TRANSFORMATION OF SPEAKER CHARACTERISTICS FOR VOICE CONVERSION

Dimitrios Rentzos Saeed Vaseghi Emir Turajlic Qin Yan Ching-Hsiang Ho*

Department of Electronics and Computer Engineering Brunel University, Middlesex UB8 3PH, UK
*Fortune Institute of Technology, Kaohsiung, Taiwan, 842, R.O.C.
(Emir.turajlic, Dimitrios.Rentzos, Saeed.vaseghi)@brunel.ac.uk, ch.ho@center.fjtc.edu.tw

ABSTRACT

This paper presents a voice conversion method based on analysis and transformation of the characteristics that define a speaker's voice. Voice characteristic features are grouped into three main categories: (a) the spectral features at formants, (b) the pitch and intonation pattern and (c) the glottal pulse shape. Modelling and transformation methods of each group of voice features are outlined. The spectral features at formants are modelled using a two-dimensional phoneme-dependent HMMs. Subband frequency warping is used for spectrum transformation where the subbands are centred on estimates of formant trajectories. The F0 contour, extracted from autocorrelation-based pitchmarks, is used for modelling the pitch and intonation patterns of speech. A PSOLA based method is used for transformation of pitch, intonation patterns and speaking rate. Finally a method based on de-convolution of the vocal tract is used for modelling and mapping of the glottal pulse. The experimental results present illustrations of transformations of the various characteristics and perceptual evaluations.

1. INTRODUCTION

The aim of voice conversion is to transform the voice of a *source* speaker towards a *target* speaker's. Voice conversion has applications in all voice output systems such as text to speech synthesis, voice editing for films, Karaoke, broadcasting and Internet voice applications.

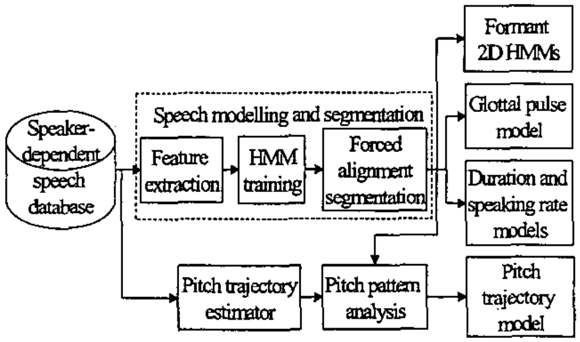
Speaker-dependent speech database
Speech modelling and segmentation
Feature extraction
HMM training
Forced alignment segmentation
Formant 2D HMMs
Glottal pulse model
Duration and speaking rate models
Pitch trajectory estimator
Pitch pattern analysis
Pitch trajectory model

Figure 1: Voice Modelling Procedure

Voice conversion requires modification of the spectral and the temporal correlates of the voice characteristics of the speaker. An effective voice conversion system would require three essential components:

(a) *Voice Feature Extraction.* Voice features include laryngeal, supra-laryngeal, intonation and delivery correlates. The laryngeal features are pitch and glottal pulse shape parameters. The supra-laryngeal parameters are the frequencies, bandwidths, and intensities of the resonance at formants. The prosodic and delivery style parameters include pitch trajectory features, duration and speaking rate.

(b) *Model Estimation.* For modelling the space of the acoustic features of the source and target speakers, a number of alternatives including vector quantized codebooks [2,4], Gaussian mixture models or hidden Markov models (HMM) may be used [1,6]. In this work, we use HMMs for estimation of the statistical distributions of spectral and temporal voice correlates of the source and the target speakers.

(c) *Voice Mapping.* Using the differences between the source and target voice features, the trajectories of the features of the source speaker are modified to towards those of the target speaker [3].

This paper is organised as follows. In section 2 the main stages of the voice conversion method are outlined. Section 3 the estimation and transformation of the spectral features. Section 4 analyses the pitch intonation and timing features. Section 5 describes the glottal pulse estimation and transformation method. Section 6 describes a number of voice features conversion experiments and section 7 concludes the paper.

2. OVERVIEW OF THE MODEL-BASED VOICE CONVERSION SYSTEM

Figure 1 illustrates the process for estimation and modelling of voice features. The voice modelling process involves the following steps:

(a) Vocal tract feature extraction, derives MFCC features and also formant candidate features comprising of the bandwidths, intensities and frequencies of the poles of LP model of speech.

