A System for Voice Conversion Based on Adaptive Filtering and Line Spectral Frequency Distance Optimization for Text-to-Speech Synthesis

Özgül Salor, Mübeccel Demirekler

Department of Electrical and Electronics Engineering
Middle East Technical University,
Ankara, 06531, Turkey
{salor, demirek}@metu.edu.tr
Web: http://www.metu.edu.tr

Bryan Pellom

The Center for Spoken Language Research
University of Colorado at Boulder,
Boulder, 80303, USA
pellom@cslr.colorado.edu
Web: http://cslr.colorado.edu

ABSTRACT

This paper proposes a new voice conversion algorithm that modifies the source speaker's speech to sound as if produced by a target speaker. To date, most approaches for speaker transformation are based on mapping functions or codebooks. We propose a linear filtering based approach to the problem of mapping the spectral parameters of one speaker to those of the other. In the proposed method, the transformation is performed by filtering the source speaker's Line Spectral Pair (LSP) frequencies to obtain the LSP frequency estimates of the target speaker. Speech signal is time-aligned into a sequence of HMM states. The filters are designed for each HMM state using the aligned data. We consider two methods for spectral conversion. A linear transformation for the LSP's was obtained using the adaptive steepest gradient descent approach. Mean values of LSP's are adjusted to match those of the target speaker. In order to prevent the LSP vectors from resulting in unstable vocal tract filters, weighted least square estimation is used. This approach optimizes differences between source and target LSP's. Weights are inverses of the source LSP variances. This approach is integrated into a Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) analysis-synthesis framework. The algorithm is objectively evaluated using a distance measure based on the log-likelihood ratio of observing the input speech, given Gaussian mixture speaker models for both the source and the target voice. Results using the Gaussian mixture model formulated criteria demonstrate consistent transformation using a 5 speaker database. The algorithm offers promise for rapidly adapting text-to-speech systems to new voices.

1. Introduction

Voice conversion is a technique that modifies a source speaker's speech to be perceived as if a target speaker has spoken it. It is a subject of considerable importance, whose applications include text-to-speech synthesis based on acoustic unit concatenation, interpreted telephony, low bit rate speech coding and imposter modeling for voice verification systems [1]. Some of the current approaches for speaker transformation are based on developing mapping functions [2], [3]. Other approaches use codebooks of the source and the target speakers for mapping [4]. Neural networks have also been used for mapping the spectral characteristics of the speakers [5]. In this paper, we consider a new approach to the problem of voice transformation. In this research, Line Spectrum Pair (LSP) frequencies have been used to model the acoustic space of the speakers. The usage of LSP parameters for voice transformation is motivated by several reasons. LSP's correlate well with formant locations and bandwidth structure. Previous studies such as [6] show that voice-personality is significantly sensitive to the formant locations. Also, LSP's exhibit a fair degree of correlation in time and posses interpolation properties that are well suited for speech synthesis applications [7].

In this work, speech signal is time-aligned into a sequence of HMM states. Filters are designed for each HMM state using the aligned data. These filters map the source LSP's to the target LSP's. Two methods for mapping LSP frequencies of the source to those of the target have been considered. One of them is based on obtaining a FIR filter adaptively for transforming each LSP separately. The second method obtains a transformation matrix for mapping LSP's of the source to those of the target. After both methods, the mean values of the LSP's are adjusted to match those of the target means. This adjusts the overall offsets of the LSP's while the linear transformations adjust the formant movements along time axis. The possiblity of coming up with an unstable vocal tract filter is avoided by an optimization performed on LSP differences between the source LSP's and the transformed LSP's. After the spectral transformation, pitch-scale transformation is made inside a TD-PSOLA synthesis framework. Both the average pitch and the pitch range of the target speaker is determined during training and applied to the transformed speech. A Gaussian Mixture Model (GMM) based evaluation metric is applied to test the algorithm.

The organization of the paper is as follows. Section 2 discusses the transformation algorithm in detail.Implementation details are given in Section 3 and Section 4 discusses the evaluations.

