VOICE CONVERSION BY CODEBOOK MAPPING OF LINE SPECTRAL FREQUENCIES AND EXCITATION SPECTRUM

Levent M. Arslan and David Talkin
Entropic Research Laboratory, Washington, DC, 20003

ABSTRACT

This paper presents a new scheme for developing a voice conversion system that modifies the utterance of a source speaker to sound like speech from a target speaker. We refer to the method as Speaker Transformation Algorithm using Segmental Codebooks (STASC). Two new methods are described to perform the transformation of vocal tract and glottal excitation characteristics across speakers. In addition, the source speaker's general prosodic characteristics are modified using time-scale and pitch-scale modification algorithms. Informal listening tests suggest that convincing voice conversion is achieved while maintaining high speech quality. The performance of the proposed system is also evaluated on a standard Gaussian mixture model based speaker identification system, and the results show that the transformed speech is assigned higher likelihood by the target speaker model when compared to the source model.

1 Introduction

There has been a considerable amount of research effort directed at the problem of voice transformation recently [1, 3, 4, 8]. This topic has numerous applications which include personification of text-to-speech systems, multimedia entertainment, and as a preprocessing step to speech recognition to reduce speaker variability. In general, the approach to the problem consists of a training phase where input speech training data from source and target speakers are used to formulate a spectral transformation that would map the acoustic space of the source speaker to that of target speaker. The acoustic space can be characterized with a number of possible acoustic features which has been studied extensively in the literature. The most popular features used for voice transformation include formant frequencies [1], and LPC cepstrum coefficients [7]. The transformation is in general based on codebook mapping [1, 3, 7]. That is, a one to one correspondence between the spectral codebook entries of the source speaker and the target speaker is developed by some form of supervised vector quantization method. In general, these methods face several problems such as artifacts introduced at the boundaries between successive speech frames, limitation on robust estimation of parameters (e.g., formant frequency estimation), or distortion introduced during synthesis of target speech. Another issue which has not been explored in detail is the transformation of the glottal excitation characteristics aside from the vocal tract characteristics. Several studies proposed solutions to address this issue recently [4, 7]. In this study, we propose new and effective solutions to both problems with the goal of maintaining high speech quality.

2 Algorithm Description

This section provides a general description of the STASC algorithm. The training speech (sampled at 16 kHz) from source and target speakers are first segmented automatically using forced alignment to a phonetic translation of the orthographic transcription. Codebooks of line spectral frequencies (LSF) are used in order to represent spectral characteristics of source and target speaker vocal tract characteristics. The reason for selecting line spectral frequencies is that these parameters relate closely to formant frequencies [5], but in contrast to formant frequencies they can be estimated quite reliably. In addition, they have a fixed range which makes them attractive for real-time DSP implementation. The LSF codebooks are generated as follows: The line spectral frequencies for source and target speaker utterances are calculated on a frame-by-frame basis and each LSF vector is labeled using the phonetic segmenter. Next, a centroid LSF vector for each phoneme is estimated for both source and target speaker codebooks by averaging across all the corresponding speech frames. A one-to-one mapping is established from the source and target codebooks to accomplish the voice transformation. The transformation will be explained in detail later in this section.

Another factor that influences speaker individuality is glottal excitation characteristics. The LPC residual can be a reasonable approximation to the glottal excitation signal. It is well known that the residual can be very different for different phonemes (e.g., periodic pulse train for voiced sounds versus white noise for unvoiced sounds). Therefore, we formulated a "codebook based" transformation of the excitation characteristics similar to the one discussed above for vocal tract spectrum transformation. Codebooks for excitation characteristics are obtained as follows: Using the segmentation information, the LPC residual signals for each phoneme in the codebook are collected from the training data. Next, a short-time average magnitude spectrum of the excitation signal is estimated for each phoneme both for the source speaker and the target speaker pitch synchronously. An excitation transformation filter can be formulated for each codeword entry using the excitation spectra of the source speaker and the target speaker. This method not only transforms the excitation characteristics, but it estimates a reasonable transformation for the "zeros" in the spectrum as well, which are not represented accurately by the all-pole modeling. Therefore, this method resulted in improved voice conversion performance especially for nasalized sounds.