(b) Cepstrum HMM training. At this stage conventional speaker-dependent HMMs of phonemic units of speech are trained on cepstrum features.

(c) Viterbi decoding and Segmentation. Using speaker-dependent HMMs and the 'forced-alignment' Viterbi decoding, the speech is processed to obtain phoneme boundary estimation and segmentation.

(d) Formant HMMs. After segmentation of speech database, the formants candidates (obtained from the poles of LP model) associated with each state of HMMs are modelled using an *M*-state formant HMM. This effectively results in a two-dimensional HMM.

(e) Glottal pulse estimation. The parameters of the LF (Liljencrants/Fant) model are used for the characterisation of the glottal pulse. The estimates of the spectral features are used in a de-convolution process to obtain an estimate of the glottal pulse shape.

(f) Pitch estimation. Pitch trajectory is estimated using autocorrelation feature analysis. This is followed by a pitch pattern analysis which is used to model parameters of the shape of the pitch trajectory.

(g) Duration model. The results of speech segmentation are used to obtain the statistics of the variations of speaking rate and phoneme duration patterns.

The statistical models of the features, for a pair of source and target speakers, are used to calculate the parameters for voice conversion. The voice conversion process illustrated in figure 2 involves the following steps:

a) Source-filter separation. Two methods are used for the source-filter separation of the speech signal. The first method is based on standard LP inverse filtering. The second is based on exact glottal pulse shape estimation and deconvolution of the speech spectrum.

b) Vocal-tract spectrum transformation. The frequency spectrum of the vocal tract of the source speaker is mapped towards the target either by frequency warping or by rotation of the poles of the LP-model.

c) Glottal pulse mapping. The shape of the glottal pulse of the source voice is changed towards the target voice through mapping the parameters the source glottal pulse towards the target glottal pulse.

d) Pitch modification based on TD-PSOLA. The speech excitation of source speaker is fed to a TD-PSOLA function for modification of pitch, duration and prosody towards those of the target speaker.

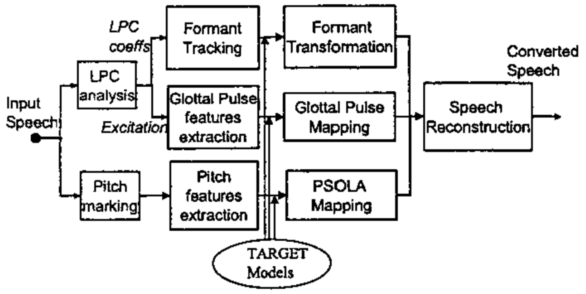

Figure 2: Voice Conversion

e) Speech reconstruction. The transformed excitation and vocal tract can be combined in frequency domain and then converted to time using inverse Fourier transform. Alternatively the modified spectrum of the vocal tract can be converted to an LP filter using an inverse Fourier transform to yield the autocorrelation function followed by Levinson algorithm to yield LP coefficients. The modified excitation is then filtered by the LP model.

3. ESTIMATION AND TRANSFORMAITON OF FORMANT MODELS

The high variability of the number of formants across time and phonemes requires accurate formant model estimation. A 2-D HMM with *N* left-to-right states distributed across frequency, and *M* states distributed across time to classify formant observations [1,6] as shown in figure 3. The formant features at time *t* are the formant frequency F_t, the bandwidth of resonance BW_t and the intensity of resonance I_t and their respective temporal difference values. The LP pole coefficients are the raw data from which the formant features are estimated. Since there is not a simple one-to-one correspondence between the poles of the LP model and the formants of the vocal tract, a process of constrained clustering and classification of poles and estimation of formants is developed. The constraints are imposed by the sorting of the elements of the formant feature vectors in terms of increasing frequency and the use of a left-right HMM along the frequency axis.

The formants within each HMM state are modelled by mixture Gaussian distributions. Given the phoneme-dependent HMM formant classifier and the associated formant bandwidths, formant trajectory estimation for a speech waveform is achieved through minimisation of a weighted mean square error objective function as

$$\hat{F}_k(t) = \min_{F_k(t)} \sum_{i=1}^{I_k(t)} w_{ki}(t) \left(\frac{(F_i(t) - F_k(t))^2}{BW_i(t)^2} \right) \qquad (1)$$

where $I_k(t)$ is the total number of poles in the t^{th} speech frame classified as formant *k* and w_{ki} is a probabilistic weight derived from the model.

