# VOICE CONVERSION THROUGH TRANSFORMATION OF SPECTRAL AND INTONATION FEATURES

Dimitrios Rentzos Saeed Vaseghi 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.
(Dimitrios.Rentzos, Saeed.vaseghi, Qin.Yan)@brunel.ac.uk, ch.ho@center.fjtc.edu.tw

## ABSTRACT

This paper presents a voice conversion method based on transformation of the characteristic features of a source speaker towards a target. Voice characteristic features are grouped into two main categories: (a) the spectral features at formants and (b) the pitch and intonation patterns. Signal modelling and transformation methods for each group of voice features are outlined. The spectral features at formants are modelled using a set of two-dimensional phoneme-dependent HMMs. Subband frequency warping is used for spectrum transformation with the subbands centred on the estimates of the formant trajectories. The F0 contour is used for modelling the pitch and intonation patterns of speech. A PSOLA based method is employed for transformation of pitch, intonation patterns and speaking rate. The experiments present illustrations and perceptual evaluations of the results of transformations of the various voice features.

## 1. INTRODUCTION

The aim of voice conversion is to transform the voice of a *source* speaker towards that of a *target* speaker. Voice conversion has application in text to speech synthesis, voice editing for films, Karaoke, broadcasting and multimedia voice applications.

An effective voice conversion system would require three essential components:

(a) *Voice Feature Extraction.* Voice features include; supra-laryngeal parameters, i.e. the frequencies, bandwidths, and intensities of the resonance at formants; laryngeal (glottal) parameters; and intonation 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 [1], hidden Markov models (HMM) or Gaussian mixture models may be used [2,3]. 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 difference between the source and target voice features, the trajectories of the features of the source speaker are modified towards those of the target speaker.

This paper is organised as follows. Section 2 presents the voice conversion method. Sections 3 and 4 describe the estimation and transformation of the spectral features and of the pitch, intonation and timing features. Section 5 describes voice conversion experiments and section 6 concludes the paper.

## 2. OVERVIEW OF THE MODEL-BASED VOICE CONVERSION SYSTEM

The parameters required for voice conversion are obtained from the statistical models of the source and target speakers. The process of estimation and modelling of voice features illustrated in Figure 1 involves the following steps:

(a) Vocal tract features extraction. Derives MFCC features and formant candidate features comprising of the bandwidths, magnitude 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 'forced-alignment' Viterbi decoder, speech is segmented and phoneme boundaries estimated.

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

(e) Glottal pulse estimation. The parameters of the LF (Liljencrants/Fant) model are used for the characterisation of the glottal pulse [7].

(f) Pitch trajectory is estimated using autocorrelation feature analysis. This is followed by a pitch pattern analysis 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 voice conversion process (Fig 2) involves the following steps:

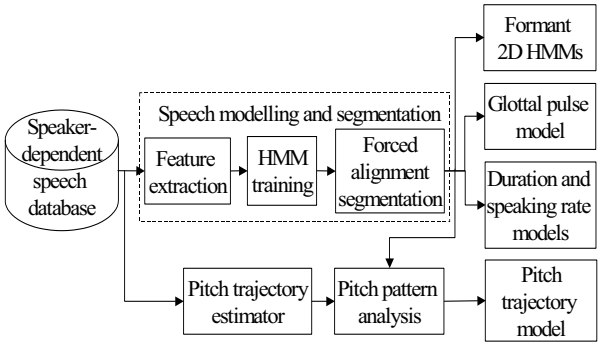


**Figure 1-** Voice modelling procedure.

a) Source-filter separation. A standard LP inverse filtering method is used for separation of vocal tract and excitation.

b) Vocal-tract spectrum transformation. The frequency spectrum of the LP model of the vocal tract of the source speaker is warped towards the target.

c) Glottal pulse mapping. The shape of the glottal pulse excitation of the source voice is warped towards the shape of the target glottal pulse.

d) Pitch modification based on TD-PSOLA. The speech excitation of the source speaker is fed to a TD-PSOLA algorithm for modification of the pitch, speaking rate and intonation features towards those of the target speaker.

e) Speech reconstruction. After transformation, the modified spectrum of the LP model of the vocal tract can be converted to LP filter coefficients. This is achieved using an inverse Fourier transform of the LP model spectrum followed the Levinson algorithm. The modified excitation is then filtered by the LP model.

