ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

VOICE CONVERSION BETWEEN UK AND US ACCENTED ENGLISH

Ching-Hsiang Ho Saeed Vaseghi Aimin Chen

The Queen's University of Belfast, N. Ireland. Email (ch.ho, s.vaseghi, a.chen@ee.qub.ac.uk)

10.21437/Eurospeech.1999-461

ABSTRACT

This paper presents an HMM-based method and experimental results for voice conversion between UK and US accented English. Phonetic-tree based tied-state triphone HMMs are used to map equivalent states of the source and target spectra. Then a linear transformation method is incorporated to estimate the most likely target spectra for a given input. The mapping is between two different sets of phoneme i.e. the 44-phoneme UK English BEEP phone set and 39-phoneme US CMU phone set. Finally, a prosody adaptation is applied to tune the prosodic parameters. The experiments are based on voice conversion between speakers speaking different unrestricted texts. Acoustic-phonetic mapping between two different accents database enables us to attempt to deconstruct accents to investigate how they are distributed among different parameters such as spectra, energy contour, pitch, and duration.

1. INTRODUCTION

Voice conversion is the mapping of the acoustic space of one speaker, the source speaker, to the acoustic space of another, the target speaker [3,4,5]. In [3] Abe, Nakamura et al describe the use of a vector quantiser code book as a one to one mapping function between the spectral vectors of the source and the target speakers. This approach was extended in [4] to a probabilistic Gaussian mixture model (GMM). In this paper these ideas are further extended to include hidden Markov models (HMMs) of context-dependent triphones and unrestricted input sentences. The factors that affect the voice characteristics of a speaker are gender, age, prosodic parameters and accent. Gender and age effect the vocal tract size and characteristics and also the pitch frequency. The simplest method for speaker adaptation involves frequency warping in which, given a set of phonetic HMMs, for the input speech a phone-dependent ML warping parameter is estimated to map the frequency spectrum of the synthesiser's voice to that of the input voice. A more detailed transformation has a full matrix linear transform for each triphone. The linear transformation matrices are estimated using a maximum likelihood criterion. The transforms are arranged in a phonetic-tree cluster structure, where the number of transforms estimated at each level depends on the amount of training data from the target speaker.

Since the accent characteristics are not only affected by phonetic variation but also by the tonal details, a prosody modification is applied in pitch synchronised analysis and synthesis framework. By spectral and prosody adaptation a mimicking of a given voice characteristics can be performed. Therefore, the effect of accent can be studied with individual phonetic and prosodic parameters.

In this paper a voice conversion system is first presented as Fig 1. Then, the voice mapping is applied between an UK and an US English database. The evaluation results are accomplished in a stage by stage process. In each stage we present perceptual results of the effect of mapping each parameters on transferring the speaker characteristics and the accent of the target speaker to the source speaker.

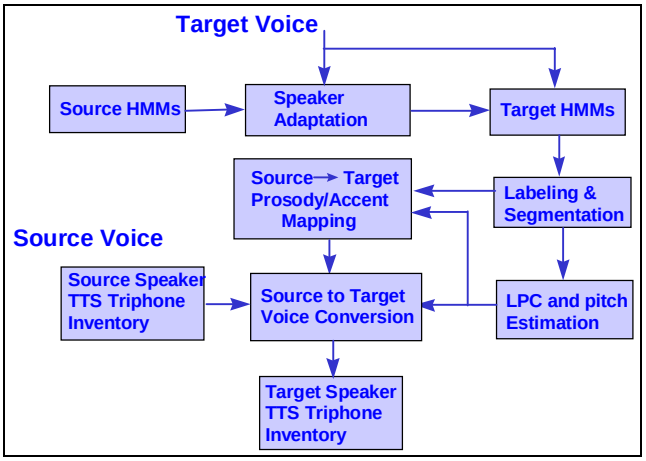

Figure 1: A voice conversion system

2. HMM-BASED VOICE MAPPING

Context-dependent triphones are the basic unit in our voice conversion system. The tied-state triphone HMMs and micro-prosody models are estimated through a decision tree clustering by a set of linguistic questions. During voice mapping, these models are used as the acoustic and prosodic space for each speaker.

2.1 US to UK Phone Mapping

For the translation of the US English database the 39-phone CMU dictionary without stress is used. BEEP dictionary which consists of 44 phones is employed to translate UK English database. By examination of the two phone sets, the following mapping are incorporated in this paper.

Rules of phoneme mapping		
BEEP	→	CMU
ax	→	ah
ea	→	eh
ia	→	ih
oh	→	aa
ua	→	uh

Figure 2: The rules of phoneme mapping

2.2 Phonetic Tree

In our study speech is modelled with context dependent triphone units [1]. The use of triphones, in addition to capturing the contextual correlation of successive speech units, alleviates the distortion effects of any timing errors in unit segmentation process. In general the quality of synthesised speech improves with increased contextual resolution. A decision-tree clustering method is employed to cluster the triphone HMMs in state tying, and to estimate the models, and the synthesis units for unseen triphones. The decision-tree for tied-stated triphone clustering is also applied to the phonetic mapping during voice conversion.