3.2 Parametric Formant Mapping Methods

Frequency warping through non-uniform adaptive sub-band spectral mapping: In the spectrum warping method the formant frequency estimates are used to divide the signal spectrum into *N* sub-bands according to the formant positions. The inputs to the spectrum mapping function are the LP-spectrum *X*(*f*), and the formant feature vector of the current source speech frame. The formant feature vector contains the formant frequencies, bandwidths and

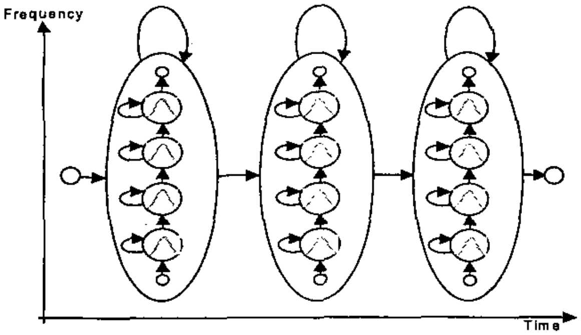

Figure 3: A 2-D HMM formant model

intensities and is used for derivation of the spectral mapping function. The equation for voice conversion through spectrum mapping is expressed as

$$Y[f,t] = \gamma(f,t)\, X[\alpha(f,t) * \beta(f,t) * f] \qquad (2)$$

where *X*, *Y*, *t*, and *f* denote the source spectrum, the transformed spectrum, and the time and frequency variables respectively. The composite frequency warping function includes the mapping functions for both the formant frequency *α(f,t)*,and bandwidth *β(f,t)* and *γ(f,t)* is the intensity shaping function used to map the energy intensities between the source and the target formant frequencies.

The frequency warping ratio *α(f,t)*, is derived from the ratio of the differences between the successive formant frequencies of the target speaker to that of the source speaker:

$$a(f,t) = \frac{F^T_{f+1,t} - F^T_{f,t}}{F^S_{f+1,t} - F^S_{f,t}} \qquad (3)$$

Similarly the bandwidth and intensity warping ratios are derived from the target to source ratios of the bandwidth and intensity of oscillation values, at each formant respectively.

Transformation of the frequency response of an LP model through rotation of the poles of the model:

In the pole rotation method, the poles' frequencies of the source speaker are associated with the formants and then rotated towards the position of the formants of the target speaker. The transformation of the formant position and shape can be achieved by moving the poles in the LP spectrum of speech. The frequency of the formant depends on the angle of the pole, and its bandwidth on the pole radius. The equations describing the desired modifications in frequencies and bandwidths are derived by

$$\varphi(p_i') = \alpha(i,t) \times \frac{Fs}{2\pi} \times \varphi(p_i) \qquad (4)$$

$$\mathrm{abs}(p_i') = \beta(i,t) \times \log(abs(p)) \times \frac{Fs}{\pi} \qquad (5)$$

where F_s is the sampling frequency, $\varphi(p)$ is the angle of pole *p*, *p* and p_i' are the original and mapped poles, and *α(i,t)*, *β(i,t)* the frequency and bandwidth mapping ratios derived using the same procedure as in the frequency warping method. The intensity of each formant is adjusted after the poles have been moved using the spectral shaping procedure described in the previous section (eq. 2).

4. ESTIMATION AND TRANSFORMATION OF PITCH INTONATION AND DURATION PATTERNS

A part of the voice identity and speaking style of a person is keyed into pitch and its characteristic broad patterns of variation over time. As with the Tilt model [7], the pitch and intonation analysis model proposed here models continuous patterns of F0 curve. The problem we are faced with is how to device a symbolic representation of the F0 curve and what set of model parameters to use. A set of features are specifically selected to represent the broad characteristics of F0 curve and to provide an effective way of modelling the most important pitch intonation characteristic features of a speaker. The pitch/intonation model is based on a rise/fall/connection (RFC) model. A RFC model is defined as representing the pitch contour using a sequence of rising and falling pitch segments with straight lines applied for the intervals without pitch values (unvoiced segments). Figure 4 shows an illustration of the pitch intonation model. A schematic representation of F0 is shown from which the pitch modelling parameters are derived. The features used for modelling the pitch and intonation patterns of a speaker are the following:

Average pitch, $F0_{av}$: The average or mean pitch value is obtained from estimates of pitch tracks of recorded examples of a voice.

