USING PHONE AND DIPHONE BASED ACOUSTIC MODELS FOR VOICE CONVERSION: A STEP TOWARDS CREATING VOICE FONTS

Arun Kumar

Centre for Applied Research in Electronics
Indian Institute of Technology
New Delhi - 110016, India
email:arunkm@care.iitd.ernet.in

Ashish Verma

IBM India Research Lab
Indian Institute of Technology
New Delhi - 110016, India
email:vashish@in.ibm.com

ABSTRACT

Voice conversion techniques attempt to modify speech signal so that it is perceived as if spoken by another speaker, different from the original speaker. In this paper, we present a novel approach to perform voice conversion. Our approach uses acoustic models based on units of speech, like phones and diphones, for voice conversion. These models can be computed and used independently for a given speaker without being concerned about the source or target speaker. It avoids the use of a parallel speech corpus in the voices of source and target speakers. It is shown that by using the proposed approach, voice fonts can be created and stored which will represent individual characteristics of a particular speaker, to be used for customization of synthetic speech. We also show through objective and subjective tests, that voice conversion quality is comparable to other approaches that require a parallel speech corpus.

1. INTRODUCTION

Voice conversion is concerned with the transformation of speech signal in the voice of a source speaker to that of a target speaker. A suitable performance measure must determine how close the converted speech is perceived to the speech of the target speaker. To address the problem in its entirety, it is necessary to consider various dimensions of speech including pitch, energy modulation, stress, rate of speaking, spectral envelope and style of speaking. Some of these dimensions are rather qualitative measures of speech. Morever, some of them are more important than others in the context of voice conversion. Therefore, classically the problem has been attempted in two parts. The first part relates to spectral envelope conversion while the second part relates to prosodic modification of the speech signal.

In this paper, we will mainly focus upon spectral envelope conversion and the implications of the proposed approach towards creation of “voice fonts”. Current voice conversion methods are limited by the requirement of a “parallel” speech corpus in the voices of source and target speakers. Here, parallel corpus means that the same set of words or sentences are uttered by both speakers. This is required to learn a mapping between the source and target spectral envelopes. Due to this requirement, the mapping function is always coupled with a specific source-target speaker pair, and has to be retrained for every new pair. We propose a novel approach for spectral envelope conversion using phone and diphone based acoustic models as a step towards overcoming this limitation. Some of the applications of ”voice fonts” will be in personalized text-to-speech synthesis and personalized multimedia mail, very low bit rate speech coding and speech-to-speech translation.

Prosodic modifications in the form of time and pitch scaling have been attempted through various sinusoidal and harmonic models [1, 2, 3]. Some of the previous attempts for spectral envelope conversion can be found in [4, 5, 6, 7, 8]. In one of the earliest attempts, Abe *et. al* in [4], used a mapping codebook method in which the source speaker’s code vector is replaced frame-by-frame with corresponding code vector of the target speaker. Vector Quantization (VQ) is applied to partition the acoustic space of source and target speakers to get these code vectors. Later in [5], a segment-based approach is used in which speech segments are used as voice conversion units to capture dynamic characteristics of the speaker individuality. The incoming speech signal from source speaker is segmented into speech units by using a speech recognizer and then these segments are replaced with corresponding speech segment of the target speaker using a table-lookup approach. In [6] Valbret *et. al.* use Linear Multivariate Regression (LMR) and Dynamic Frequency Warping (DFW), to convert the spectral envelope. They have used pitch-synchronous overlap add analysis (PSOLA) for prosodic modifications. In [7], Yannis *et. al.* have used a probabilistic conversion function to convert the spectral envelope of voiced frames of speech to that of the target speaker. They have used Harmonic + Noise Model (HNM) for time and pitch scale modifications of speech [3]. They first partition the acoustic space of the source speaker with Gaussian Mixture Models (GMM) and then learn the parameters of the conversion function using a parallel speech corpus in source and target speaker’s voices.

Rest of the paper is organized as follows. We describe our approach in Section 3. However, Section 2 first describes how this approach can be used to create voice fonts. Various experiments are described in Section 4. We discuss the results corresponding to these experiments in Section 5. Finally we conclude in Section 6.

