COPY SYNTHESIS OF FEMALE SPEECH USING THE JSRU PARALLEL FORMANT SYNTHESISER

Wendy J. Holmes

GEC Hirst Research Centre,
East Lane, Wembley, Middlesex, U.K. HA9 7PP

ABSTRACT

This paper presents work using a new method of copy synthesis for the JSRU parallel formant synthesizer, to obtain good quality synthesis of both female and male speech. Natural speech was analysed to derive values for each of the synthesizer control parameters at regular (10 ms) intervals. The formant frequencies were obtained using excitation synchronous formant analysis and labelling techniques previously reported [10,11]. The amplitude control parameters were derived by a two-stage process: first the spectral amplitudes at the formant frequencies were measured from FFT analyses, and then the measured amplitudes were transformed using a table of "amplitude correction values" (ACVs) to obtain amplitudes which are appropriate as control parameters. Some modifications made to the synthesizer to improve the production of female speech are also described. The results have demonstrated that it is possible to obtain good quality female speech from the JSRU synthesizer.

1. INTRODUCTION

In copy synthesis, synthesizer control parameters are specified frame by frame with the aim of obtaining a synthetic utterance which sounds as much like a particular natural utterance as possible. The parallel-formant synthesizer developed at the Joint Speech Research Unit [1,2,3] has been successfully used for copy synthesis of male speech (for example, [4]). However, very little copy synthesis of female speech has been attempted and no successful synthesis of female speech has previously been reported. The aim of the work described in this paper was to demonstrate that good quality female speech can be obtained from the JSRU synthesizer, using a largely automatic procedure of analysis for synthesizer parameters. The present work on copy synthesis is intended to provide the first stage towards automatic extraction of segment models for synthesis by rule.

1.1. The JSRU synthesizer

The details of the JSRU synthesizer have been described elsewhere (e.g. [1,2,3]). A brief summary will be given below, to provide a background for explaining the parameters which need to be derived for copy synthesis and a method to extract appropriate values from natural speech.

The principles of operation of the synthesizer are based on a conventional model of speech production, which envisages speech as the output of a time-varying filter driven by a substantially separable excitation function. There is a voiced and an unvoiced excitation generator, both of which are arranged to produce a spectral envelope which is substantially flat over the frequency range of the formants. The time-varying filter consists of a parallel network with six branches containing resonators (as shown in Figure 1). This filter shapes the excitation spectrum to model both the response of the vocal tract and the natural variation of the excitation spectral envelope. The synthesizer is controlled by 12 signals, of which nine control the formant frequencies and amplitudes, and the other three control voiced excitation fundamental frequency, glottal pulse open/closed ratio and the degree of voicing. These control signals are typically specified every 10 ms. The work described in this paper used a software implementation of the synthesizer, which had been recorded in "C" from the Fortran listing given in [2]. The quantitative interpretation of the control signals is determined by pre-set parameters in a separate text file.

Each of the six parallel branches in the synthesizer has as input its own mixture of voiced and unvoiced excitation, depending on the degree of voicing control. Three of the branches represent formants F1, F2 and F3 and have variable frequency and amplitude controls. The resonator labelled FN contributes to the frequency region below F1 and is particularly important for nasal sounds. It is fairly heavily damped and behaves like a low-pass filter. Although its frequency is controllable, this study, like most previous work with the synthesizer, has kept the FN resonator frequency fixed at 250 Hz. There is a special connection between the amplitude control for F1 and the amplitude control for FN (called ALF), which is designed so that ALF effectively completely controls the low frequency region up to about 300 Hz. The remaining two branches in the synthesizer represent voiced and unvoiced alternatives (in the form of a single resonator and an 800 Hz wide band-pass filter) for the fourth formant region, which has fixed centre frequency (3500 Hz in the standard implementation) and a variable amplitude control. Each formant resonator is followed by its own fixed spectral weighting filter, to allow for more complex modelling and ensure proper mixing of the outputs from the branches. In voiced speech, all the resonators except FN have excitation synchronous bandwidth modification, to model the increase in formant bandwidths which occurs during the glottal open period. The sum of the outputs from the six spectral weighting filters is low-pass filtered at about 4.5 kHz and de-emphasized by the final output filter before producing the speech waveform.

1.2. Control parameters for copy synthesis

To use the synthesizer for copy synthesis, a method is required for analysis of natural speech to derive values for the following parameters every frame (usually 10 ms):

- degree of voicing (V)
- fundamental frequency (FO)
- frequencies of the first three formants (F1,F2,F3)
- amplitudes of these formants (A1,A2,A3)
- amplitude of the fixed high frequency formant (AHF)
- amplitude in the low frequency region (ALF)

The glottal pulse open/closed ratio is difficult to derive from natural speech so was left at its default value.

