SPEAKER TRANSFORMATION USING SENTENCE HMM BASED ALIGNMENTS AND DETAILED PROSODY MODIFICATION

Levent M. Arslan and David Talkin

Entropic Research Laboratory, Washington, DC, 20003

ABSTRACT

This paper presents several improvements to our voice conversion system which we refer to as Speaker Transformation Algorithm using Segmental Codebooks (STASC)[2]. First, a new concept, sentence HMM, is introduced for the alignment of speech waveforms sharing the same text. This alignment technique allows reliable and high resolution mapping between two speech waveforms. In addition, it is observed that energy and speaking rate differences between two speakers are not constant across all phonemes. Therefore a codebook based duration and energy scaling algorithm is proposed. Finally, a more detailed pitch modification is introduced that takes into account pitch range differences between source and target speakers in addition to mean pitch level differences. The proposed changes made a significant impact on the quality of transformed speech. Subjective listening tests showed that intelligibility is maintained at the same level as natural speech after the speaker transformation.

1. INTRODUCTION

There has been a considerable amount of research effort directed at the problem of voice transformation recently [1, 2, 3, 4, 6]. 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 parametric spectral transformation that would map the acoustic space of the source speaker to that of the target speaker. The transformation is in general based on codebook mapping [1, 3, 7]. That is, a one to one correspondence between spectral codebook entries of the source speaker and the target speaker is developed by some form of supervised vector quantization method. Therefore, it is crucial for the success of the mapping to obtain good alignments between source and target speaker utterances. Traditionally, a phonetic alignment or dynamic time warping algorithm is applied to extract the corresponding speech units from these utterances. In this paper, we are introducing a new method for the alignment process using sentence HMMs. Using this method, we were able to improve the quality of our system when compared to our previous approach of using phonetic alignments.

In addition, we propose a set of prosody mapping techniques to match target speaker prosody characteristics. Using the proposed methods, we were able to improve the performance of our system when compared to our previous approach of using phonetic alignments.

2. ALGORITHM DESCRIPTION

This section provides a general description of the STASC algorithm. We will describe the algorithm under two main sections: i) transformation of spectral characteristics, ii) transformation of prosodic characteristics.

2.1. Spectral Transformation

In STASC, vocal tract characteristics of source and target speakers are represented by codebooks of line spectral frequencies (LSF). The reason for selecting LSFs 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 dynamic range which makes them attractive for real-time DSP implementation. In order to obtain LSF codebooks first we need to align source and target speaker utterances so that we can formulate a one-to-one mapping. Our previous method for performing the alignment task involved a forced alignment of source and target speaker utterances to a phonetic translation of the orthographic transcription. One disadvantage of this method is that it requires the transcription of speech and a phonetic dictionary. This can be a tedious task in a multi-lingual system (e.g. automatic language translation). Another disadvantage is that speaker independent models may not match very well with specific recording conditions, and may result in poor alignments. In this paper, we are proposing a new method, sentence HMM based alignment, to overcome these difficulties.

Sentence HMMs

The proposed sentence HMM method does not require the phonetic translation of the orthographic transcription for the training utterances, however it assumes that source and target talkers are speaking the same sentences. Phonetically balanced sentences can be selected in order to minimize the amount of training data required. After the training data is collected, silence regions at the beginning and end of each utterance are removed. Next, cepstrum coefficients, log-energy and probability of voicing along with their delta coefficients are extracted for each analysis frame in each utterance. Utterance-mean subtraction is applied to the parameter vector to obtain a more robust spectral estimate. Based on the parameter vector sequences, sentence HMMs are trained for each training utterance of the target speaker. The initial number of states in sentence HMMs is set proportional to the duration of each utterance. The training is done using segmental k-means algorithm followed by Baum-Welch algorithm. During initial training, the states with similar mean spectral vectors are collapsed into single states to avoid the use of unnecessary states. The initial covariance matrix is estimated over the complete training data-set, and is not updated during the training since the amount of data corre-

sponding to each state is not sufficient to make a reliable estimate of the variance. Next, the best state sequence for each utterance is estimated using Viterbi algorithm. The average LSF vector for each state is calculated both for the source and target speakers using frame vectors corresponding to the state index. Finally these average LSF vectors for each sentence are collected to build source and target speaker codebooks. In Figure 1, the alignments to state indices are shown for the sentence "She had your" both for the source and target speaker utterances. From the figure, it can be observed that very detailed acoustic alignment is performed quite accurately using sentence HMMs.