Since the decision tree for phonetic clustering should be consistent among different speakers, a general phonetic tree which is trained by a large database are used [2]. As each cluster of target and source speaker features are modelld by mixture Gaussian it is appropriate to use a least mean square error estimation method for spectral mapping.

2.2 Context-dependent Micro-Prosody Tree

In the system considered in this paper speech prosody is modelled using the concept of the decision tree *statistical micro-prosody* model. Micro-prosody is defined as prosodic relations between successive phonetic segments. Micro-prosody parameters are considered as signals whose states depend on the current and the neighbouring phones, for example the probability of pitch frequency can be modelled as

$$p\left(F_{0_n}, \lambda_n | (\lambda_{n-1}, F_{0_{n-1}}), (\lambda_{n+1}, F_{0_{n+1}}), stress\right) \quad (2)$$

where the prosody of a phone is affected by the neighbouring phones, their prosodic conditioning and the stress.

For modelling and training of prosodic parameters a hierarchical decision tree-based prosody clustering structure is used in which linguistic knowledge and statistical training methods are combined. At the lowest level for each leaf a set of parameters are estimated to maintain the correct 'micro-prosodic' relationship between the energy, the duration and the pitch of successive triphones in a sentence.

These statistics are then used to adapt the prosodic parameters from source speaker to target speaker and maintain the correct relation between prosody of successive triphone units in synthesised speech.

3. VOICE CONVERSION SYSTEM

Voice conversion is a two stage process. Firstly the spectral of the source utterance is converted to the spectral of the target speaker. Here, a set of tied-state triphone HMMs is used to adapt the source model to the target. The best triphone for concatenation is estimated by minimising the least square error or maximising the likelihood of the adapted model. Then, prosodic parameters are adapted by linear transformation for each clustered micro-prosody. A pitch synchronised synthesiser is used to evaluate the final results.

3.1 Features

Line spectral frequencies (LSFs) are used as spectral feature for voice conversion for two main reasons. Firstly, LSFs have been shown to possess very good linear interpolation characteristics. A smoothing between the concatenation units can be easily achieved. Secondly, LSFs are well related to the formant frequencies and bandwidths, also a weighted distance measure can be used to emphasis spectral frequencies around formant location.

3.1 Voice Spectral Mapping Function

Using a Wiener-type least squared error optimisation method, the mapping function between the source spectrum $X(w)$ and the target spectrum $Y(w)$ for the i^{th} speech class is of the form

$$Y(\omega) = \frac{\left|\hat{Y}(\omega)\right|}{\left|X(\omega)\right|} X(\omega) \tag{3}$$

where $\hat{Y}$ stands for the estimated spectrum given X. In [4] a Gaussian mixture model is described for mapping the source spectrum to the target spectrum. Extending the mapping function here to context-dependent phonetic HMMs, with M-mixture Gaussians per state model, the mapping between the corresponding states of phonetic HMMs yields

$$E[Y|X] = \sum_{k=1}^{N_M} \sum_{i=1}^{N_S} \sum_{m=1}^{M} P(c_{kim}, s_{ki}, \lambda_k | X)[\upsilon_{kim} + \Gamma_{kim}\Sigma_{kim}^{-1}(X_{kim} - \mu_{kim})] \tag{4}$$

where μ, υ, are the mean of X and Y, Σ is the covariance matrix of X and Γ the cross-covariance of X and Y. For each phone segment of the source speech, the spectral mapping from target candidate is estimated by dynamic time warping with the cost function, Eq 5, which is the weighted mean square error. In practice instead of using a mean target spectra for coversion we might select the single example that is closest to the mean using a weighted opitisation criteria such as

$$\varepsilon = \sum_{\text{all } i} \left\| w_i \left(Y_i - E\left(Y | X_i\right) \right) \right\|^2 \tag{5}$$

where higher weights are assigned to the closely spaced LSFs which are corresponding to formant location.

An alternative method of spectral conversion is to use a linear speaker transform as in speaker adaptive speech recognition as

$$\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{x}_k \tag{6}$$

where the linear transformation is a full matrix. The solution for A can be obtained using probabilistic optimisation. Eq 6 can be extended to a tree of matrix transforms, where the number, and contextual resolution of transforms increase as more data becomes available.

To evaluate the spectral transformation, the vocal tract filter is applied to the magnitude spectrum of the original source signal. Then, inverse DFT is used to produce the synthesis target voice [5].

3.2 Micro-Prosody Modification

For a given sequence of concatenated speech. For example consider the modification of, $F_{0a|b,c}$; the fundamental pitch frequency of a triphone 'a', within the contexts of phones b and c. The modification factor required to maintain the statistical relationships of the pitch of the phone b within the contexts of a and c is obtained from a triphone tree cluster model. This modification factor is then used to change the pitch of concatenated speech sequence.