The flow diagram of the STASC voice transformation algorithm is shown in Figure 1. The incoming speech is first sampled at 16 kHz and preemphasized with the filter $P(z) = 1 - 0.95z^{-1}$. Next, 18th order LPC analysis is performed to estimate the prediction coefficients. Based on the prediction coefficients, an inverse filter, $A(z)$, is formulated as:

$$A(z) \quad = \quad 1 - \sum_{k=1}^{P} \mathbf{a}_k z^{-k}. \qquad (1)$$

This filter is used to estimate $g_s(n)$ which is an approximation of the excitation signal for the source speaker. Next, line

spectral frequencies, $\mathbf{w}$, are derived from the prediction coefficients. In order to find the corresponding LSF vector for the target speaker, the most likely phones in the source codebook are estimated, and weights are assigned to each of them based on their relative likelihoods. Next, the same set of weights is used to construct the target LSF vector from target phone codebook.

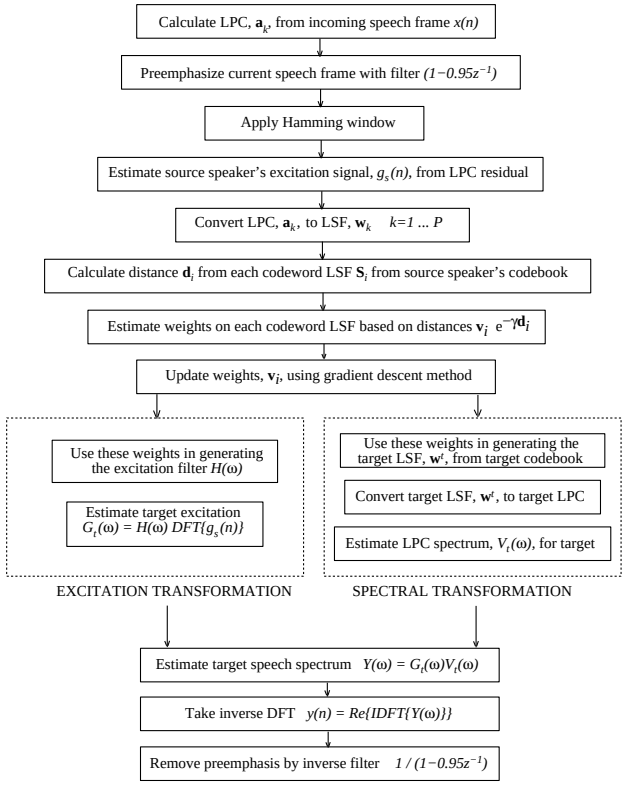

Figure 1: Flow-diagram of STASC voice conversion algorithm.

Codebook Weight Estimation Method

Line spectral frequency vector $\mathbf{w}$ is compared with each LSF centroid, $\mathbf{S}_i$, in the source codebook and the distance, $\mathbf{d}_i$, corresponding to each codeword is calculated. The distance calculation is based on a perceptual criterion where closely spaced line spectral frequencies which are likely to correspond to formant locations are assigned higher weights. The weights of the line spectral frequencies are calculated based on the formulation proposed in [6],

$$\begin{aligned} \mathbf{h}_k &= \frac{1}{\text{argmin}(|\mathbf{w}_k - \mathbf{w}_{k-1}|, |\mathbf{w}_k - \mathbf{w}_{k+1}|)} \quad k = 1, \ldots, P \\ \mathbf{d}_i &= \sum_{k=1}^{P} \mathbf{h}_k |\mathbf{w}_k - \mathbf{S}_{ik}| \qquad i = 1, \ldots, L \end{aligned} \tag{2}$$

where L is the codebook size. In addition to above weighting, for voiced segments, lower order LSFs, and for unvoiced segments, higher order LSFs are weighted more by an exponential weighting factor. Based on the distances from each codebook entry, an approximate line spectral frequency vector can be expressed as a weighted sum of the source codebook line spectral frequencies [2]:

$$\begin{aligned} \mathbf{v}_i &= \frac{e^{-\gamma \mathbf{d}_i}}{\sum_{l=1}^{L} e^{-\gamma \mathbf{d}_l}} \qquad i = 1, \ldots, L \\ \tilde{\mathbf{w}}_k &= \sum_{i=1}^{L} \mathbf{v}_i \mathbf{S}_{ik} \qquad k = 1, \ldots, P \end{aligned} \tag{3}$$

where the value of γ for each frame is found by an incremental search with the criterion of minimizing the perceptual weighted distance between the approximated LSF vector $\tilde{\mathbf{w}}$ and original LSF vector $\mathbf{w}$. However this set of weights may still not be the optimal set of weights that would represent the original speech spectrum. In order to improve the estimate of weights a gradient descent algorithm is employed. The previously estimated weights are used as the initial seed to the gradient descent algorithm. The weights $\mathbf{v}_i$ are constrained to have positive values after each iteration of the gradient descent algorithm to prevent unreasonable estimates. The weight estimation algorithm can be summarized as follows:

