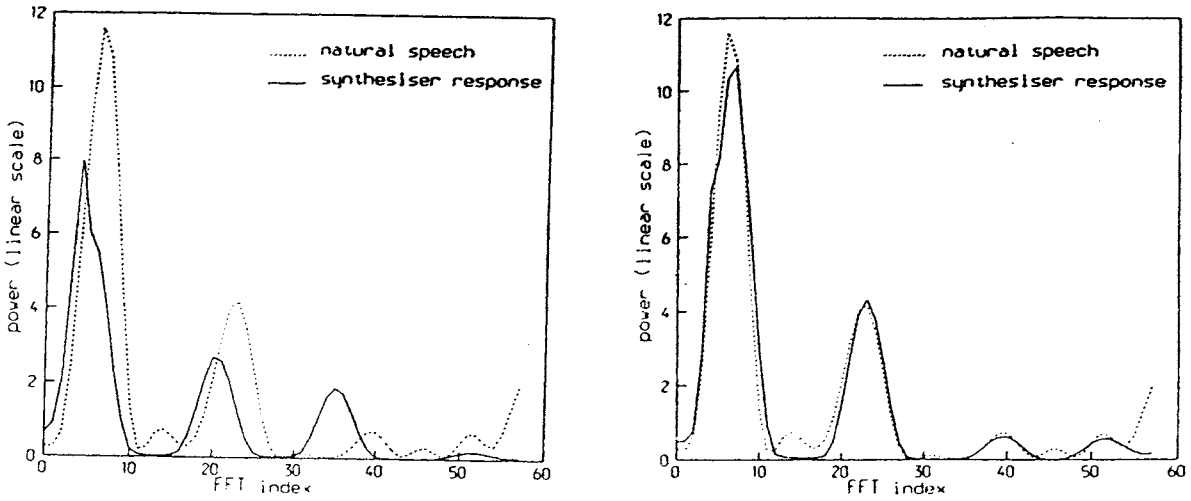
Calculation of natural speech spectrum: The natural speech spectrum was derived in different ways for voiced and unvoiced speech, and a general measure of low-frequency amplitude was used for optimising ALF. For voiced speech, a closed-phase excitation-synchronous FFT analysis was performed, using speech samples from the first 2.5 ms after glottis closure markers. The speech samples were padded with zeros and a 12.8 ms FFT performed in order to obtain reasonably closely spaced frequency samples. The measured powers in appropriate analysis frames were averaged to provide a result for each 10 ms frame. For unvoiced speech, a 12.8 ms Hamming-windowed FFT was taken every 5 ms, and measured powers for pairs of windows were averaged together to derive a result for every 10 ms frame. For estimating ALF, the sum of the powers of the low-frequency components was taken from 25.6 ms Hamming-windowed FFTs at 10 ms intervals.

Calculation of synthesiser response spectrum: As the synthetic speech spectrum for each frame has to be calculated many times during the re-estimation, the spectral response was not calculated from the waveform. Instead a spectral representation was computed directly from the synthesiser control parameters, by taking a sum of the transfer functions of the individual formant resonators. The response of each resonator was pre-computed for a suitable range of formant frequencies and stored at the same 78.125 Hz spacing as was used for the natural speech analysis, up to the 4.5 kHz limit of the synthesiser. Then for any set of control parameters it was simple to calculate the combined response as required. During voiced speech the formant responses were convolved with an approximation to the spectrum of a 2.5 ms rectangular window, so that the peaks in the calculated spectrum were a similar shape to those obtained from the windowed natural speech. A spectral measure was obtained for ALF by summing the responses in the same low-frequency region as was used for the natural speech analysis.

