Experiments in Female Voice Speech Synthesis Using a Parametric Articulatory Model

Dongbing Wei and C. C. Goodyear
Department of Electrical Engineering and Electronics
University of Liverpool, U.K.

1. ABSTRACT

A parametric vocal tract model and a two dimensional articulatory parametric subspace for a female voice are presented. The parameters of the model, which determine the vocal tract shape, can be found uniquely for VV transitions by mapping directly from f1 and f2 onto this subspace, while a modified technique involving f3 is available for voiced VC and CV diphones. The area functions of the vocal tract, generated by these parameters, are used to drive a time-domain synthesiser. Synthesis to give female speech, copied from either male or female natural speech, may be performed.

2.INTRODUCTION

There is a continuing demand for natural sounding synthetic speech for man-machine interfaces. Most of the work in producing synthetic voices has so far been concentrated on the male voice. However, a variety of voices has always been demanded in different speech synthesis applications. Originally, the synthetic female voice did not sound convincing. Achieving a good female voice has attracted much effort and the quality of the synthetic speech continues to improve [1][2].

Articulatory speech synthesis is known to be a method capable of producing very good quality speech. Due to the non-uniqueness of acoustic to articulatory mapping, the selection of vocal tract shapes during speech synthesis presents a substantial computational load. In previous work, we have proposed a method which effectively defines a two dimensional subspace of the articulatory parameter space, accessed by formant frequencies f1 and f2 and giving smoothly connected vocal tract shapes among the vowels. Mapping from acoustic features to articulatory parameters is unique and much simpler in this subspace. The method has achieved high quality synthetic speech for a male speaker with very low computational cost. Further details are reported in [6-8].

In this paper, a synthesizer for a female sounding voice based on the above technique is proposed. The method includes the use of a simple nine parameter vocal tract model, a two dimensional vowel subspace for a female voice and an interpolation algorithm. We have found this method to be capable of synthesising a high quality female sounding voice with very low cost in computational load and storage.

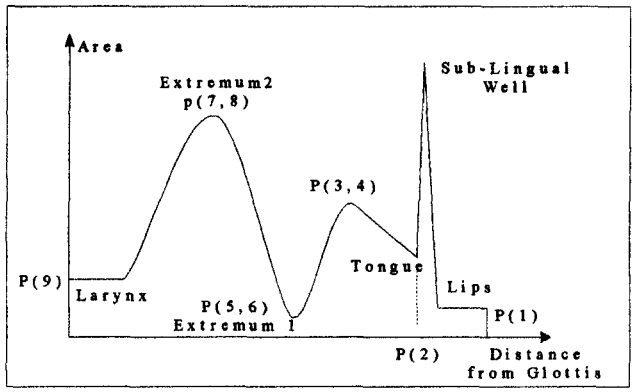

Figure 1. Nine-parameter vocal tract model

3. PREVIOUS WORK AND ACHIEVEMENTS

The parameters of our vocal tract model [6][7] shown in figure 1 comprise P1, the lip area, P2-P4, the area at the tip of the tongue and the area and location at the rear of the tongue blade, P9, the area near the larynx and P5-P8, the positions and areas at two extrema between P4 and the larynx. These are used to generate a smooth curve for the area function and can be adjusted to match known shapes quite well.

The vocal tract area functions for nine vowels, i.e. /i:/, /æ/, /ɐ/, /ɔ/, /u/, /ə:/, etc., uttered by a male speaker were measured by a magnetic resonance imaging (MRI) technique. The corresponding articulatory parameters of the above model for these nine vowels have then been calculated and are known for each of the points in our two dimensional articulatory vowel subspace in figure 4.

Copy synthesis of a male voice was successfully achieved by driving our articulatory model with smoothly evolving vocal tract parameters derived from formant frequencies. The evaluation of the parameters in the subspace at each pitch period was performed using a two dimensional interpolation method. The spectrograms of the natural and synthetic speech are shown in figure 6 and figure 7 respectively.