Pitch range: A value of three times the standard deviation is used to model the pitch range of variation about the mean value. A multiple of the standard deviation is considered as less sensitive to estimation errors than maximum and minimum pitch values.

Pitch slopes: Three different kinds of pitch slopes are considered over the duration of an intonation phrase. These are: (a) the average pitch slope (*∂F0phrase*). It is found from the slope of F0 curve across the entire length of an intonation phrase. (b) The initial pitch slope. It is obtained form the slope of the first pitch segment of a phrase, which is usually rising (*∂F0initial*). (c) Finally, the final pitch slope (*∂F0final*), is obtained from the final part of the intonational phrase. The final pitch slope of a phrase plays an important role in differentiating between a question or a statement and between different regional and national accents.

Pitch accent parameters: a pitch accent is defined as a segment of rising or falling pitch. The average slope of pitch accents is considered as a characteristic parameter of a speaker's voice. Accent slopes are obtained from the RFC analysis of the speakers' F0 contours.

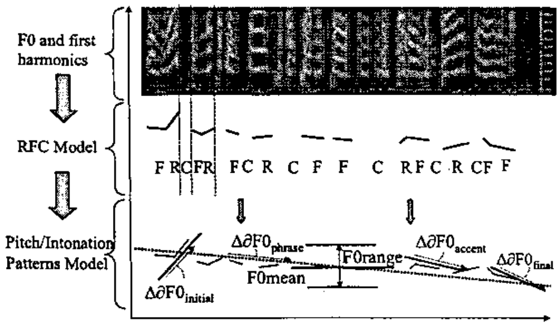
F0 and first harmonics
RFC Model
F RCFR FC R C F F C RFC R CF F
Pitch/Intonation Patterns Model
$\Delta\partial F0_{phrase}$
$\Delta\partial F0_{accent}$
F0range
$\Delta\partial F0_{final}$
F0mean
$\Delta\partial F0_{initial}$

Figure 4: The pitch intonation model

The variables that affect the style of delivery of speech also include speaking rate, phoneme duration pattern, hesitation pattern and pattern of pitch modulation over time. Specifically the following duration and timing parameters of speaker's voice are quantified:

(a) Speaking rate: the number of spoken words per second.
(b) Phoneme duration pattern.
(c) Pause rate: number of pauses per second.

A pitch transformation method, based on time-domain pitch synchronous overlap and add method (TD-PSOLA), is developed that allows independent modification of any number of the pitch intonation and timing parameters.

5.ESTIMATION AND TRANSFORMATION OF GLOTTAL PULSE

Laryngeal correlates affect the voice quality of speech. Depending on the laryngeal parameters the voice can have a modal (normal), pressed, breathy, creaky, harsh or whispery quality. The quality of voice is affected by the shape, volume and frequency of the pulse flow. The LF (Liljencrants/Fant) model is used for glottal pulse modelling. The glottal pulse has two distinct time characteristics: the open quotient ($OQ=T_c/T_0$) is the fraction of each period the vocal folds remain open and the skew of the pulse or speed quotient ($\alpha=T_p/T_c$) is the ratio of T_p, the duration of the opening phase of the open phase, to T_c the total duration of the open phase of the vocal folds. To complete the glottal flow description, the pitch period, the rate of closure ($RC= (T_c-T_p)/T_c$) and the minimum to maximum pulse amplitude ratio (E_e/E_c) are also included.

Estimation of the five LF parameters its parameters requires first an estimation of the glottal closure instant. The estimation of the glottal closure instant (GCI) exploits the fact that the minimum phase signal has an average group delay value proportional to the shift between the start of the signal and the start of the analysis window, and that at the instant when the two coincide, the average group delay is of zero value.