2. VOICE FONTS

The Webster’s dictionary defines “font” as “an assortment or set of type all of one size and style”. In the same token, we define “voice font” as “a set of descriptors of a person’s voice”. Just as text fonts make the text appear in a particular style, voice fonts will make the speech sound in a particular individual’s voice. Some of the most important descriptors of voice fonts will comprise spectral envelope for acoustic units and prosodic features such as fundamental frequency and rate of speaking [9].

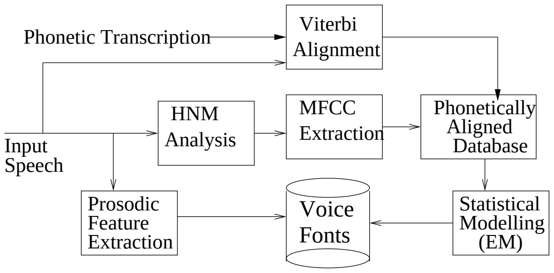

Fig. 1. Creation of Voice Fonts

The motivation for voice fonts is to create a compact representation of the speaker individuality, which can be used to create personalized voice from speech in another voice. This will require easy extraction and storage of speaker specific descriptors from the speech signal. Voice fonts will also allow the use of the underlying descriptors as independent input into a text-to-speech synthesis system to synthesize speech in the voice of the particular speaker. The standardization of the descriptors constituting the voice font of a person will be an important goal to realize the potential applications of voice fonts.

Voice fonts will allow an independent representation of the spectral envelope which can be used for voice conversion without having to learn a mapping in advance between the source and target envelopes. By modifying the underlying descriptors of voice fonts to match with those of a target speaker, speech signal can be synthesized in the target speaker's voice. Our approach attempts to achieve this goal by using speech units based acoustic models for voice conversion.

3. PROPOSED APPROACH

As mentioned, most of the approaches proposed in the literature, require a parallel speech corpus in the voices of source and target speakers to learn a mapping between their respective spectral envelopes. This parallel speech corpus is created by asking the source and target speakers to speak the same words or sentences. Due to this requirement, the mapping function is always associated with a specific source-target pair and has to be retrained for every new pair.

We propose to partition and model a speaker's acoustic space with explicit speech units, like phones or diphones, instead of modeling it with vector quantization or Gaussian Mixture Models. The spectral envelope of each such speech unit can then be converted into the corresponding spectral envelope of the target speaker for the same speech unit. This method of spectral envelope conversion does not require a parallel speech corpus to learn a mapping between source and target speakers. Instead, the spectral envelope of the speech units can be modeled from independent speech corpus for a particular speaker.

3.1. Extraction of individuality features

The creation of voice fonts, by extracting the underlying descriptors, is schematically shown in Fig. 1. We in particular focus upon three of these descriptors in this work, *viz.*, pitch frequency, spectral envelope of various acoustic units and average speaking rate. We record continuous speech utterances from a speaker whose individuality features are to be computed. Pitch frequency is computed for every voiced frame using a standard autocorrelation based pitch detector, described in ITU-T standard G.729. Average of the pitch frequency so computed constitutes one descriptor of the voice font for the speaker. For spectral envelope representation, we use 16 dimensional Mel-Frequency Cepstral Coefficients (MFCC) extracted from a 16 kHz speech signal by discrete regularized cepstrum method using a warped frequency scale [10]. The speech utterances, spoken by the speaker, are phonetically aligned by using Hidden Markov Models (HMM) of a speech recognizer through Viterbi alignment. All the MFCC vectors corresponding to a particular phone are then grouped together using these alignments and modeled through Gaussian Mixture Models (GMM), in general. GMM models the distribution of an observed vector $\mathbf{x}$, a p-dimensional MFCC vector in this context, as a weighted sum of several Gaussian distributions. This can be represented as follows:

$$p(\mathbf{x}) = \sum_{j=1}^{M} \alpha_j * N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{1}$$

where $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ defined by

$$N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2}$$

In (1), the terms α_j are normalized positive weights such that $\sum_{j=1}^{M} \alpha_j = 1$ and $\alpha_j > 0$. The parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and α_i are computed using the classic Expectation Maximization (EM) algorithm. The conditional probability, $P(P_i|\mathbf{x})$, that a vector $\mathbf{x}$ belongs to a phone P_i can be computed as follows:

$$P(P_i|\mathbf{x}) = \frac{\beta_i \sum_{j=1}^{M} \alpha_{ij} N(\mathbf{x}, \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})}{\sum_{i=1}^{N} \beta_i \sum_{j=1}^{M} \alpha_{ij} N(\mathbf{x}, \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})} \tag{3}$$

where β_i represents probability of occurrence of phone P_i and N denotes the total number of phones. β_i are computed from the aligned speech database of the speaker.