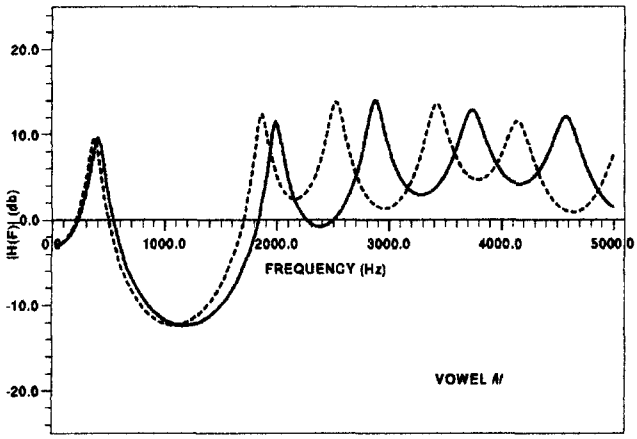

Figure 2. The spectral of the vowel /i/.
(.... Male voice, - female voice)

4. EXPERIMENTAL WORK

Much of the preparatory work had concentrated on the specification of vocal tract parameters to produce the proper formants. The success of the vowel subspace in producing a male voice [6-8] has been taken as the basis on which to define a similar female articulatory subspace. We chose this route in the absence of MRI data for a female vocal tract. The parameter vowel subspace can then be defined by the method introduced in [7].

The vocal tract length, which was 21 sections for our male voice, has been shortened to 19 sections. For the vowel / i:/ the vocal tract male distance parameters were scaled by the same factor and used in our synthesiser as a approximation for a female /i:/ configuration. These were then adjusted by an optimisation procedure to match the formants of an available female voice. By applying a similar scaling procedure to the male parameters of the other eight primary vowels, the primary points shown in figure 5 were produced. The adjustment has been well controlled so that the formant values of the nine vowels are now about 10%-15% higher than the ones we measured earlier from a male voice and the resulting synthetic vowels are female sounding. The spectra of the synthetic vowels /i:/ and /e:/ produced in this way are show in figure 2 and 3, repectively.

The same scaling procedure was used on the secondary points in the male articulatory subspace, shown in figure 4, to define a female subspace which is also accessed by f1 and f2. The resulting formant frequencies f1, f2 and f3 for the vowels and secondary points in the new vowel subspace plotted in figure 5, are about 10%-15% higher than in the original male vowel subspace. This shift is in line with that reported by Karlsson[2].

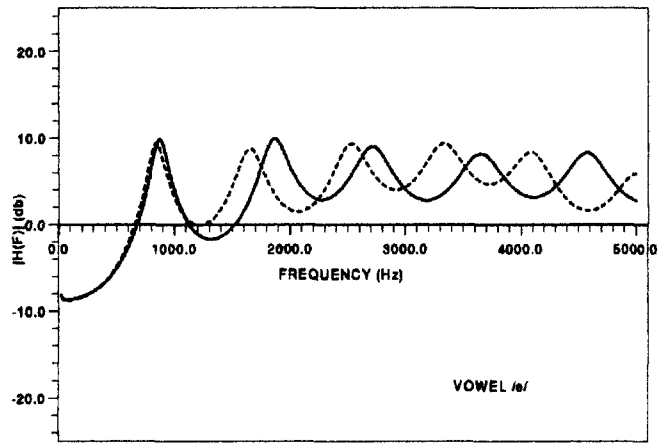

Figure 3. The spectral of the vowel /e/.
(.... Male voice, - female voice)

5. THE INTERPOLATION

An interpolation parameter surface P(f1, f2), though a set of m scatter primary and secondary points (f1, f2, P_r), for r=1, 2,...m, using a modification of Shepard's method [9], was constructed. The surface is continuous and has continuous first derivatives.