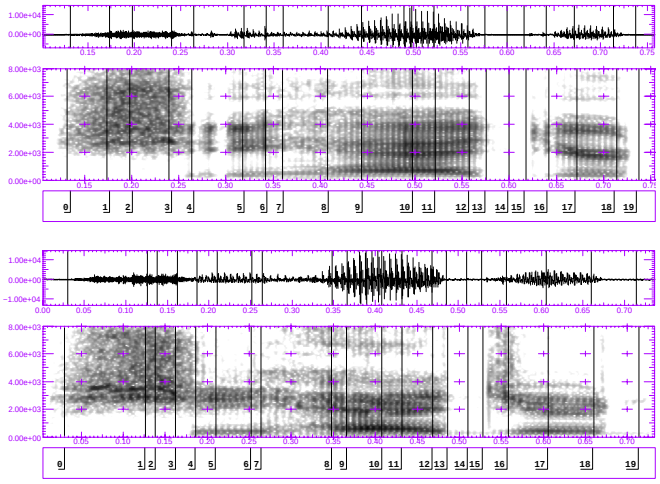

Figure 1: Sentence HMM based state alignments for source and target speaker utterances "She had your".

Source to Target mapping

The flow diagram of the STASC voice transformation algorithm is shown in Figure 2. The incoming speech is first sampled at 16 kHz. Next, 18th order LPC analysis is performed to estimate the prediction coefficients vector $\mathbf{a}$.

Based on the source-filter theory, the incoming speech spectrum $X(\omega)$ can be represented as

$$X(\omega) \quad = \quad G_s(\omega)V_s(\omega), \tag{1}$$

where $G_s(\omega)$ and $V_s(\omega)$ represent source speaker glottal excitation and vocal tract spectrums respectively for the incoming speech frame $x(n)$.

The target speech spectrum $Y(\omega)$ can be formulated as:

$$Y(\omega) \quad = \quad \left[\frac{G_t(\omega)}{G_s(\omega)}\right]\left[\frac{V_t(\omega)}{V_s(\omega)}\right] X(\omega) \tag{2}$$

where $V_t(\omega)$ and $G_t(\omega)$ represent codebook estimated target vocal tract and glottal excitation spectrums respectively. The source speaker vocal tract spectrum $V_s(\omega)$ can be estimated from the original LPC vector $\mathbf{a}$:

$$V_s(\omega) \quad = \quad \left|\frac{1}{1-\sum_{k=1}^{P}\mathbf{a}_k e^{-jkw}}\right|^{\frac{1}{2}}. \tag{3}$$

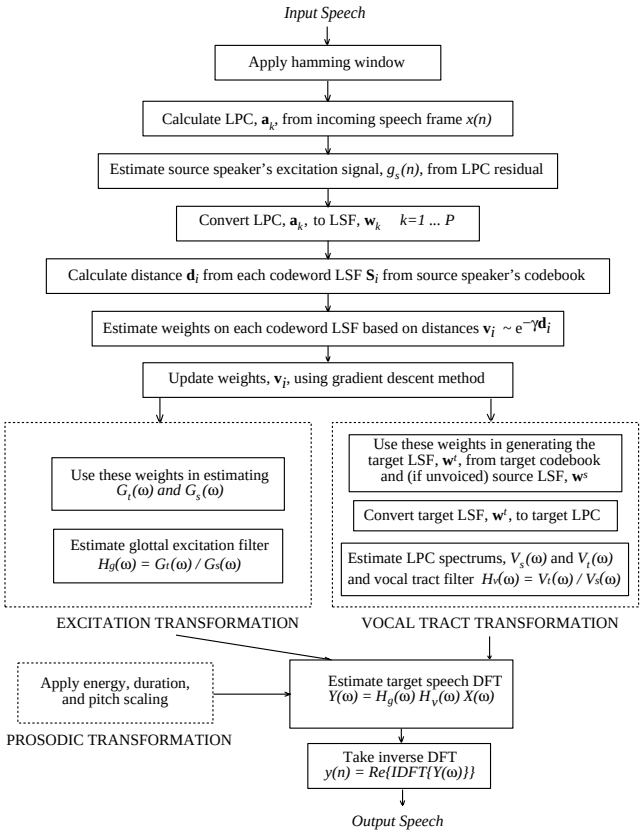

Figure 2: STASC voice conversion algorithm flowchart.

Glottal Excitation Spectrum Mapping

In order to estimate the target excitation and vocal tract parameters, first the incoming source speech spectrum (LSF representation) is approximated as a weighted combination of source codebook LSF vectors:

$$\tilde{\mathbf{w}}_k^s \quad = \sum_{i=1}^{L} \mathbf{v}_i \mathbf{S}_{ik} \qquad k = 1, \ldots, P \tag{4}$$

where $\mathbf{S}_i$ is the i^{th} codeword LSF vector and $\mathbf{v}_i$ represents its weight. The codebook weight estimation procedure is described in detail in [2]. The estimated set of codebook weights 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:

$$H_g(\omega) = \left[\frac{G_t(\omega)}{G_s(\omega)}\right] = \sum_{i=1}^{L} \mathbf{v}_i \frac{\mathbf{U}_i{}^t(\omega)}{\mathbf{U}_i{}^s(\omega)} \tag{5}$$

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.