The LF parameters are then obtained from an iterative application of a dynamic time alignment method to an estimate of the glottal pulse sequence. The method uses a pitch synchronous linear prediction model where the vocal tract parameters are obtained from periods of zero-excitation coinciding with the close phase of a glottal pulse cycle. The parameterization process can be divided into two stages:

(a) Initial estimation of the LF parameters. An initial estimate of each parameter is made. For example t_e corresponds to the instant when the glottal derivative signal reaches its local minimum. E_e is the magnitude of the signal at this instant. t_p can than be estimated as the first zero crossing from the left of t_e.
(b) Constrained non-linear optimisation using a minimum mean squared error criterion. A dynamic time warping (DTW) method is employed. A time alignment technique is used because, in the LF parameter extraction process perceptually the important aspects of the glottal pulse derivative are its timing parameters.

Having estimated the glottal pulse parameters and separated the spectra of the glottal source and the vocal tract, the glottal pulse is transformed by using the target's glottal parameters to generate a new excitation.

6.EXPERIMENTS

In this section the experimental procedures and results are described. The two main themes of the paper evaluated in the experiments are: (a) analysis of the speaker features to be transformed (b) experiments on the transformation of those features for voice conversion.

A number of experiments were conducted on a set of five speakers including three male American English speakers, and two female American English speakers. The test speakers' databases consist of 140 spoken sentences per speaker, with a sampling rate of 10 kHz. The speech is pre-emphasised with a first order pre-emphasis filter and segmented into 25 ms long overlapping segments with an overlap of 15 ms. Each speech segment is windowed and modelled by an LP model order of 13. An LP model of 13 can model up to 6 formants, this is considered to be the maximum number of formants in speech.

Evaluation of Feature Models

The Hidden Markov model Toolkit (HTK) developed at Cambridge University is used for training HMMs and decoding speech [5]. In the first stage of speech processing each 25 ms speech segment is converted to a 39 dimensional feature vector comprising of 13 cepstrum, 13 delta cepstrum and 13 delta-delta cepstrum features. The HTK software is then used to train speaker-dependent phoneme-dependent HMMs, with each HMM composed of 3 states each modelled by 20 Gaussian mixture distributions. The average phoneme recognition rate with these speaker-dependent HMMs is about 98%. Then, the HTK implementation of, the Viterbi state-decoder is used with the phonetic transcription supplied, to obtain the phonetically labelled segment boundaries of. This combination results in a minimisation of errors in the estimation of phonetic segment boundaries.

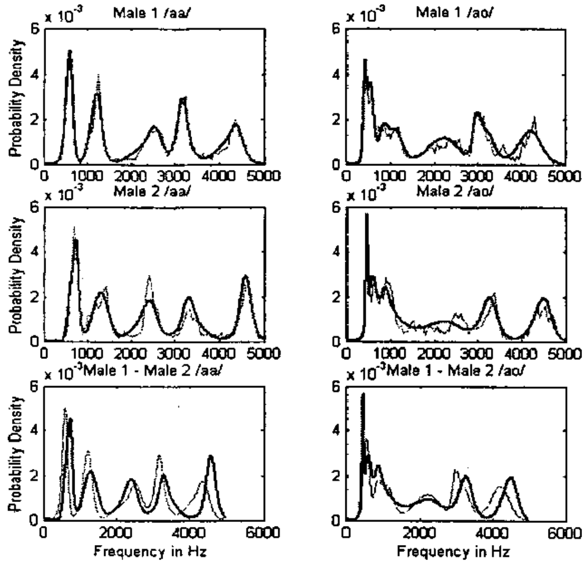

Figure 5: Frequency histograms – Formant HMM Models

The phonetic segment boundaries are then used for training formant HMMs. The distributions of the formants are modelled by phoneme-dependent HMMs trained on formant feature vectors obtained from the poles of LP model of speech segments. The elements of each LP feature vector, comprising of the frequency (Hz), bandwidth (Hz) and intensity (dB) of resonance at the poles, are ordered in terms of ascending frequency of the poles. Formant HMMs for each segmented state of phonemes are trained, again using HTK. Each formant HMM has five states to model five formants in speech. The distribution of formants in each state is modelled by a 4 mixture Gaussian pdf. The overall result is a set of speaker-dependent phoneme-dependent formant 2-D HMMs with (3) states along time and (5) states along frequency. Figure 5 shows the comparative plots, for two vowels /aa/and /ao/, of histograms of the pole frequencies superimposed on the Guassian mixture model extracted from the states of HMMs of formants. Figure 5 clearly illustrates that the Gaussian mixture models obtained from the states of HMMs of formants provide a smooth close fit to the histograms. Hence HMMs, in addition to being a good model of cepstrum features, are also a good probability model of the distributions of the formants along the frequency axis. The two bottom illustrations of figure 5 compare the formant models for the two speakers. It can be seen that their differences are formant specific. Similar comparisons between the speakers voice features show clear differences in the pitch intonation and timing features.