## 3. ESTIMATION AND TRANSFORMAITON OF FORMANT MODELS

### 3.1 Formant Models

A 2-D HMM with $N$ left–to-right states across frequency, and $M$ states across time is used to model and label formant observations [2,3]. The formant features at time $t$ are the frequency $F_t$, the bandwidth $BW_t$ and the magnitude $I_t$ of the resonance and their temporal difference values. The LP pole coefficients are the raw data from which the formant features are estimated. Since there is not a 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 sorting 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 pdf.

*Estimation of formant trajectories*: The poles of each speech segment are ordered in increasing frequency and then classified and labelled with a formant classifier using the set of speaker-dependent and phoneme-dependent HMMs of formants and a Viterbi decoder[3,4]. Formant trajectory estimation for a speech waveform is achieved through minimisation of a weighted mean square error objective function:

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

where $I_k(t)$ is the total number of poles in the $t^{th}$ speech frame

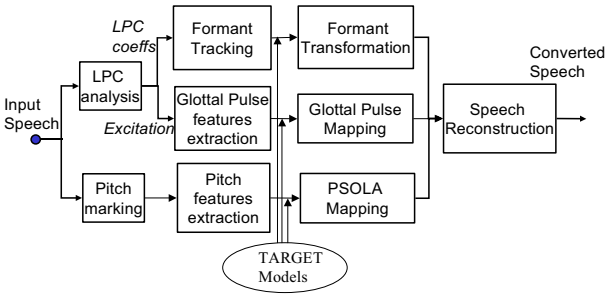


**Figure 2** - Voice Conversion procedure

classified as formant $k$ and $w_{ki}$ is a probabilistic weight derived from the model. Similarly, through the viterbi decoder at each formant, a trajectory of the associated bandwidths is obtained. The estimates of the magnitude of resonance at formants are obtained by sampling the magnitude frequency response of the LP model.

### 3.2 Formant Transformation

Formant transformation is performed through non-uniform adaptive sub-band spectral mapping. The formant frequency estimates are used to divide the signal spectrum into $N$ sub-bands centred on formant trajectories. 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 magnitudes and is used for derivation of the spectral mapping function [3]. The equation for voice conversion through spectrum mapping is expressed as

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

where $X$, $Y$, $t$, and $f$ denote the source spectrum, the transformed spectrum, and the time and frequency variables respectively. The frequency warping function includes the mapping functions for both the formant frequency $\alpha(f,t)$,and bandwidth $\beta(f,t)$. The magnitude frequency shaping function $\gamma(f,t)$ is used to map the spectral magnitude between the source and the target speakers.

The frequency warping function $\alpha(f,t)$, is derived from the ratio of the differences between the successive formant frequencies of the target speaker to that of the source speaker as

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

Similarly, the bandwidth and intensity warping functions are derived from the ratios of the bandwidth or intensity of the target and source speakers at each formant respectively.

## 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 broad characteristic patterns of variation over time. The pitch and intonation analysis model proposed here models continuous patterns of F0 curve. 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 and intonation characteristic features of a speaker. The F0 contour is extracted from autocorrelation-based pitchmarks. It is followed by a rise/fall/connection (RFC) analysis [5]. 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 3 shows an illustration of the pitch intonation model, where a schematic representation of F0, is used for the derivation of the features modelling the pitch and intonation patterns of a speaker. These features are described as follows:

*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, $F0_{range}$*: 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 phrase

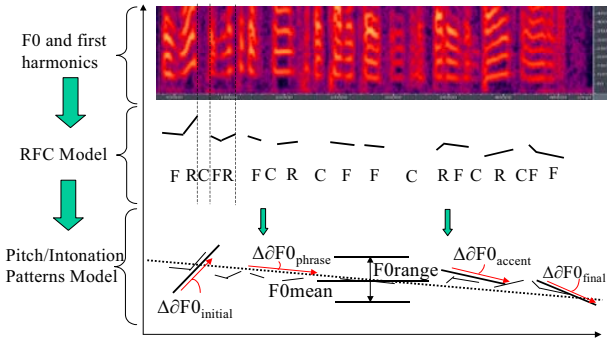


**Figure 3 -** The pitch and intonation model

slope ($\partial F0_{phrase}$), found from the slope of F0 curve across the entire length of an intonation phrase, (b) The initial pitch slope ($\partial F0_{initial}$), obtained form the slope of the first pitch segment of a phrase (c) the final pitch slope ($\partial F0_{final}$), 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 and a statement and between different regional and national accents.

*Pitch accent*: 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.

*Duration parameters*: The variables that affect the style of delivery of speech also include speaking rate, phoneme duration pattern and pause pattern.