The basic method is global in that the interpolated value of the parameter vector P at any point depends on all the data, but the modification method is local. The behaviour of this interpolation technique is well suited to the required mapping, where sharp changes in the parameters should be avoided.

6. SYNTHESIS OF FEMALE VOICE

Figure.6 and figure.7 show the spectrograms of the utterance "Hide away a gory rag" by a male speaker and the copy synthetic speech obtained by our articulatory model [8], respectively.

To obtain a female version of this utterance, the following parameters are required for every pitch period:

(1) the area functions of the new vocal tract,

(2) a suitable glottal pulse,
(3) the proper female fundamental frequency.

The pitch markers of the male speech were extracted from the natural male utterance and pitch-synchronous autocorrelation analysis was used to estimate values for the first three formants at each pitch cycle. The first two formant frequencies were scaled up about 12% and the fundamental frequency was scaled by a factor 1.7. The area function profile is obtained from the nine-parameter articulatory model(as shown in fig.1) in which the parameters can be obtained by one of the follow procedures:

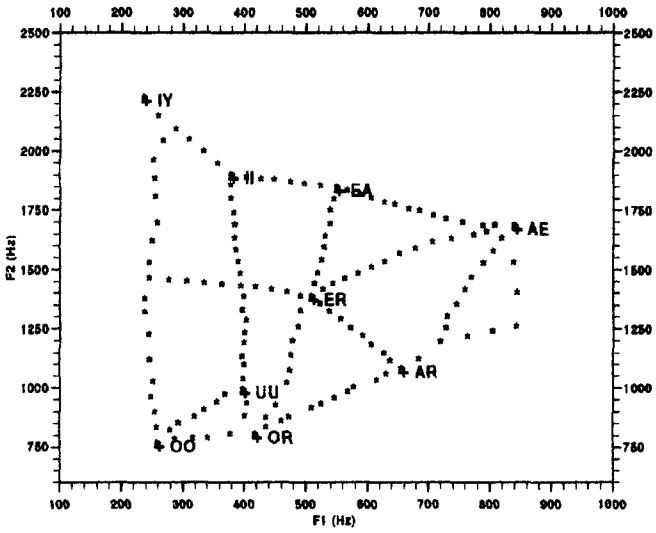

Figure 4. The vowel sub-space for male voice.
(+ Primary points; * secondary points.)

a) the parameters of the vocal tract model (which produced the synthetic male speech shown in figure7), were evaluated from the first two formant frequencies using the interpolation method in the subspace in figure 4 for the voiced diphones. These parameters can be adjusted by the same algorithm which defined the vowel subspace figure 5. The synthesiser controlled by these adjusted parameters produces speech with a convincing female quality.

b) the estimated first two formant frequencies are scaled up by a factor about 1.12 to match a female voice. With the new formants values the parameters of the model were evaluated by the Shepard interpolation method referred to above based on the proposed female vowel subspace shown in figure 5.

This procedure was modified for the VC and CV diphones in the way described in [8].

The excitation source is also considered to be one of the important factors in female voice synthesis. The glottal pulse used in the female voice synthesis system here is defined by the LF-Model. Their parameters were adopted from Karlsson's investigations of a female voices source[4] and their amplitudes were adjusted to match the energy in each pitch cycle with that in the natural speech.

Note that although the pitch was scaled by the factor 1.7, the total duration of the utterance was unchanged.