Codebook Weight Update by Gradient Descent

$$\begin{aligned} Initialize: \quad E^0 &= \infty \\ n &= 1; \\ Loop \\ \mathbf{e} &= \mathbf{h} \cdot (\mathbf{w} - \mathbf{S}\mathbf{v}^{n-1}) \\ E^n &= \sum_{k=1}^{P} |\mathbf{e}_k| \\ \mathbf{v}_i^n &= \mathbf{v}_i^{n-1} + \eta \mathbf{e}^T \mathbf{S}_i \quad i = 1, \ldots, L \\ \mathbf{v}_i^n &= \max(\mathbf{v}_i^n, 0) \quad i = 1, \ldots, L \\ n &= n + 1 \\ until \quad E^n &> E^{n-1} - 1.0e^{-4} E^{n-1} \end{aligned}$$

where $\mathbf{S}$ is a $P \times L$ size matrix whose columns represent a codeword LSF vector, and η is a constant which controls the rate of convergence. In our implementation, in order to reduce computation η is adjusted after each iteration based on the reduction in error E^n with respect to E^{n-1}. If there is significant amount of reduction in error then η is increased, otherwise it is decreased. It was also observed that only a few codebook entries were assigned significantly large weight values (i.e. $\mathbf{v}_i^0$). Therefore in order to save computational resources the gradient descent algorithm was performed on only 5 most likely codeword weights. Using the gradient descent method, a 15-20% additional reduction in average Itakura-Saito distance between the original and approximated spectra was achieved. The average spectral distortion (SD), which is a commonly used measure for spectral quantizer performance evaluation, was also reduced from 1.8 dB to 1.4 dB.

Glottal Excitation Mapping

The estimated set of codebook weights can be regarded as information about the phonetic content of the current speech frame. It can be utilized in two separate domains:, i) transformation of the glottal excitation characteristics, ii) transformation of the vocal tract characteristics. For transformation of the glottal excitation, the set of weights is used to construct an overall filter which is a weighted combination of excitation

codeword filters:

$$H(\omega) = \sum_{i=1}^{L} \mathbf{v}_i \frac{\mathbf{U}_i{}^t(\omega)}{\mathbf{U}_i{}^s(\omega)} \quad (4)$$

where $\mathbf{U}_i{}^t(\omega)$ and $\mathbf{U}_i{}^s(\omega)$ denote average target and source excitation spectra for the i^{th} codeword respectively. The target excitation spectrum $G_t(\omega)$ can be obtained by applying this filter to the DFT of the source speaker excitation signal $g_s(n)$:

$$G_t(\omega) = H(\omega)\mathrm{DFT}\{g_s(n)\}. \quad (5)$$

Spectral Mapping

The same set of codebook weights ($\mathbf{v}^i,\ i = 1, \ldots, L$) are applied to target LSF vectors ($\mathbf{T}_i,\ i = 1, \ldots, L$) to construct the target line spectral frequency vector $\tilde{\mathbf{w}}^t$:

$$\tilde{\mathbf{w}}_k^t = \sum_{i=1}^{L} \mathbf{v}_i \mathbf{T}_{ik}, \qquad k = 1, \ldots, P \quad (6)$$

Next, target line spectral frequencies are converted to prediction coefficients, $\mathbf{a}^t$, which in turn are used to estimate the target LPC vocal tract filter:

$$V_t(\omega) = \left| \frac{1}{1 - \sum_{k=1}^{P} \mathbf{a}_k e^{-jk\omega}} \right|^{\beta}. \quad (7)$$

The weighted codebook representation of the target spectrum results in expansion of formant bandwidths. In order to cope with this problem a new bandwidth modification algorithm is proposed.