Conversion illustrations

Formant Tracking – Comparison of formant tracks to LP Spectrograms: The formant tracks are obtained by classification and estimation of the formants using the probabilistic formant model. The success of the formant tracker is illustrated by the superimposition of the estimates of formant tracks of speech segments on their LP-spectrograms. In Figure 6 the first illustration shows the superimposed formant tracks for the source, and the second for the target, both saying “too narrow”. It can be seen that the estimates of the formant tracks accurately follow the actual trajectory of the tracks as indicated by the trajectories of LP-spectrograms. In the third illustration the two tracks and the track of the converted sentence are time aligned and displayed. From this figure it can be seen that the converted track follows the target. Figure 7 shows the mapping of the frequency spectra of source speaker to a target speaker for two frames. The success of the transformation is almost perfect.

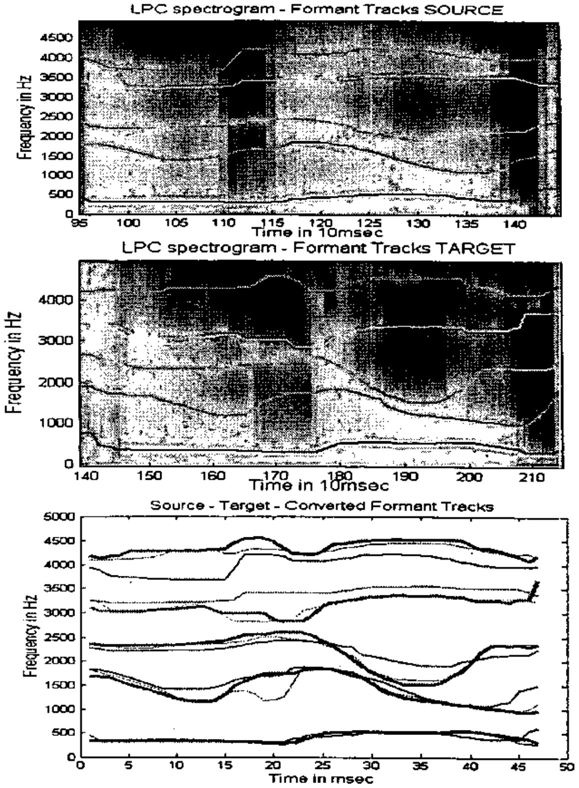

Figure 6: LPC spectrogram with the superimposed formant tracks and transformation of formant track

Pitch Slopes: An illustration of the F0 contour before and after it has been transformed is shown in figure 8. All parameters, i.e. average pitch, pitch range pitch slopes and duration are mapped.

Perceptual experiments

Two different perceptual tests, based on mean opinion score (MOS) were performed to evaluate the vòice conversion method and their results are shown in tables I and II. The tests were focused on formant and pitch features transformation. Tests comparing the importance of the

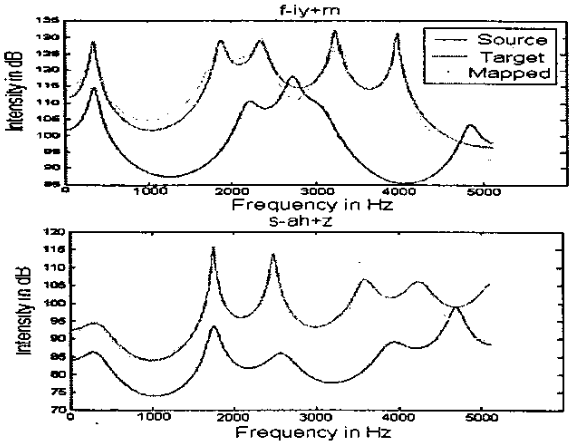

Figure 7: Transformation of frequency spectra for /iy/, /ah/. Warping ratios for /iy/: α=[0.97; 0.85; 0.87; 1.06; 0.82], β=[0.7823; 0.7; 0.9; 0.4; 0.4], γ=[5; 9.5; 2.5; 9; 15.5]

voice features and including transformations of the glottal pulse will be performed in the next stage of the research.