Although voicing and the fundamental and formant frequencies are often difficult to estimate accurately, there are various techniques available for directly obtaining values from natural speech. However, in the case of the formant amplitudes the requirement is to determine values from the
As in Breen's method, the present study uses estimated formant speech for copy synthesis, before explaining the dimensions required in sections of this paper will describe the process of analysis of natural frequencies to derive spectral amplitudes from an FFf and then the ACV look-up table and then the method for deriving the table. The which will be called synthesiser control amplitudes have already been obtained. The training data was obtained from Seeviour et al's [5] analysis by synthesis FFf components defined by the formant frequencies (which had already been estimated). The MLPs then transformed the measured amplitudes into synthesiser control amplitudes obtained from an initial analysis. The training technique which was used does not require training material. The availability of training material for the MLPs, where the desired amplitudes into speech with a particular spectral content. Thus the amplitude parameters are quite difficult to determine, but they must be fairly accurate as they represent both the response of the vocal tract and variation in the excitation spectral envelope.

1.3. Deriving amplitude control parameters

Much previous work on deriving amplitude parameters for the synthesiser has used analysis by synthesis [4,5,6,7]. Holmes [4] used a human in the adjustment loop, comparing spectrograms and waveforms in filtered bands. Seeviour et al. [5] and Dupree [6] used an automatic method for optimising all parameters by analysis by synthesis, based on comparisons of the log power spectrum of the natural speech with log power spectra of several versions of the synthetic speech. During voiced speech, these spectra were derived from excitation synchronous analysis with 3.2 ms windows. However, analysis by synthesis is a computationally expensive technique. To reduce the amount of computation, Lowry et al. [7] have used a method in which formant frequencies were first estimated and then analysis by synthesis was used to derive the amplitude parameters only, based on spectra from 30 ms analysis windows.

Breen [8] has used multilayer perceptrons (MLPs) to optimise synthesiser control amplitudes obtained from an initial analysis. The initial analysis derived the spectral amplitude of each formant by taking FFT components defined by the formant frequencies (which had already been estimated). The MLPs then transformed the measured amplitudes to be suitable as synthesiser control amplitudes. This method requires the availability of training material for the MLPs, where the desired synthesiser control amplitudes have already been obtained. The training data was obtained from Seeviour et al's [5] analysis by synthesis technique.

As in Breen's method, the present study uses estimated formant frequencies to derive spectral amplitudes from an FFT and then transforms these to synthesiser control amplitudes. However, the technique which was used does not require training material. The transformation is accomplished by means of a look-up table of quantities which will be called "amplitude correction values" (ACVs). The following sections of this paper will describe the process of analysis of natural speech for copy synthesis, before explaining the dimensions required in the ACV look-up table and then the method for deriving the table. The results are discussed.

2. METHOD OF ANALYSING NATURAL SPEECH

The analyses used in these experiments require the times of glottis closure to have been detected. This information was obtained from the waveform output from a laryngograph [9]. However, it would also have been possible to estimate times of glottis closure directly from the speech waveform. After estimating times of glottis closure, the parameters were obtained every 10 ms as follows:

2.1. Degree of voicing

In the automatic analysis procedure a binary classification for voicing was used, classifying each natural speech frame as either completely voiced or completely unvoiced. In cases where there was clearly mixed excitation in the natural speech, the degree of voicing parameter could be hand-edited appropriately. The automatic voicing classification was based on the spacing between the closure markers. A frame was classified as unvoiced if the closure markers were 20 ms apart or more.

2.2. Fundamental frequency

A fundamental frequency contour was obtained by taking the reciprocal of the spacing between the glottis closure markers when they were less than 20 ms apart, with interpolation between values at the end of one voiced region and those at the start of the next.

2.3. Formant frequencies

For voiced speech, it was found that in general good formant tracks were obtained by applying a 10 pole larynx synchronous covariance LPC analysis at 10 kHz sampling rate over the first 2 ms of the closed phase [10]. For unvoiced speech, it is difficult to define where formants should be, as energy is more spread out over the frequency spectrum. However, for providing synthesiser control parameters, the frequencies were found to be less critical for unvoiced than for voiced speech, as long as the amplitudes were correct for the frequencies that were used. A 10 pole covariance LPC analysis taken at 5 ms intervals was found to be quite successful. A formant-labeling algorithm [11] was applied to the output of the LPC analyses to determine values for F1, F2 and F3 every 10 ms. Any gross errors of formant frequency measurement were subsequently corrected by hand editing.