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

## 5.EXPERIMENTS

A number of experiments were conducted on a set of two male and two female American English speakers taken from the WSJ database. 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, able to model up to 6 formants, considered to be the maximum number of significant formants in speech.

### 5.1 Feature Models

The Hidden Markov model Toolkit (HTK) is used for training HMMs and decoding speech [6]. 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 a Gaussian mixture distribution with 20 components. The average phoneme recognition rate with these speaker-dependent HMMs is about 94%. The Viterbi state-decoder is then used with the phonetic transcription supplied, to obtain the phonetically labelled segment boundaries. This combination results in a minimisation of errors in the estimation of phonetic segment boundaries.

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

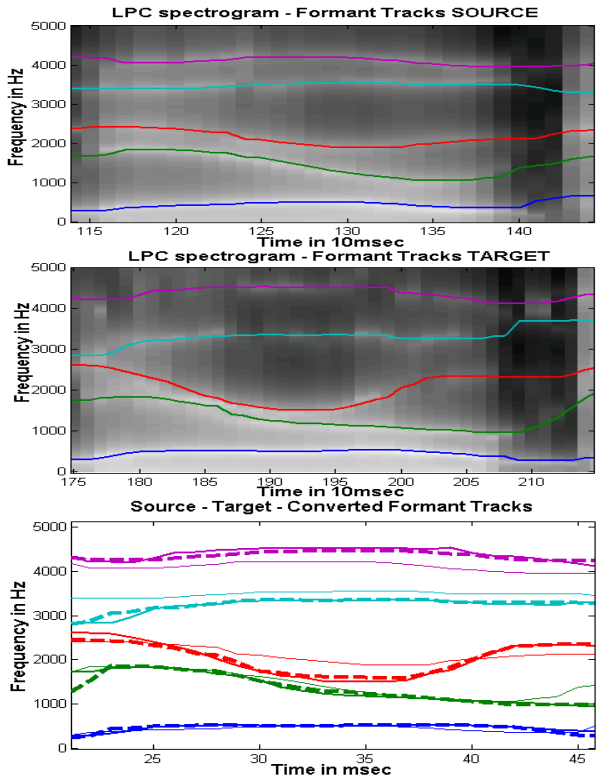


**Figure 4 -** Top and middle: source and target LP spectrogram and formant tracks, bottom: time-normalised source(thin line), target(solid line) and transformed(dashed) speech formant tracks

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 magnitude (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 component mixture Gaussian pdf. The overall result is a set of speaker-dependent phoneme-dependent formant 2-D HMMs with three states along time and five states along frequency. HMMs, in addition to being a good model of cepstrum features, are also good probability models of the distributions of the formants along the frequency axis [3].

### 5.2 Voice Conversion Illustrations

*Formant Mapping*: In Figure 4, the first illustration shows the superimposed formant tracks for the source, and the second for the target, both saying "two 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 the LP-spectrograms. In the third illustration the two formant tracks and the track of the converted sentence are time aligned and displayed. It can be seen that the converted track follows the target's. Figure 5 shows the mapping of the frequency spectra of a source speaker to a target speaker for two frames. Note that the transformed and target spectra are very close.

*Pitch intonation modification:* An illustration of the F0 contour before and after transformation is shown in Figure 6. All parameters, i.e. average pitch, pitch range, pitch slopes and duration are mapped.

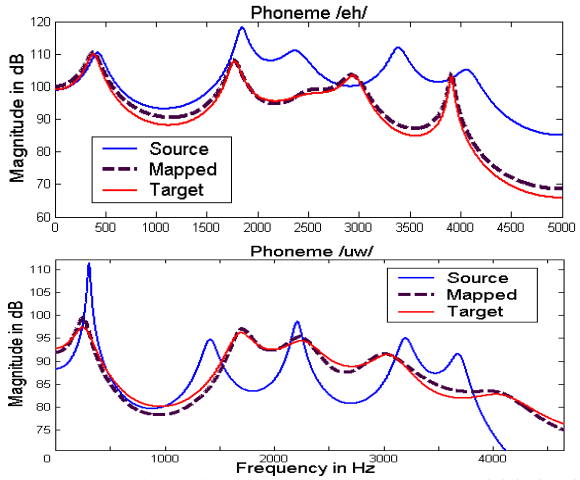


**Figure 5 -** Transformation of frequency spectra for /eh/, /uw/. Warping ratios for /eh/, (fig. 4, frame 122): α=[0.9 0.96 1.05 0.87 0.96], β=[1 1.25 1.33 1.37 0.25], γ=[1 0.4 0.25 0.4 0.9].