Note that GMM is not being used to partition the acoustic space of the speaker as in [7]. In our method, acoustic space of a speaker is partitioned explicitly into phones using the alignments and GMM is used for finer modeling of each phone. When M=1, GMM reduces to computing the mean and covariance matrix of the MFCC vectors for the given phone. Note that although the above procedure has been described for phone based speech units, the same procedure will be used for diphones, or allophones in general. We have used phones and diphones in our experiments, described in Section 4. GMM parameters (α, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) and the probabilities of occurrence β_i for all the phones, combined together will represent the second descriptor of the voice font for the speaker.

The third descriptor of the voice font, *viz.*, average rate of speaking, is not incorporated in the current model. We set it manually for the time scaling of the speech signal in order to match with that of the target speaker.

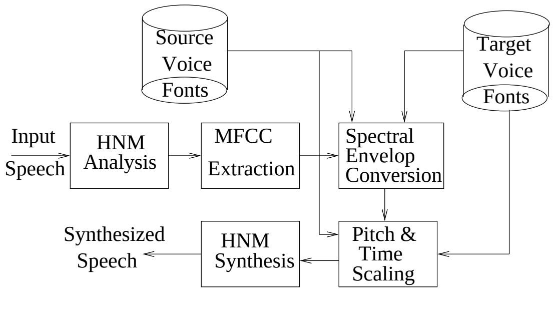

Fig. 2. Voice conversion procedure

3.2. Voice conversion

We perform voice conversion from one speaker to another speaker by converting three descriptors of the voice fonts as described earlier in this section.

We have implemented an integrated signal processing module which converts these three descriptors of the incoming speech signal. This module has been depicted in Fig. 2. It uses HNM to analyze the speech signal. HNM assumes the speech spectrum to be divided into two bands separated by maximum voiced frequency. The lower band is modeled as a sum of harmonically related sinusoids with slowly varying frequency and amplitude, and the upper band is modeled by modulated noise. This can be represented as

$$\hat{s}(t) = \sum_{k=-L(t)}^{k=L(t)} A_k(t) \exp(jktw_o(t)) + e(t) \tag{4}$$

where $w_o(t)$ is the fundamental frequency, $A_k(t)$ are the harmonic amplitudes and $L(t)$ is number of harmonics in the voiced band of the speech signal at time t. $e(t)$ models the noise part of the signal.

For every analysis frame, harmonic amplitudes A_k are computed using $w_o(t)$ by least square optimization between the actual speech signal and the harmonic part of $\hat{s}(t)$. The noise part is modeled by an all-pole filter of order 15, whose coefficients are extracted from a 40 ms speech signal centered around the analysis frame. More details about this procedure can be found in [3]. Spectral envelope corresponding to the voiced part of speech signal is obtained in the form of discrete MFCC vector by using the approach presented in [10].

To convert the spectral envelope of the input speech signal into that of the target speaker, we use following conversion function

$$\mathbf{x}_t = \sum_{i=1}^{K} P(P_i|\mathbf{x}_s)\boldsymbol{\phi}_i \tag{5}$$

$$\boldsymbol{\phi}_i = \sum_{j=1}^{M} \alpha_{ij} \left[\boldsymbol{\mu}_{tij} + \boldsymbol{\Sigma}_{tij}\boldsymbol{\Sigma}_{sij}^{-1}(\mathbf{x}_s - \boldsymbol{\mu}_{sij})\right] \tag{5a}$$

In (5), $\mathbf{x}_s$ and $\mathbf{x}_t$ represent the source and target MFCC vectors, $\boldsymbol{\mu}_{sij}$, $\boldsymbol{\Sigma}_{sij}$, $\boldsymbol{\mu}_{tij}$ and $\boldsymbol{\Sigma}_{tij}$ represent the mean vector and covariance matrix corresponding to GMM component j of phone (diphone) i for the source and target speakers respectively. The probability, $P(P_i|\mathbf{x}_s)$, is given in (3). Note that M is the number of GMM components to model a phone and N is total number of phones. In general, $K \leq N$, so that only first K most likely phones are used, corresponding to $\mathbf{x}_s$.

