Proceedings of the Institute of Acoustics

AUTOMATIC PARAMETER DERIVATION FOR SYNTHESIS-BY-RULE ALLOPHONE MODELS

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 for deriving individual segment tables, and also describes the method for combining specified sets of individual models to form allophone models. In the next stage, clustering procedures will be applied to determine the required extended allophone inventory to improve the quality of the 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 approach of Holmes, Mattingly and Shearme [1] to generate frame-by-frame control data for the JSRU parallel-formant synthesiser [2]. 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 [3,4]. 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 rather than being a problem with the synthesiser itself. However, as a preliminary to the present study, it has been shown possible to demonstrate good quality synthesis using the Holmes-Mattingly-Shearme (HMS) segment structure, when the segment parameters were optimised by hand for a particular utterance. Every occurrence of each phoneme in the utterance was modelled with a different segment specification, and the parameters for each table were estimated to model frame-by-frame copy synthesis control signals obtained previously [4]. 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.

AUTOMATIC PARAMETER DERIVATION FOR SYNTHESIS-BY-RULE

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 would make it simple to add new voices to a system, as all that would be 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 modelling every occurrence of each phoneme with a separate segment table when estimating the parameters, and then to apply a second stage of combining similar models to derive a set of allophone models. The first stage, which will be referred to as "automatic segmental copy synthesis", has been described in [5]. This paper describes the segmental copy synthesis in greater detail, and includes more recent developments. The paper then discusses the method used to combine the resulting individual segment models according to a pre-determined allophone inventory. In future work, clustering techniques will be applied to determine the set of allophone models required to adequately model the required range of utterances.

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 [6]. 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 avoids the frame-to-frame continuity problems which have occurred when it is applied on a frame-by-frame basis [7], as the inherent continuity constraints of the SbR model are imposed.

The following sections of this paper give a brief description of the JSRU synthesiser and SbR method used. The model optimisation scheme is then described in more detail and the automatic segmental copy synthesis results discussed, before describing the method of combining the individual models to obtain allophone models and initial synthesis-by-rule results.

2. THE JSRU SYNTHESISER AND SYNTHESIS-BY-RULE METHOD

The JSRU synthesiser [2] 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 [4].

AUTOMATIC PARAMETER DERIVATION FOR SYNTHESIS-BY-RULE

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 [8]. 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.

The HMS system uses phonetic context to vary the parameter values at the nominal boundary where two elements join. For example, at a boundary between a consonant and a vowel, the behaviour of any one parameter will be determined by the target value of the vowel and the identity of the consonant. The consonant table will contain the following entries:

- its own target value
- the proportion of the vowel target to be used in deriving the boundary value
- a "fixed contribution " (FC) to the boundary value
- the transition duration within the consonant (in frames)
- the transition duration within the vowel (in frames).

The boundary value is calculated as: boundary = FC + proportion × vowel target value. 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. The system of boundary value calculation means that, for one vowel, the boundary value will be different for different consonants. The relative influence of any one phonetic element on another is represented in the HMS system by associating each element with a 'rank'. Stop consonants generally cause transitions in all other types of sound and therefore have the highest rank, whereas vowels have very low rank and semivowels an intermediate rank. For any sequence of two elements, the table with the higher rank is used to determine the nature of the transition. As published, the HMS system is symmetrical, treating vowel-consonant and consonant-vowel transitions in exactly the same way. However, in the work described in this paper, the tables have been extended to include separate transition specifications for initial and final transitions.

The allophone models output from the combination process need to contain all the above information. However, at the individual model optimisation stage, the tables can be simplified as each table is used in only one context. At this stage it is therefore only necessary to derive for the formant frequency and amplitude parameters in each element:

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

3. AUTOMATIC SEGMENT MODEL DERIVATION

Individually optimised 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). The table entries for the degree of voicing parameter are set using phonetic knowledge, with suitable values assigned to each allophone for different types of phonetic context if required. Formant frequency and amplitude parameters are derived using the iterative optimisation procedure, as described below.