In the first test five words were selected, and the formants of a source speaker were modified towards those of the voice of a target speaker. The information about the target speakers was derived from the targets speakers' HMMs. The subjects were asked to listen to three utterances; the original, the transformed and the same word uttered by the target and asked which of the two speakers uttered the transformed version. The subjects were asked to give a score from 1 to 10 with 1 signifying the least similarity between the transformed speech and the source or target speech, and 10 signifying the highest similarity. Two variants of that test were conducted. In the first instance only the formants were changed and in the second the other parameters of the resonance at the formants, namely the bandwidths and intensities, were also transformed. From table I shows that formants play an important part in the perception of a speaker thus formant transformation is needed to achieve voice conversion. It is also interesting to note the significance of the additional formant parameters (bandwidth, intensity and spectral tilt) in speech conversion.

In the second test (table II) entire sentences were converted.

Table I: Formant Conversion: Mean Opinion Scores

	Source	Target
Only Frequency	4.75	5.58
All Formant Parameters	3.91	6.20

Score of similarity of changed segment to original and target speakers

Table II: Voice Conversion Mean Opinion Scores

	Source	Target
Words	3.87	6.62
Sentences	3.5	6.13

Formant and average pitch change of entire sentences. Score of similarity of changed speech to original or target speakers

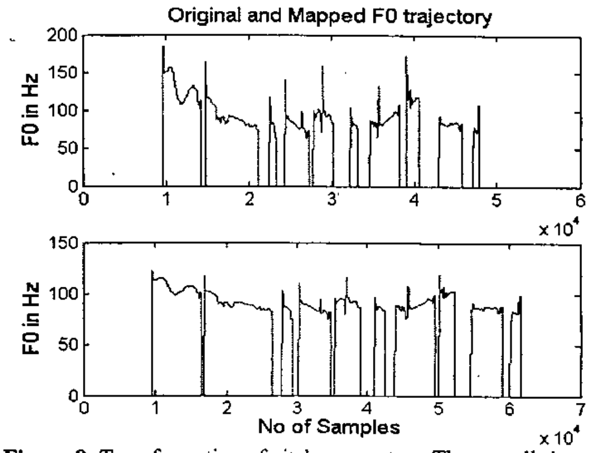

Figure 8: Transformation of pitch parameters. The overall shape is preserved. Modified parameters: average pitch from 85Hz to 100 Hz, duration increase by 30%, pitch range decrease by 50%, phrase slope from –30Hz/sec to –12Hz/sec.

Ten sentences were converted and American males were chosen as source and target speakers. The subjects were asked to decide which of two speakers (source or target) was more likely to have uttered the converted sentence and again give scores from 1 to 10. The final test shows that the proposed formant transformation system is successful for voice conversion. On average, the sentences resembled more like the target speaker's (6.13) and less like the source speaker's (3.5), although the scores were lower than in the single words case (6.62 and 3.87 respectively).

7. CONCLUSION

This paper presented the voice features that need to be converted for voice conversion. Methods for transforming those features were analysed. It was shown that successful voice conversion can be achieved.

REFERENCES

[1] Acero A. (1999), "Formant Analysis and Synthesis using Hidden Markov Models", Proc. of the Eurospeech '99.

[2] Arslan L.M. and Talkin, D, (1997) Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum", EUROSPEECH 1997 Proceedings.

[3] Kain A., Macon M., "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction". Proceedings of ICASSP, May 2001.

[4] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., (1988) Voice Conversion Through Vector Quantization, Proceedings of the IEEE ICASSP 1988, pp. 565-568.

[5] Woodland, P.C. Young, S.J. (1993). The HTK Continuous Speech Recogniser. Proceedings Eurospeech 1993, pp2207-2219.

[6] Dimitrios Rentzos, Saeed Vaseghi, Qin Yan, Ching-Hsiang Ho*, Emir Turajlic, "Probability Models of Formant Parameters for Voice Conversion", in Proc. Eurospeech 2003, pp. 2405-2408.

[7] Taylor P. A. (2000), "Analysis and Synthesis of Intonation Using the Tilt Model" Journal of Acoustical Society of America" Vol. 107 pp. 1697-1714.