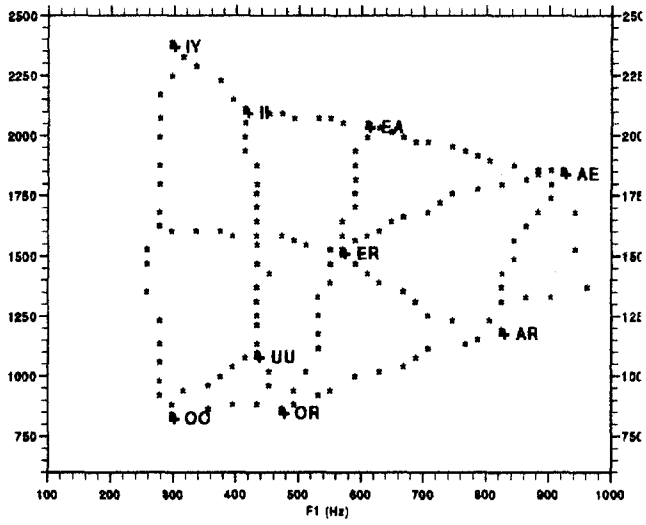

Figure 5. The vowel sub-space for female voice.
(+ Primary points; * secondary points.)

Figure 8 shows a spectrogram of synthetic female speech.

7. DISCUSSION

The proposed method for female voice synthesis is not a simple transformation of the original male voice speech. A parametric articulatory model and a parameter subspace for female voice have been defined based on our male speaker articulatory speech model. The parameter vowel subspace in figure 5 provides smoothly evolving articulatory parameters for copy synthesis of a voiced female speech. The system may also be used for copy synthesis of female voiced speech utterances, driven by the values of acoustic and articulatory features (1), (2) and (3) in section 6, which can be obtained directly by analysing this female natural speech, together with the mapping of figure. This seems a successful alternative to measuring the vocal tract shape by the MRI technique with a female speaker[6-8].

8. ACKNOWLEDGEMENT

The authors are grateful to EPSRC for financial support.

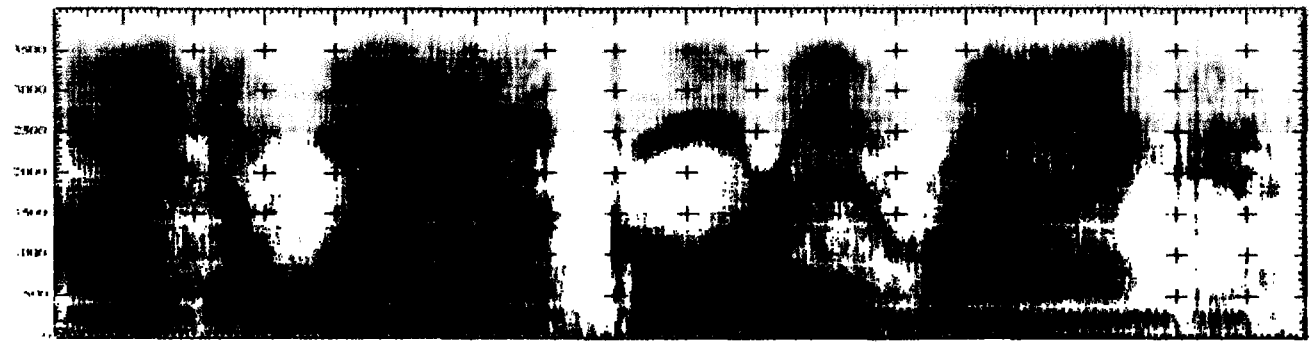

Figure 6. The spectrum of the natural male voice speech.