### 5.3 Perceptual Experiments

Two sets of voice conversion experiments, based on mean opinion score were performed (table I); one for a pair of male source and target speakers and one for a pair of female. Ten listener subjects took part in all experiments. In each experiment one speaker pair was used and four sentences were transformed. Each time the opposite transformation was also performed i.e. the source was used as a target and the target as a source resulting in eight sentences for each of the three conversions. Furthermore, the experiment was set up to convert each test sentence using three different levels of transformation resulting in a total of forty-eight test sentences. In the first level, all voice features were transformed. In the second level, only the formant features, including the frequency, bandwidth and magnitude of resonance at formants, were transformed. Finally, in the third level only the phoneme duration, speaking rate, the intonation features and the pitch mean and variance were transformed.

The listeners were asked to compare each converted sentence to a sentence spoken by the source and one spoken by the target. The listeners were asked to give a score from 1 to 10, with 1 signifying the least similarity and 10 signifying the highest. A number of unprocessed source and target sentences were also randomly included to provide the unconverted similarity score.

*a) Transformation of all features:* On average, the sentences resembled more like the target speaker's (with similarity scores of 6.25 for male and 5.53 for female) and less like the source speaker's (with similarity scores 2.75 and 3.4). These similarity scores demonstrate the success of voice conversion in particular when compared with the average similarity score of the unprocessed sentences to a sentence of the same speaker of 8.1 for males and 8.9 for females. The conversion from a female to a female speaker resulted in a lower performance due to the voice characteristics of the particular speakers rather than their gender.

| Features converted | Male | | Female | |
|---|---|---|---|---|
| | Source | Target | Source | Target |
| No conversion | 8.1 | 2.2 | 8.9 | 2.7 |
| All features | 2.75 | 6.25 | 3.4 | 5.53 |
| Formants | 5.19 | 4.56 | 5.66 | 3.85 |
| Pitch,sp.rate,intonation | 4.86 | 4.18 | 4.55 | 4.05 |

**Table I:** Voice Conversion Mean Opinion Scores

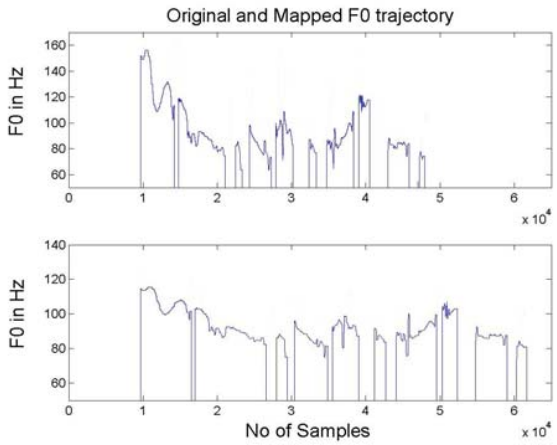


**Figure 6 -** Transformation of pitch parameters. The overall shape is preserved. Modified parameters: average pitch from 85Hz to 100 Hz, duration increase by 30%, pitch range decrease by 50%, phrase slope from –30Hz/sec to –12Hz/sec.

*b) Transformation of selected features:* Both the conversion of formant features or the delivery style and intonation features go someway into making the converted source speech similar to the target but are not successful on their own (the similarity scores are 4.56 and 4.18 for male and 4.05 and 3.85 for female respectively). It is clear that the pitch intonation features play an important role in the conversion and that their importance varies depending on the particular set of speakers being transformed.

## 6. CONCLUSION

This paper presented a model-based voice morphing method. The features used for characterisation of a speakers voice included spectral features, intonation and delivery style features. The statistical models used for modelling the trajectories of voice characteristic features were presented and a method was outlined for transformation of the voice of source speaker to a target speaker. Experimental evaluation shows that the method described can change the perception of a source speaker to another speaker. Further work is focussed on modelling and transformation of glottal pulse characteristics.

## REFERENCES

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

[2] 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.

[3] Rentzos D., Vaseghi S., Yan Q., Ho C.H., Turajlic E. "Probability Models of Formant Parameters for Voice Conversion", in Proc. Eurospeech 2003, pp. 2405-2408.

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

[5] Taylor, P (1995), "The rise/fall/connection model of intonation" Speech Communication 15, pp169-186

[6] Young, S. Woodland, P. (1996). HTK Hidden Markov Model Toolkit. Entropic.

[7] Fant,G. (1986), Glottal flow: models and interaction, Journal of Phonetics, 14, pp. 393-399.