Bandwidth Modification Method

The bandwidth modification algorithm makes use of the knowledge that average formant bandwidth values of the target speech should be similar to that of source speech. Once an estimate of the source speech bandwidths is obtained, the bandwidths of the target speech can be forced to be similar to this estimate by modifying the distance between line spectrum pairs representing each formant. The algorithm can be formulated as follows. First, estimate the average formant bandwidth across first four formant frequencies from both the source speech spectrum and current estimate of the target speech spectrum to find the ratio r:

$$\bar{\mathbf{b}_s} = \frac{1}{4}\sum_{i=1}^{4} \mathbf{b}_{si} \qquad \bar{\mathbf{b}_t} = \frac{1}{4}\sum_{i=1}^{4} \mathbf{b}_{ti}$$

$$r = \frac{\bar{\mathbf{b}_s}}{\bar{\mathbf{b}_t}}. \quad (8)$$

In the above formulation bandwidths are approximated by the difference of closely spaced LSFs. Next, find the line spectral pairs in the target LSF vector that correspond to each formant frequency location $\mathbf{f}_i$, $i = 1 \ldots 4$. Finally, apply the estimated bandwidth ratio to adjust the line spectral pairs:

$$\begin{aligned} \mathbf{w}_j^i &= \mathbf{w}_j^i + (1 - r) * (\mathbf{f}_i - \mathbf{w}_j^i) \\ \mathbf{w}_{j+1}^i &= \mathbf{w}_{j+1}^i + (1 - r) * (\mathbf{f}_i - \mathbf{w}_{j+1}^i) \qquad i = 1 \ldots 4 \end{aligned} \quad (9)$$

where $\mathbf{w}_j^i$ and $\mathbf{w}_{j+1}^i$ represent a line spectral frequency pair around $\mathbf{f}_i$. In order to prevent the estimation of unreasonable bandwidths the minimum bandwidth value is set to be 50 Hz. Using this processing method resulted in more accurate bandwidth estimation for the vocal tract filter $V_t(\omega)$ based on detailed observations and subjective listening tests.

Combined Output

The vocal tract filter is next applied to the spectrum of the estimated target excitation signal to get an estimate of the spectrum corresponding to the preemphasized target speech:

$$Y(\omega) = G_t(\omega)V_t(\omega). \quad (10)$$

Next, inverse DFT is applied to produce the synthetic target voice,

$$y(n) = \mathrm{Real}\{\mathrm{IDFT}\{Y(\omega)\}\}. \quad (11)$$

Finally preemphasis is removed from the speech by applying inverse preemphasis filter:

$$P^{-1}(z) = \frac{1}{1 - 0.95z^{-1}}. \quad (12)$$

The next section discusses the evaluations conducted to test the performance of the STASC algorithm.

3 Evaluations

In order to evaluate the performance of the STASC algorithm we used a simple speaker identification system. The idea is that if we can make the speaker identification (ID) system select the target speaker after processing source speaker utterance, it means that the voice conversion algorithm is performing well. Of course besides checking for the binary decision between the two speakers, one would like to have a confidence measure on the decision as well. For this reason, the log-likelihood ratio of the target speaker to that of the source speaker is adopted as an objective measure in our evaluations. The performance measure θ_{st} can be formulated as:

$$\begin{aligned} \theta_{st} &= \log \frac{P(\mathbf{X}|\lambda_t)}{P(\mathbf{X}|\lambda_s)} \\ &= \log P(\mathbf{X}|\lambda_t) - \log P(\mathbf{X}|\lambda_s) \end{aligned} \quad (13)$$

where $\mathbf{X}$ is the observation vector sequence, λ_t is the target speaker model, and λ_s is the source speaker model. The speaker ID system employs Gaussian mixture models in order to represent the source and target speaker characteristics. The 24 dimensional feature vector consists of 12 Mel-Cepstrum coefficients and their delta coefficients. Initial vector quantization was done using binary split vector quantization method. This was followed by 2 iterations of Forward-Backward training. During data collection sessions each speaker was asked to read a different story to the tape recorder. The recorded speech was approximately one hour long for each speaker. Forty-five minutes of the recording was used as training data (both for speaker ID models and voice transformation codebooks), and fifteen minutes of speech was set aside for testing. The average likelihood of the first speaker's speech data for

Speaker ID Evaluation of Voice Conversion Algorithm		
$Testcase$	θ_{st} before conversion	θ_{st} after conversion
$Sp1->Sp2$	-5.59	+5.47
$Sp1->Sp3$	-4.29	+3.22
$Sp2->Sp1$	-6.22	+1.51
$Sp2->Sp3$	-6.55	+3.98
$Sp3->Sp1$	-3.57	+0.47
$Sp3->Sp2$	-4.70	+4.53

Table 1: The speaker ID evaluation for voice conversion. $Sp1$: first speaker, $Sp2$: second speaker, $Sp3$: third speaker.