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 [4]. Initial values for the transition durations were set using phonetic knowledge. The re-estimation was also guided by the use of phonetic knowledge, 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. Transition durations were also constrained to be in a suitable range according to the phonetic identity of the speech segment.

Within any restrictions imposed on allowed values, the target, boundary and transition duration 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 the other table entries were optimised. The boundary values were optimised next, with the transition durations optimised last as these were considered to have the least influence on the spectrum of the resulting synthetic speech. Further iterations of target, boundary and transition duration 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 optimum for each frame. For the chosen frequency table entry, a record was kept of the best formant amplitude for each frame, and then in a second stage the amplitude entries were optimised to match the frame-by-frame amplitude tracks. 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 intervals 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. In sounds involving mixed voicing, the voiced analysis was used for the lower

formants and the unvoiced analysis for the higher formants. For each formant, the point of change from using voiced to using unvoiced analysis was defined as the degree of voicing at which that formant would receive an equal mix of voiced/unvoiced excitation in the synthesiser. 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.

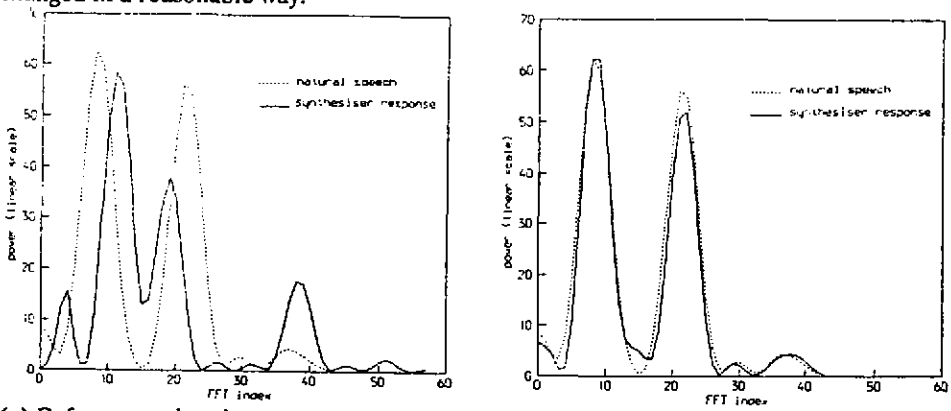
Calculation of synthesiser response spectrum: As the synthetic speech spectrum for each frame has to be calculated many times during the re-estimation, the spectral response was not calculated from the waveform. Instead a spectral representation was computed directly from the synthesiser control parameters, 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 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 and transition durations 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.

3.3. Experimental results

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 /3/ 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, when the initial estimates for the formant parameters were deviated by amounts of up to around 400 Hz, the re-estimation program would return values near to those originally used to generate the speech.

The re-estimation program has been tested on a natural speech database of 75 isolated words and five sentences for one female speaker. The words in this database were chosen to provide a few examples of every phoneme, covering a number of acoustically different phonetic contexts. The utterances were synthesised with the prosody copied from the natural versions, and comparisons made with the natural speech low-pass filtered to 4.5 kHz. 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. For example, Fig. 1 shows spectral cross-sections in the target region of an /3/ 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. The algorithm is able to perform well on sonorant sounds without many restrictions on the range of allowed values. For consonant sounds, the results are best when phonetic knowledge is applied to limit the range of allowed values to ensure that the synthesiser parameters are used in an appropriate way to model the spectral characteristics of these sounds. This is an area where investigations are continuing, to further improve synthesis quality and robustness of the algorithm to perform well on all speech sounds in any context.