2. Spectral Conversion Training Algorithm

2.1. Overview

We focus on two different approaches based on adaptive linear filtering techniques for the training part of the transformation. A simple diagram of a supervised adaptive filter is shown in Figure 1. The impulse response of the filter is adapted in time so as to minimize the error between the output of the filter and the de-

This work was supported in part by TÜBITAK, Scientific and Technical Research Council of Turkey, through a Graduate Research Fellowship.

sired response. The idea in both of our approaches is that, an FIR filter can be found to map the LSP's of each HMM-state of each phoneme of the source speaker to those of the target speaker. LSP's extracted at every frame are considered as temporal sequences, so n in Figure 1 corresponds to the frame number. The theory behind adaptive filtering requires that the input and the desired response are zero mean signals in time n. Therefore, mean values of both are subtracted before starting the training process. Once these filters are determined during the adaptation (training) process, they are used as transformation filters to obtain the estimates of the target speaker's LSP's. The advantage of this approach is that only 35 sentences (approximately 2 minutes of speech data) from the target speaker are enough for the all transformation filters to reasonably converge for each phoneme using the Least Mean Squares (LMS) adaptation approach. Moreover, the algorithm is computationally simple and fast for both training and synthesis phases.

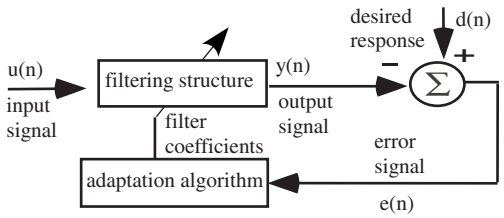

Fig. 1. Adaptive Filtering. In our application, the primary signal u(n) is the LSP's of the source speaker, reference signal d(n) is the LSP's of the target speaker

2.2. Training Data

For each analysis frame, an observation vector of 16 LSP frequencies is calculated. Dynamic Time Warping (DTW) based on minimizing the Euclidian distance between the LSP's of the source and the target is applied to time align the source and the target utterances. 35 phonetically balances sentences for both source and the target are considered for training from METU microphone database [9]. Their HMM state-level Viterbi alignments are obtained using the University of Colorado's speech recognizer system, Sonic [10], trained on Turkish [9].

2.3. FIR Filter Adaptation Based on LMS Algorithm

This approach considers each of the 16 LSP frequencies obtained from frames of source and the target speakers separately. A block scheme of the training algorithm is shown in Figure 2. Using DTW, which minimizes the Euclidian distance between the LSP's of the source and target, the input and the desired signal are time-aligned. The filters have 10 tap weights for adaptation. The training phase results in one set of tap weights of an FIR filter for each LSP frequency within each context-independent phoneme HMM-state. The target and the source utter the same sentences for training. Their orthographic transcriptions are known and state-level alignments are also provided to the system. Here n corresponds to the frame number. Then, the LMS algorithm is applied to estimate the transformation FIR filter. This process is repeated for each HMM-state of each phoneme. The LMS update equations are given as $\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\mathbf{u}(n)\mathbf{e}^*(n)$, where $\mathbf{w}(n)$ is the tap weight vector of the adaptive FIR filter at frame number n, μ is the step size, $\mathbf{u}(n)$ is the filter input (source LSP's) and $\mathbf{e}(n)$ is the error between the output of the filter frame n and the desired response (target LSP's). The step-size parameter is selected data-dependent as $1.8/tr(\mathbf{R})$, where $\mathbf{R}$ represents the autocorrelation matrix of the input and $tr(\cdot)$ is the trace of a matrix operation. This guarantees the convergence of the LMS algorithm [11].

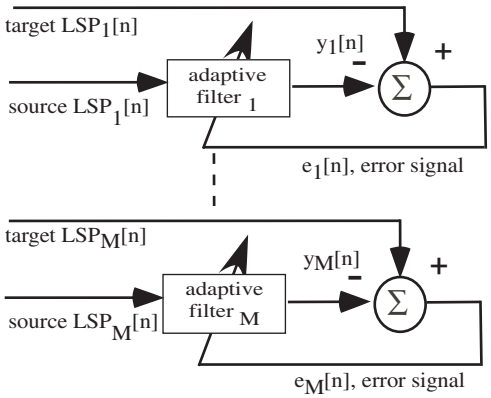

Fig. 2. LMS adaptation scheme for transforming LSP frequencies, M=16.