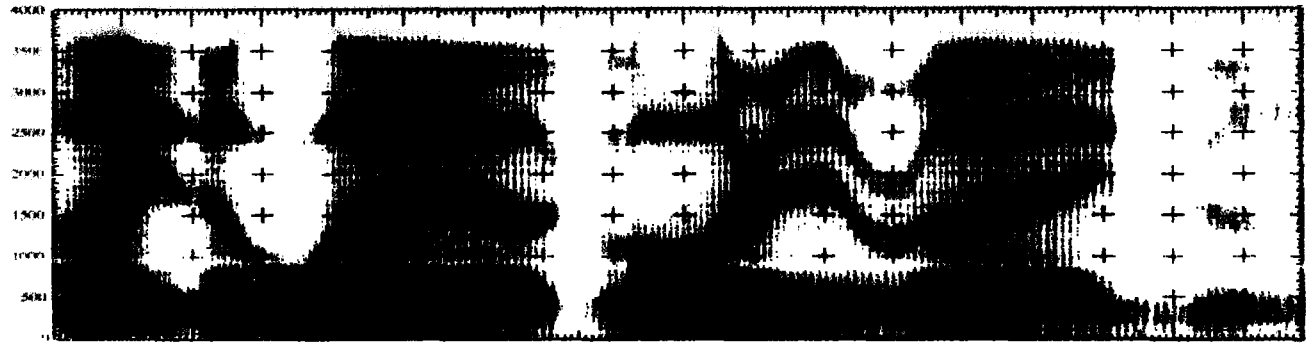

Figure 7. The spectrum of the synthetic male voice speech.

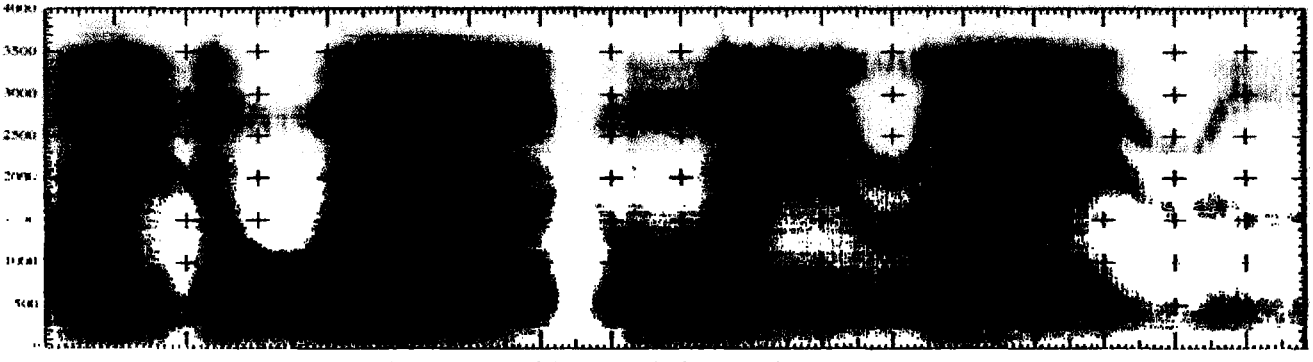

Figure 8. The spectrum of the synthetic female voice speech

REFERENCES

[1] Holmes, W.J. Copy Synthesis of female speech using the JSRU Parallel formant synthesizer. Eurospeech'89, Paris, Vol.2, pp513-516.

[2] Karlsson, I. Female Voices in Speech Synthesis. Journal of Phonetics 19, 1991 pp111-120

[3] Price, P. Male and Female voice source characteristics: Inverse filtering results. Speech Communication, 8, pp261-277.

[4] Karlsson, I. A Female Voices for Text-to-Speech System. Eurospeech'89, Paris, Vol.1, pp349-351.

[5] Carlson, G., Fant, G., Gobl, C., Granstrom, B., Karlsson, I and Lin, Q-G.,Voice source rules for Text-to-Speech Synthesis., Proc. ICASSP89, Vol.1 pp223-227, 1989.

[6] Wei, D., Devaney, J.W. and Goodyear, C.C., Voiced Diphone Synthesis Using a Parametric Articulatory Model and Formant Based Mapping. Proc. of European Conference on Speech Communication and Technology, EUROSPEECH95. Madrid, Spain, 1995.

[7] Wei, D. and Goodyear, C.C., A Novel method of Mapping from Acoustic Feature Articulatory parameters, International Conference Multimodel Interface'96. Beijing, 1996.

[8] Goodyear, C.C and Wei, D. 'Articulatory Copy Synthesis Using a Nine-Parameter Vocal Tract Model. ICASSP'96, Atlanta.

[9] Shepard, D. A two-dimensional interpolation function for irregulary spaced data. Proc 23rd. Nat. Conf. ACM, Braudon/System Press Inc., Princeton.