4. COMBINING INDIVIDUAL SEGMENT MODELS INTO ALLOPHONE MODELS

4.1. Method

The automatic segment model derivation stage described above produces as output a set of individually optimised models where every occurrence of every allophone has its own model. For each allophone, the model combination stage averages all examples to form a single model which can be used in place of any of the individual models. The target values in the combined model were derived by simply calculating the average of the individual target values. When deriving the transition information in the combined model, it is necessary for each individual example to take into account the rank of the allophone in relation to the ranks of the two adjacent elements. For both the initial and final transitions, the examples which will contribute to the average will be those which are dominant in the context in which they were derived, and will therefore control the transitions when SbR is carried out using the rank system. Within this constraint, the transition durations were obtained by taking a weighted average, where each contribution to the average was weighted by the magnitude of the difference between the target and boundary values. Thus transitions of large magnitude, where the duration is most important, have more influence than transitions of small magnitude. As part of the combination process, the individually optimised boundary values need to be converted to the context-sensitive form of fixed contributions and proportions. This was accomplished by using linear regression to find the one value of fixed contribution and the one value of proportion which provide the best fit to all the individual boundary values. The target values used in solving the linear regression equation were the averaged values to be used in the combined models, as these are the ones which are actually used for SbR in the final system.

4.2. Experimental results

The model combination algorithm has been applied to the individual models obtained for the 75 isolated words (see Section 3.3), using an allophone inventory of 111 phonetic elements. This inventory represents the minimal set of elements that were considered to be definitely needed, and is intended to be extended later as required. It is based on the original inventory of 66 phonetic elements [9] used in the JSRU experimental text-to-speech system. Additions have been made to model the different degrees of voicing required for voiced fricatives, affricates and stops in different contexts, and also for phonemes in contexts which are known to be acoustically very different from the standard form of that phoneme (for example, /t/ in an /str/ cluster). The 75 words were synthesised using the allophone models and results compared with the natural speech and with the synthetic speech obtained from the individually optimised models, again with the prosody copied from the natural utterances. The combined models represent an average intensity and so, to reproduce an individual utterance, it was necessary to use "loudness modifiers" to adjust the overall intensity of each element to be appropriate for its occurrence in that utterance.

As long as the individual models were reasonably good, the synthetic speech obtained from the combined models was in most cases clearly intelligible and still retained many of the voice qualities of the original speaker. However, not surprisingly, for many examples it did sound considerably worse than the segmental copy synthesis using the individual models. The cases where the quality was worst were instances of phonemes for which it would seem reasonable to add an extra allophone (for example, some of the consonant clusters had not

at this stage been separated out as definitely requiring separate models). One way of improving the quality would be to add new allophones for those environments where the combined models are obviously inadequate. A better approach would be to automatically determine the required inventory from the data by using clustering procedures, and this is the approach which is being adopted. It will also be important to develop methods for coping with any occurrences of an allophone which have been pronounced differently from the majority of examples of the same allophone in similar phonetic contexts, and which should therefore be omitted from the combination process.

5. CONCLUSIONS

The automatic procedure for deriving individual segment tables for high quality copy synthesis has been shown to be successful on a range of utterances from one female speaker. The model combination algorithm has been applied to the individual models to obtain SbR models with an inventory of 111 phonetic element allophones. The synthetic speech produced by the combined models is intelligible and of reasonable quality, although in many instances it is significantly worse than that obtained with the individually optimised models. To obtain really good quality, more allophones are still required, and so clustering procedures will now be developed to derive the additions to the inventory from the requirements of the data.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] Holmes W J and Pearce D J B (1990) "Automatic derivation of segment models for synthesis by rule", Proc. ESCA Workshop on Speech Synthesis, Autrans, France, September 1990, pp. 5-8.

[6] 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.

[7] Seeviour P M, Holmes J N and Judd M W (1976) "Automatic generation of control signals for a parallel formant speech synthesizer", Proc. IEEE ICASSP, pp. 690-693.

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

[9] Edward J A (1982) "Rules for Synthesising the Prosodic Features of Speech", JSRU Research Report No. 1015.