2.4. Steepest Descent Based Transformation Matrix Adaptation

This approach tries to obtain a transformation matrix, instead of transforming LSP's by using adaptive filters of Section 2.3, which maps source LSP's to the target LSP's. The idea is similar to the MLLR transformation approach to speaker adaptation, which tunes the HMM-mean parameters to a new speaker in a speech recognition system [12]. The transformation matrix is estimated adaptively in a similar manner to the adaptive filtering problem of Section 2.3. Equation 1 shows the transformation, $\mathbf{W}(n)\mathbf{u}(n) = \mathbf{y}(n)$, where $\mathbf{W}(n)$ is the transformation matrix, the vector $\mathbf{u}(n)$ is Mth order LSP's of the source at frame n and the vector $\mathbf{y}(n)$ is the estimated target LSP vector at frame n.

$$\begin{bmatrix} w(1,1) & w(1,2) & \cdots & w(1,M) \\ w(2,1) & w(2,2) & \cdots & w(2,M) \\ \vdots & \vdots & \ddots & \vdots \\ w(M,1) & w(M,2) & \cdots & w(M,M) \end{bmatrix} \begin{bmatrix} u_1(n) \\ u_2(n) \\ \vdots \\ u_M(n) \end{bmatrix} = \begin{bmatrix} y_1(n) \\ y_2(n) \\ \vdots \\ y_M(n) \end{bmatrix} \quad (1)$$

Let $\mathbf{d}(n) = [d_1(n), d_2(n), ..., d_m(n)]^T$ be the desired LSP vector from the target speaker. To optimize the filter design, we have chosen to minimize the mean-square value of each element of the estimation error vector, $\mathbf{e}(n) = \mathbf{d}(n) - \mathbf{y}(n)$. Then the cost function to be minimized is

$$J = E\{\mathbf{e}^T[n]\mathbf{e}[n]\}$$

where $E(\cdot)$ denotes the statistical expectation operator.

It is intuitively reasonable that successive operations to the rows of the transformation matrix $\mathbf{W}$ in the direction of the negative of the gradient vector (i.e. in the direction of the steepest descent of the error performance surface) should eventually lead to the minimum mean squared error J.

This leads to the gradient row vector in the direction of the Lth row of the transformation matrix $\mathbf{W}$, $\mathbf{w}_L[n] = [w(L,1), w(L,2), \ldots, w(L,M)]$.

Lets define

$$J_L = E\{e_L[n]e_L^*[n]\} \quad L = 1, \ldots, M$$

Then the gradient vector can be found as

$$\overline{\nabla}\, J_L[n] = -2(E\{\mathbf{u}[n]d_L[n]\} - E\{\mathbf{u}[n]\mathbf{u}^T[n]\}\mathbf{w}_L[n])$$

which will be used to update the Lth row of $\mathbf{W}$ as follows:

$$\mathbf{w}_L[n+1] = \mathbf{w}_L[n] + \frac{1}{2}\mu[-\overline{\nabla} J_L[n]] \quad L = 1, \ldots, M. \quad (2)$$

The transformation matrix $\mathbf{W}$ is determined using the update in Equation 2 for every L value. This transformation is found for every state of every phoneme during the training phase.

2.5. LSP Distance Adjustment for Stability

Both of the methods proposed in Sections 2.3 and 2.4 may result in unstable vocal tract filters (misordered LSP's). This problem should be solved in order to guarantee a high quality transformed voice. During the training process ensemble means and variances of LSF curves are obtained for both source and the target. Time means obtained over these curves are used to adjust the offsets of the LSF values of the transformed speech after one of the methods in Sections 2.3 or 2.4 is applied. This might also cause instability since LSP means are adjusted independent of each other. This problem can be addressed by adjusting the LSP distances after solving the following quadratic optimization problem:

$$min \sum_{i=1}^{N} \frac{1}{v_i}(d_i - \Delta_i)^2$$

$$LSP_S^i + \Delta_i < LSP_S^{i+1} + \Delta_{i+1},\ 1 < i < N-1$$

$$0 < LSP_S^1 + \Delta_1,\ LSP_S^N + \Delta_N < \pi$$

where LSP_S^i is the i^{th} source LSP, d_i is the difference between the i^{th} mean LSP's of the source and the target. The algorithm searches for the best Δ_i, which is nearest to the original difference d_i, but also satisfies the constraints for stability of the vocal tract filter. Weighting with the inverse of the variance v_i lets the more strictly located LSP's to be optimized more near to the original difference. This optimization results in a higher quality voice.
We have tested the results of this optimization without applying the methods in the previous sections, but after subtracting the means of the LSP's of the source and adding those of the target. This method prevented unstabilities and resulted in a higher quality transformed voice.

3. Implementation of the Conversion System

A block diagram of the overall synthesis system is shown in Figure 3. 16th order LPC parameters are estimated pitch synchronously with a window length of twice the pitch period and overlap of one pitch period. The system can synthesize speech either using the residual signal of the source speaker or the target speaker. The performances in both cases are given in Section 4. The mean pitch and the pitch range of the source speaker are both modified using the technique in [4]. When the system is in target residual mode, target residual signal is used for voiced frames and source residual is used for the unvoiced frames. One target residual for every HMM state of every phoneme is extracted and stored during training.