the first speaker model, $\log P(\mathbf{X}|\lambda_1)$, was -70.53. After using STASC for transformation to the second speaker, first speaker model likelihood reduced to -72.62, and second speaker model likelihood increased from -76.12 to -67.15. This is expressed in the Table 1 in terms of log-likelihood ratio as an increase from -5.59 to +5.47. The transformation was not as successful for every speaker combination. For instance after conversion from third speaker (Sp3) to first speaker (Sp1) the likelihoods showed smaller differences (θ_{31} : $-3.57 \rightarrow +0.47$). However, in all cases the likelihoods moved in the right directions for source and target speakers (i.e., away from the source speaker, and towards the target speaker). In Figure 2, the illustration

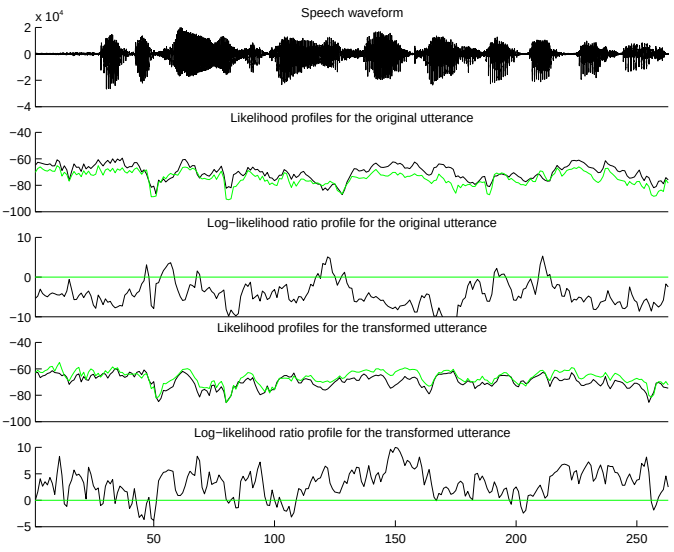

Figure 2: Illustration of speaker conversion algorithm performance in terms of speaker ID system likelihoods across time (*solid line*:source speaker likelihood, *dashed line*: target speaker likelihood).

of the algorithm performance using speaker likelihood criterion on a sample test utterance is shown. Here, it can be seen that the voice conversion performance also depends on the context, and for some phonemes it is more successful, whereas it does not perform as well for others. Part of this can be explained by the fact that same VQ indices are not forced to be used in speaker ID system, and another mixture combination from the source speaker may represent the target speaker characteristics in some cases.

4 Conclusion

In this study, a new voice conversion algorithm is developed. The algorithm is based on a codebook mapping idea, however it uses a weighted average of codewords to represent each speech frame which results in smoother transition across successive frames. Both vocal tract characteristics and glottal excitation characteristics are transformed within the same framework which makes the algorithm computationally tractable. In addition, average prosodic characteristics are modified by time-scale and pitch-scale modification algorithms. As a result, high quality speech which characterizes the target speaker was obtained after the STASC algorithm was employed for voice conversion. The performance of the algorithm was tested by a standard speaker ID system as well as informal listening tests. The objective evaluations verified that the target speaker characteristics are captured after spectral transformation is employed. The algorithm produces very convincing speech output with high quality. The current algorithm performs the prosody modification and spectrum+excitation transformation sequentially. As a future study, the integration of these two components in a single framework will be investigated.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. "Voice Conversion through Vector Quantization". In *Proc. IEEE ICASSP*, pages 565–568, 1988.

[2] L.M. Arslan, A. McCree, and V. Viswanathan. "New Methods for Adaptive Noise Suppression". In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 812–815, Detroit, USA, May 1995.

[3] G. Baudoin and Y. Stylianou. "On the transformation of the speech spectrum for voice conversion". In *Proceedings ICSLP*, pages 1405–1408, Philadelphia, USA, 1996.

[4] D.G. Childers. "Glottal source modelling for voice conversion". *Speech Communication*, 16(2):127–138, February 1995.

[5] J.R. Crosmer. *Very low bit rate speech coding using the line spectrum pair transformation of the LPC coefficients*. PhD thesis, Elec. Eng., Georgia Inst. Technology, 1985.

[6] R. Laroia, N. Phamdo, and N. Farvardin. "Robust and Efficient Quantization of Speech LSP Parameters Using Structured Vector Quantizers". In *Proc. IEEE ICASSP*, pages 641–644, 1991.

[7] K.S. Lee, D.H. Youn, and I.W. Cha. "A new voice transformation method based on both linear and nonlinear prediction analysis". In *Proceedings ICSLP*, pages 1401–1404, Philadelphia, USA, 1996.

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