Distance calculation: The spectral distance measure for re-estimating targets was evaluated over the duration of the associated phonetic element, but for boundaries the distance was evaluated over two consecutive elements. It is important that the distance measure gives most weight to the degree of match round spectral peaks, and is not greatly affected by differences in the trough regions. Therefore the distance calculation was applied to the natural speech and synthesiser response on a linear power scale. The distance measure was calculated separately for each formant, but only over the appropriate frequency range for that formant to reduce the danger of weak formants being disturbed by slight mis-match of adjacent strong formants. Before calculating a distance score, a global scale factor was applied to the measured speech spectra (with a different scale factor for voiced speech, unvoiced speech and the ALF measure) so that the speech spectrum and calculated response covered a similar intensity range. The distance for one frame was the sum of the squared differences between the two spectra measured over the appropriate frequency range. The fourth root of the distances for individual frames were summed, to give an appropriate degree of weight to differences in weak frames relative to strong frames. This corresponds quite well to psychological measures of loudness.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

Various tests have been carried out, comparing calculated synthesiser response spectra with measured spectra of both rule-generated synthetic speech and natural speech from one female speaker. Firstly, it was verified that the synthesiser response calculated did actually provide a close approximation to spectra measured from synthetic speech for the same control signals. Spectral cross-sections for single frames were compared, and distance scores obtained for various matches. For voiced speech, there was no significant error on any formant for the range of values tested. For unvoiced speech the formant frequencies were generally within 100 Hz, but some variation was to be expected in unvoiced speech due to the random nature of the excitation. Some comparisons were then made between single frames of natural speech and calculated synthesiser response spectra for various values of control parameters, to check that it was possible to obtain a good match to natural speech spectra using these calculated synthetic spectra. It was found that, particularly during sonorant sounds, a very close match was possible and that, as synthesiser response frequencies and amplitudes were changed, the calculated distance scores also changed in a reasonable way.



(a) Before re-estimation (b) After re-estimation

Fig. 1 - Spectral cross-sections for natural speech and synthesiser response in one frame of an /ə/ vowel

Having evaluated the distance metric, the re-estimation algorithm could be tested. The first experiments were carried out with speech which had been generated by rule, so that the correct table values were known. It was found that the initial estimates for the formant parameters could be deviated by amounts of up to around 400 Hz and the re-estimation program could return values near to those originally used to generate the speech. The re-estimation program has so far been tested on a small number of natural speech utterances and preliminary results seem promising. Comparisons based on informal listening tests and study of spectrograms have shown that good quality segmental copies of natural speech can be obtained by the automatic method described, with the voice quality of the original speaker preserved. Currently the performance is best for sonorant sounds. Fig. 1 shows spectral cross-sections in the target region of a /ə/ vowel. A comparison is shown between natural speech and the calculated synthesiser response, both before and after re-estimation, and it can be seen that the formants have been moved from being quite a bad match for the natural speech to being a very close fit. Results for fricatives have also been quite successful, although for stop consonants it appears that the results could be improved by applying more phonetic knowledge to further limit the range of allowed values.

The automatic procedure for deriving individual segment tables for high quality segmental copy synthesis has been shown to be successful on a limited number of utterances. Future work will involve testing on a larger database to obtain many examples of context-dependent phonetic element tables for each allophone. These tables will then be combined to form allophone models for improved synthesis by rule. Next, formal listening tests will be conducted to assess the quality of both the segmental copy synthesis and the synthesis by rule.

ACNOWLEDGEMENTS

This work was performed for Marconi Speech and Information Systems as part of the Alvey Integrated Speech Technology Demonstrator programme with CSTR and HUSAT.

REFERENCES

Bridle J S and Ralls M P (1985) "An approach to speech recognition using synthesis-by-rule", in "Computer Speech Processing", F. Fallside and W. A. Woods (Eds.), Prentice-Hall International.

Holmes J N (1973) "The influence of glottal waveform on the naturalness of speech from a parallel-formant synthesizer", IEEE Trans. Audio, Electroacoust., Vol. AU-21, pp. 298-305.

Holmes J N (1985) "A parallel-formant synthesizer for machine voice output", in "Computer Speech Processing", F. Fallside and W. A. Woods (Eds.), Prentice-Hall International.

Holmes J N (1988) "Speech Synthesis and Recognition", Van Nostrand Reinhold (UK).

Holmes J N, Mattingly I G and Shearme J N (1964) "Speech Synthesis by Rule", Language and Speech, Vol. 7, pp. 127-143.

Holmes W J (1989) "Copy synthesis of female speech using the JSRU parallel-formant synthesiser", Proc. European Conf. on Speech Communication and Technology, pp. 513-516.