Speaker	Age	Gender	Median Pitch	Sentence Set
A	27	female	239 Hz	1,2,3
B	28	female	222 Hz	1
C	21	female	258 Hz	2
D	24	male	128 Hz	2,3
E	23	male	140 Hz	3

Table 1. Speakers used in the voice transformation experiments.

4. Algorithm Evaluation

4.1. Experimental Database

Speech from 5 adult (2 male and 3 female) speakers of Turkish from the METU speech database [9] is used for the evaluations. The speech is sampled at 16 kHz. The speech data consists of sets of 40 sentences randomly selected from 2460 phonetically balanced sentences. Table 1 shows the speakers and the sets of the sentences that they have uttered. Each set has 40 sentences. All possible transformations between the source and the target are allowed for the evaluations.

4.2. Speaker Modeling for Algorithm Evaluation

We consider assessing the performance of the proposed algorithm using Gaussian mixture model (GMM) speaker models, since it has been shown that GMMs are useful for modeling speaker identity [7]. The same approach in [7] is used.
Let $\mathbf{X}_s = (\mathbf{x}_s(1), \mathbf{x}_s(2), \ldots \mathbf{x}_s(T))$ represent a sequence of T observation vectors obtained from an input speech utterance and λ_s and λ_t represent GMMs of the source input and the target voice respectively. Then the log-likelihood ratio (LLR) of observing $\mathbf{X}_s$ given the source and the target voice models can be expressed by,

$$LLR = \log \frac{p(\mathbf{X}_s|\lambda_t)}{p(\mathbf{X}_s|\lambda_s)} = \log \left(\frac{\prod_{t=1}^{T} p(\mathbf{x}_s|\lambda_t)}{\prod_{t=1}^{T} p(\mathbf{x}_s|\lambda_s)} \right). \quad (3)$$

The measure reflects how well the input voice scores to the target model relative to the source model. This is a two-way classifier system, which results in a negative value when the input is more likely to be the source and a positive value when the input is more likely to be the target.

4.3. Evaluations

The continuous speech utterances in the database were analyzed every 10 msec using a 25 msec Hanning window. Silence sections were removed from each utterance and observation vectors consisting of 20 mel-cepstral coefficients were calculated. A total of 32 Gaussian densities were used to model each voice and 10 iterations of the EM algorithm were used to estimate the voice model parameters. We first evaluate the proposed algorithms in Sections 2.3 and 2.4 by considering 35 sentences from one speaker as the source and 35 sentences from another speaker as the target voice. Later, the evalaution is repeated with held-out 5 sentence from source and their transformed forms. The log-likelihood ratio scores obtained from every sentence were then averaged for both cases. The results of these evaluations are shown in Figure 4. It is seen that the proposed algorithm increases the log-likelihood of the source speaker towards that of the target speaker with both algorithms and for transforming using both source and target residual cases. Overall, we see that log-likelihood ratio increases from -6.61 to 0.92 after transformation.

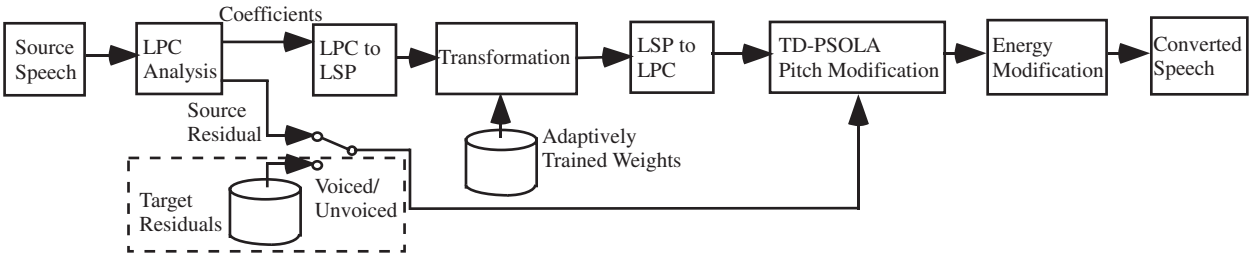

Fig. 3. System architecture diagram.