2.4. Formant amplitudes

Formant amplitudes were obtained in two stages. First spectral amplitudes were derived from FFT analyses, by measuring amplitudes...
from the components defined by the formant frequencies. Then amplitude correction was applied based on the ACV table. For the formants F1-F4, there was one type of analysis for voiced and one for unvoiced speech, with the amplitude measured at the formant frequency. For the ALF parameter there was a more general measure of amplitude in the low-frequency region, which was the same for both voiced and unvoiced speech. For ALF, one two-dimensional table (dependent on F1 and FO) designed specifically was used to investigate effects of synthesizer control amplitude, fundamental frequency, and formant frequency, and of adjacent formants. Thus it was possible to verify the theoretical justification of ACV requirements given in Section 4.1 and also to develop the special rules for accommodating interactions when formants move close together.

4. DERIVING AMPLITUDE CORRECTION VALUES

To derive amplitude correction values, it was necessary to study the relationship between the input synthesizer control amplitude for each formant and the measured amplitude in the synthetic speech. This was done by performing the reverse of the copy synthesis process: test signals were constructed by specifying synthesizer control parameters, then 'speech' was synthesized from these parameters and the amplitude measured at the formant frequencies using the methods described in Section 2 above. Then ACVs were simply calculated:

\[ ACV = \text{measured amplitude} - \text{input amplitude (dB)} \]

This technique was used to investigate the effects of synthesizer control amplitude, fundamental frequency, and formant frequency, and of adjacent formants. Thus it was possible to verify the theoretical justification of ACV requirements given in Section 3 and also to develop the special rules for accommodating interactions when formants move close together.

This method has the advantage that the ACV table is set up without requiring any previously analyzed natural speech. New ACVs may be required if changes are made to the synthesizer's preset parameter file, but are not needed for different speakers.

5. SYNTHESIZER MODIFICATIONS FOR FEMALE SPEECH

5.1. Formant damping

With the default bandwidth modification in the synthesizer, it was found that at typical female fundamental frequencies (200 Hz and above) there was much more interaction between successive excitation pulses than appeared to typically occur during natural female speech. This also meant that ACVs were affected by fundamental frequency for high fundamental frequencies, both while deriving ACVs and during copy synthesis of female speech. Therefore, the bandwidth modification of F1,
F2 and F3 was increased during the open phase to obtain more formant damping.

5.2. Frequency of F4

The default value of the fixed frequency fourth formant is 3500 Hz, which is quite suitable for male speech. However, for the female speech used in these experiments, F3 was often above 3000 Hz so it was not very realistic to have F4 set to 3500 Hz. Therefore the centre frequencies of both the voiced and unvoiced F4 filters were increased to 4000 Hz for female speech. In the case of voiced speech, it was also appropriate to increase the frequency of the pole for simulating the skirt of PS (see Figure 1) to 4700 Hz.

6. COPY SYNTHESIS RESULTS

The copy synthesis process has been tested on five sentences containing a variety of speech sounds (including vowels, semi-vowels, plosives, affricates and fricatives) spoken by two male and two female speakers. The quality of the copy synthesis was assessed by making comparisons with the natural speech low-pass filtered at 4.5 kHz, as this is the approximate frequency of the overall low-pass filter in the synthesizer. Comparisons were made by listening to the speech and by visual examination of waveforms and spectrograms.

It has been found that it is possible to obtain high quality synthesis for female as well as for male speech. Of the sentences tested, the highest quality was obtained for the utterance "Why are you early you owl?", which contains only sonorant sounds. However good results were also obtained for sentences containing non-sonorants. For example, Figure 2 shows spectrograms of natural and synthetic speech for a female speaker saying "six plus three" and it is evident that the main features of the natural speech have been reproduced in the synthetic version. In the /s/ sound, although the four formants in the synthesizer have not produced the continuous flat spectral trend which is seen in the natural speech from around 1 kHz to 4.5 kHz, this did not appear to be a problem perceptually. Listening to the speech, the synthetic version was perceptibly different from the original when the two were played in quick succession. However, the voice quality of the synthetic speech was very similar to that of the original speaker and, when played in isolation sounded acceptable as a recording of natural speech.

Figure 2 - Spectrograms showing copy synthesis results for a female speaker saying "Six plus three"

(a) Natural speech low-pass filtered at 4.5 kHz
(b) Synthetic speech

CONCLUSIONS

The results of these experiments have successfully demonstrated the technique of copy synthesis for the JSRU synthesizer using amplitude correction tables to transform initial amplitudes measured from FFT analyses. Providing that the formant frequencies for the voiced regions have been estimated accurately, all the amplitudes can be measured completely automatically by this method. In unvoiced regions, values for "formant" frequencies are difficult to estimate. However, the exact values do not seem to be so critical and the amplitude analysis process will ensure that the amplitude is correct for the specified frequency. The copy synthesis method has been used to obtain good quality female synthetic speech, so demonstrating that there is no intrinsic limitation in the JSRU parallel formant synthesizer for producing female speech.

8. ACKNOWLEDGEMENTS

This work was carried out under the Alvey project MMI-009, "Speech Interfacing and Phonetic Algorithms".

9. REFERENCES