It can be noticed that this conversion function is motivated by the conversion function presented in [7], given as follows

$$\mathbf{x}_t = \sum_{i=1}^{M} P(C_i|\mathbf{x}_s)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_i)\right] \tag{6}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are mean and covariance matrix of GMM component C_i. Parameters $\boldsymbol{\nu}_i$ and $\boldsymbol{\Gamma}_i$ are learned through least square optimization performed over mapped source and target MFCC vectors, learned from a parallel speech corpus.

The function (5) used in the proposed approach differs from (6) in following aspects. Since it is based upon explicit phone acoustic models, it does not require a parallel speech corpus to learn its parameters, $(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. This makes it suitable to be used for voice fonts. Furthermore, to compute its parameters, (6) involves the inversion of very large matrices which makes it computationally very expensive. This problem does not arise in (5).

After converting the spectral envelope, pitch and time scaling are performed using the HNM framework. A constant pitch scaling factor is obtained by dividing average pitch frequency of the target speaker with that of the source speaker. This information can easily be extracted from the voice fonts of the source and target speakers as pitch frequency is a descriptor of voice fonts. A constant time-scaling factor is used to match the average speaking rates of source and target speakers.

4. EXPERIMENTS

We performed voice conversion experiments on continuous sentences to measure the robustness of the proposed approach and compare its performance with other approaches.

About 30 minutes of speech was recorded from two male speakers in the form of continuous Hindi sentences. About half of this data was recorded in the form of a parallel speech corpus where both the speakers spoke same sentences, while in the other half, they were given different sentences to speak. All the sentences were Viterbi aligned with an HMM based Hindi speech recognizer which, with an N-gram language model, produces about 88% word recognition rate on a general dictation task [11]. The parallel speech corpus, comprising of about 22,000 aligned and mapped MFCC vectors and 61 component Gaussian Mixture Models, was used to train the function, given in (6), for the speakers. The second half of the database was used to create voice fonts comprising of phone and diphone based GMM models and average pitch frequency independently for both the speakers. For phone based models we used 61 phones (number of phones in the speech recognition system), and for diphone based models, we used 335 diphones extracted from the database by considering only those diphones which had more than 100 vectors aligned to each of them. For simplicity only one component per GMM was used in both the models. To reduce the computations further, in all the experiments individual cepstral coefficients were assumed to be uncorrelated, resulting in diagonal covariance matrices.

For a quantitative comparison, we computed spectral envelopes of the source, target and converted speech sentences over aligned frames. The envelope is based on an LPC all-pole filter, whose coefficients were computed from 40ms of voiced speech. This was performed for various phonetic segments. Spectral Distortion for these envelopes was computed using the following expression:

$$SD = \left[\frac{1}{K}\sum_{k=0}^{K-1}\left[20\log|\mathbf{S}(k)| - 20\log|\hat{\mathbf{S}}(k)|\right]^2\right]^{1/2} \tag{7}$$

where $S(k)$ and $\hat{S}(k)$ represent the spectral envelope of the target and converted speakers respectively.

We performed listening tests in order to measure relative perceived quality of the synthesized speech sentences. In the first test, called Degradation Category Rating (DCR), we asked the subjects to grade the synthesized speech on a scale of 1 to 5, representing decreasing level of degradation from the target speech. In second test, we call opinion test, we asked the subjects to grade the synthesized speech on a scale of 1 to 10, considering the closeness of the speaker individuality of the synthesized speech with that of the target speaker. These listening tests involved twelve synthesized sentences and six subjects.

Results corresponding to the above experiments are discussed in Section 5. Some of the synthesized sentences can be accessed at *http://in.geocities.com/ashish_verm/conversion.html*.

5. RESULTS AND DISCUSSION

The spectral envelopes for four aligned voiced frames are shown in Fig. 3. The average spectral distortion (SD), corresponding to the frames are listed in Table 1. It can be noticed from the results that the proposed approach, using phone and diphone based acoustic models, performs similar to that using function (6), (shown with '+' symbol). The spectral distortion is reduced by 2 to 4 dB, which is about 33% to 68% relative improvement, from the initial distortion between source and target speaker. In some cases, our approach performs better than (6).