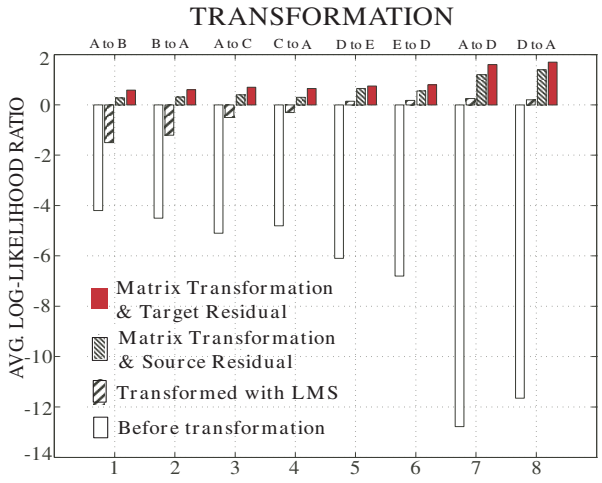

Fig. 4. Improvements in log-likelihood ratio score of speech utterances to desired target voice after transformation.

5. Conclusions

In this paper, an adaptive filter theory-based training and Line Spectral Pair (LSP) frequency distance optimization algorithm for spectral voice conversion method was formulated. In the proposed method, the transformation is performed by filtering the source speaker's LSP's to obtain the LSP's of the target speaker. In the first method, adaptive FIR filters for the LSP's of each HMM state of each phoneme are obtained. In the second method, a transformation matrix for the LSP's is obtained using the adaptive steepest descent approach. Based on the evaluation metric, the matrix transformation approach was found to provide improved spectral conversion when compared with the LMS filtering approach. These approaches are integrated into a TD-PSOLA analysis-synthesis framework. Energy scaling and pitch modification, which takes into account both pitch range and mean pitch level differences, are introduced to the final synthesis structure. The algorithm is objectively evaluated using a distance measure based on the log-likelihood ratio of observing the input speech given Gaussian mixture speaker models for both the source and the target voice. Using this criteria, the transformation matrix method resulted in an average log-likelihood increase from -6.61 to 0.92 on the average after transformation. Such a scheme for transformation could be a fast and simple voice conversion technique, which requires limited amount of training data. We are planning to integrate this approach into our Turkish TTS engine [13, 15] working on the Festival [14] framework for transforming diphone databases to different voices. We are planning to integrate a codebook-based residual transfomation to obtain better results.

6. References

[1] Pellom B., Hansen J., "An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Imposters", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1999)*, Phoenix, Arizona, USA, March 1999.

[2] Stylianou Y., Cappe O., "A System for Voice Conversion Based on Probabilistic Classification and a Harmonic Plus Noise Model", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1998)*, Seattle, Washington, USA, 1998.

[3] Kain A., Macon M., "Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelop Mapping and Residual Prediction", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2001)*, Salt Lake City, USA, June, 2001.

[4] Arslan L., Talkin D., "Speaker Transformation Using Sentence HMM Based Alignments and Detailed Prosody Modification"*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1998)*, Seattle, Washington, USA, 1998.

[5] Watanabe T., Murakami T., Namba M., et.al., "Transformation of Spectral Envelope for Voice Conversion Based on Radial Basis Function Networks", *International Conference on Spoken Language Processing (ICSLP-2002)*, Denver, USA, September 2002.

[6] Takagi T., Kuwabara H., "Contributions of the Pitch, Formant Frequency and Bandwidth to the Perception of Voice-Personality"*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1986)*, Tokyo, Japan, 1986.

[7] Pellom B., Hansen J., "Spectral Normalization Employing Hidden Markov Modeling of Line Spectrum Pair Frequencies", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1997)*, Munich, Germany, April 2000.

[8] Reynolds D., Rose R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp 72-83, Jan. 1995.

[9] Salor Ö., Pellom B., Ciloglu T., et.al., "On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language", *Proceedings of the International Conference on Spoken Language Processing (ICSLP-2002)*, Denver, USA, Sep, 2002.

[10] Pellom B., "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, March 2001.

[11] Haykin S., "Adaptive Filter Theory", Prentice-Hall, Inc., New Jersey, 1996.

[12] Leggetter C.J., "Improved Acoustic Modeling for HMMs Using Linear Transformations"*Ph.D. Thesis, Cambridge University*, 1995.

[13] Salor O., Pellom B., Ciloglu T., "New Corpora and Tools fo Turkish Speech Research"*Conference on Signal Processing and Applications (SIU-2002)*, Pamukkale, Turkey, 2002.

[14] The Festival Speech Synthesis System, *www.festvox.org*.

[15] Salor O., Pellom B., Demirekler M., *Implementation and Evaluation of a diphone based Text-to-Speech System for Turkish*. Submitted to EUROSPEECH 2003.