For prosody adaptation, the mean and variance of the distribution of the micro-prosody parameters of the source speaker s is adapted to that of the target speaker t using the following relation [5]

$$F_{0\,a|b,c}^{\,t} = \alpha\, F_{0\,a|b,c}^{\,s} + \beta \tag{7}$$

where the notation $F_{0a|b,c}$ denotes the pitch of the triphone a within the context of neighbouring phones b and c, and the adaptation coefficients a and b are given by

$$\alpha = \sqrt{\frac{\sigma_t^2}{\sigma_s^2}}, \quad \beta = \mu_t - \alpha\mu_s \tag{8}$$

where m and s^2 denote the context-dependent mean and variance of prosodic parameters. This relation is used for mapping of pitch, energy and duration parameters.

4. EVALUATIONS

For evaluation purposes two databases of with unrestricted text were used. The UK English (UKE) database contains about 4.8 hours reading by a male speaker in which there are around 8000 words and 18000 triphones. The US English (USE) database is taken from one of the WSJ0 speaker-dependent male databases. This database is made up of about 3.6 hours reading. There are around 7000 words and 13000 different triphones.

For each database, around 8700 tied-state triphone HMMs with 3 left-to-right states are trained. Each HMM state contains 12/15 mixture Gaussians. During training, the features used in this experiment are 12 MFCCs with energy and their differentiation and acceleration. Given these HMMs, the automatic segmentation is completed through forced alignment. Given the statistical prosody models and HMMs, the TTS triphone inventory is selected from the segmented speech. During voice conversion, LSF features which are derived from 40 order LPC coefficients are applied.

The performance of voice conversion is evaluated in two stages. The first stage is the spectral mapping

without prosody modification. The most likely set of features for conversion are selected from the target inventory. Then, the selected spectra and the unmodified source residual signal are used to synthesis new speech for evaluation. The second stage is the spectral mapping followed by the prosody mapping. Here, the TD-PSOLA method is employed. . During transformation the short-term (ST) signals with around 2 local pitch periods are mapped under the pitch-synchronisation. Prosody parameters of the source signal are first kept the same. Then, each parameter is modified individually. These parameters are duration, energy contour, and pitch period. Based on the statistical models, the prosody mapping is estimated from source and target speech database. In each stage a comparison of the spectral mapping method is also made with the direct triphone mapping. In our experiment the American English spoken by the source speaker was first transformed to UK English spoken by the target speaker.

4.1 Objective Evaluation

For the objective test, the likelihood of each sentence given the sentence target HMM sequence is computed by the recogniser. In Table 1, ORG denotes the likelihood for unmodified source speech. S1 and S2 denote the source speech after spectral transformation only. P1 and P2 denote TD-PSOLA transformation applied. In S1 and P1 the optimal target candidate is estimated by decision tree clustering followed by a spectral mapping. In S2 and P2, except for the unseen triphone, only the direct phone mapping is applied.

	ORG	S1	S2	P1	P2
c0201	-404	-359	-358	-373	-368

Table 1: Evaluation results for utterance c0201.

4.2 Subjective Evaluation - Spectral Mapping

In our investigation the decision tree clustering plus spectral mapping method produced more natural synthesised speech than direct triphone mapping, although the likelihood score is slightly lower. According to the accent evaluation, it is agreed that the modified utterance contains a limited amount of characteristics of the target speaker. That is to say the speech sounds like neither source speaker nor target speaker. On the other hand, the UK accent can be clearly heard in the synthsised speech. Most of the listeners in our lab confirm that it sounds like an American speaker with UK accent.

4.3 Subjective Evaluation - Prosody Mapping

Compared to the spectral mapping, TD-PSOLA transforms the ST windowed signal. From the listening evaluation it is clear that the residual signal does contain a lot of speaker characteristics. However, the sound is not very pleasant. It is likely that the prosodic parameters should be highly correlated to the residual excitation signal in performing natural sound. Furthermore, the prosody mapping is performed simultaneously. It seems more natural when pitch period modifications are applied. To evaluate the exact accent effects from prosody, a more detailed phonetic and context dependent study is needed.

5. CONCLUSION

In this paper a HMM based voice conversion framework is presented. A first study of the accent effects has been performed. Several aspects are going to be investigated in the next stage. A more detailed phonetic rule can be used to improve the decision tree to study speaker accent. A different speaker adaptation method can be applied to the spectral mapping. A better prosody model is needed for the prosody adaptation. For each speaker parameters such as spectral features, excitation signal, pitch, duration and energy, an extensive and systematic study will be accomplished.

REFERENCES

[1] R.E. Donovan. (1996) Trainable Speech Synthesis, PhD Thesis, Cambridge University Engineering Department.

[2] H. Hon, A. Acero, X. Huang, J. Liu, M.Plumpe. (1998) Automatic Generation of Synthesis Units for Trainable Text-to-Speech systems, ICASSP.-98, page 293-296.

[3] M. Abe, S. Nakamura, K. Shikano,H. Kuwabara (1988), Voice Conversion Through Vector Quantisation, proc. ICASSP-88, pages 655-658.

[4] Y. Stylianou, O Cappe (1998), Voice Conversion Based on Probabilistic Classification and Harmonic Noise Model, proc. ICASSP-98, pages 28-284.

[5] L. M. Arslan, D. Talkin (1998), Speaker Transformation Using Sentence HMM Based Alignment and Detailed Prosody Modification, IEEE Proc. ICASSP98.