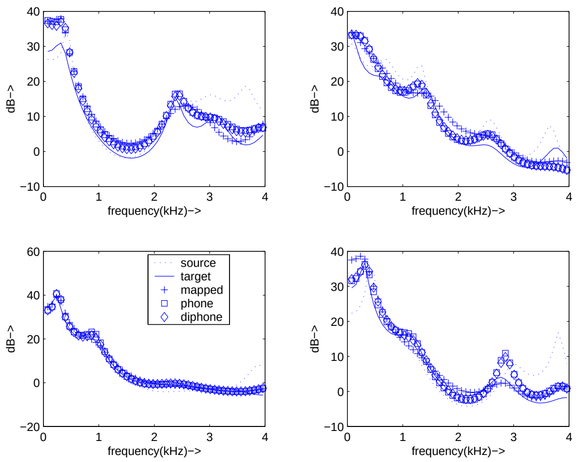

Fig. 3. Spectral evnelope plots

The results corresponding to the subjective tests are shown in last two rows of Table 1. As can be noticed from these results, the perceived quality of the synthesized speech using both the approaches are quite similar.

	Initial	Trained	Phone	Diphone
Spec. Dist.(dB)	9.17	5.53	5.51	5.13
Spec. Dist.(dB)	6.30	4.22	3.48	3.50
Spec. Dist.(dB)	5.84	2.11	1.84	1.85
Spec. Dist.(dB)	7.47	3.09	3.08	3.03
DCR score	-	3.25	3.25	3.20
Opinion Test	-	6.41	6.11	6.31

Table 1. Objective & subjective test resuls; Initial: between source and target, Trained: conversion function (6), Phone: phone models, Diphone: diphone models

6. CONCLUSION

We have proposed creation of voice fonts which can be used to personalize speech for an individual for various applications. We have shown the viability of this concept by proposing an approach for spectral conversion which makes use of acoustic unit based models and hence avoids the requirement of a parallel speech corpus. We have shown through various experiments that the quality of the converted speech is as good as using a mapping based approach.

7. REFERENCES

[1] Quatieri T. F. and McAulay R. J., "Speech Transformations based on sinusoidal representation," *IEEE Transaction on Acoustics, Speech and Signal Processing*, pp. 1449-1464, Vol. ASSP-34, No. 6, December 1986.

[2] Moulines E. and Charpentier F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, pp. 453-467, Vol. 9, No. 5, December 1990.

[3] Stylianou Y., Laroche J. and Moulines E., "High-quality speech modification based on a harmonic + noise model," in *Proceedings of. EUROSPEECH*, Madrid, Spain, 1995.

[4] Abe M., Nakamura S., Shikano K. and Kuwabara H, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conference Acoustics, Speech, Signal Processing*, pp. 655-658, April 1998.

[5] Abe M., "A segment-based approach to voice conversion," in *Proc. IEEE Int. Conference Acoustics, Speech, Signal Processing* , pp. 765-768, 1991.

[6] Valbret H., Moulines E. and Tubach J. P., "Voice transformation using PSOLA technique," *Speech Communication*, pp. 175-187, Vol. 11, June 1992.

[7] Stylianou Y., Cappe O., and Moulines E., "Continuous probabilistic transform for voice conversion," *IEEE Transaction on Speech and Audio Processing*, pp. 131-142, Vol. 6, No. 2, March 1998.

[8] Baudoin G. and Stylianou Y., "On the transformation of the speech spectrum for voice conversion", in *Proc. IEEE Int. Conference on Spoken Language and Processing*, pp. 1405-1408, Philadelphia 1996.

[9] Furui S, "Research on Individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication*, pp. 183-197, Vol. 5, 1986.

[10] Cappe O., Laroche J. and Moulines E., "Regularized estimation of cepstrum envelop from discrete frequency points," in *Proc. IEEE ASSP Workshop on Application of Signal Processing to Audio and Acoustics*, NY, October 1995.

[11] Neti C., Rajput N. and Verma A., "A large vocabulary continuous speech recognition system for Hindi," *National Conference on Communication*, Mumbai, January 1992.