Vocal Tract Spectrum Mapping

The same set of codebook weights ($\mathbf{v}^i,\ i = 1, \ldots, L$) are applied to target codebook LSF vectors ($\mathbf{T}_i,\ i = 1, \ldots, L$) to construct

the target LSF vector $\tilde{\mathbf{w}}^t$:

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

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

$$V_t(\omega) = \left| \frac{1}{1 - \sum_{k=1}^{P} \mathbf{a}_k e^{-jk\omega}} \right|^{\frac{1}{2}}. \tag{7}$$

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 used and is described in [2].

Combined Output

The vocal tract filter and glottal excitation filters are next applied to the magnitude spectrum of the original source signal to get an estimate of the DFT corresponding to target speech:

$$Y(\omega) = H_g(\omega) \frac{V_t(\omega)}{V_s(\omega)} X(\omega). \tag{8}$$

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

$$y(n) = \mathrm{Real}\{\mathrm{IDFT}\{Y(\omega)\}\}. \tag{9}$$

2.2. Prosodic Transformation

In STASC, a frequency domain pitch synchronous analysis synthesis framework is adopted in order to be able to realize both spectral and prosodic transformations simultaneously. In addition to the spectral transformation discussed in the previous section pitch, duration, and amplitude are modified to mimic target speaker prosodic characteristics. Each analysis frame length is set to be constant for unvoiced regions. For voiced regions the frame length is set to two or three pitch periods depending on the pitch modification factor. It is observed that when the pitch modification factor is less than one using smaller frame lengths reduces artifacts introduced by the modification.

Pitch-Scale Modification

The pitch modification involves matching both the average pitch value and range for the target speaker. This is accomplished by modifying the source speaker fundamental frequency, f_0^s, by a multiplicative constant a and an additive constant b:

$$f_0^t = a f_0^s + b \tag{10}$$

The value for a is set so that the source speaker pitch variance σ_s^2, and target speaker pitch variance σ_t^2 match, i.e.,

$$a = \sqrt{\frac{\sigma_t^2}{\sigma_s^2}} \tag{11}$$

Once the value for a is set, the value for the additive constant b can be found by matching the average f_0 values.

$$b = \mu_t - a\mu_s \tag{12}$$

where μ_s and μ_t represent source and target mean pitch values. Therefore, the pitch scale modification factor β at each frame can be set as

$$\beta = \frac{a f_0^s + b}{f_0^s} \tag{13}$$

in order to achieve the desired target speaker's mean pitch value and range.

Duration-Scale Modification

The duration characteristics can vary across different speakers significantly due to a number of factors including accent or dialect. Although modifying the speaking rate uniformly to match the target speaker duration characteristics reduces timing differences between speakers to some extent, it is observed that this is not sufficient in general. It is very well known that the variation in duration characteristics between two speakers is heavily dependent upon context. Therefore it is highly desirable to develop a method for automatically estimating the appropriate time-scale modification factor in a certain context. In STASC, a codebook based approach to duration modification is implemented. The sentence HMM alignments used for spectral mapping can also be used to generate the appropriate duration modification factor for a given speech frame. In order to accomplish this, first duration statistics are estimated based on state durations for both the source speaker and the target speaker for all the entries in the codebook. Then the same codebook weights developed for spectral mapping can be used to estimate the appropriate time-scale modification factor γ:

$$\gamma = \sum_{i=1}^{L} \mathbf{v}_i \frac{d_i^t}{d_i^s}, \tag{14}$$

where d_i^t and d_i^s represent source and target speaker durations for the i^{th} codeword.

A major application for current time-scale modification algorithms is to slow down the speech for accurate transcription by humans. The problem with most of those systems is that they use a constant time scale modification factor when changing the speaking rate. However, not all the phonemes are scaled to the same extent when a speaker modifies his/her speaking rate. Therefore, the same approach proposed here for transforming duration characteristics across speakers can be applied to speaking rate modification algorithms if the statistics for slow, normal and fast speaking styles are generated prior to the application. It should be noted that large amount of training data is essential to the success of the duration modeling method proposed here.

Energy-Scale Modification

In addition to pitch and duration, energy is another important component which characterizes the prosody of a speaker. In order to match target speaker's stress characteristics we applied a codebook based energy mapping as well. The RMS energy is scaled with a variable η at each time frame. The scaling factor can be expressed as follows:

$$\eta = \sum_{i=1}^{L} \mathbf{v}_i \frac{e_i^t}{e_i^s}, \tag{15}$$

where e_i^t and e_i^s represent average source and target speaker energies for the i^{th} codeword.

Finally, the pitch-scale modification factor β, the time-scale modification factor γ, and the energy scaling factor η are applied within a pitch-synchronous overlap-add synthesis framework to perform prosodic modification.

3. EVALUATIONS

In order to evaluate the performance of the STASC algorithm we performed an objective scoring test and a subjective listening experiment. The objective scoring test involved the comparison of three conditions with respect to target speech: untransformed source speech, transformed speech using previous STASC, and transformed speech using new STASC. The objective scores are based on sentence HMM alignments between target speech and each processing condition. The objective scores are evaluated under 3 categories: i) vocal tract spectrum match, ii) duration match, and iii) RMS energy match. Vocal tract spectrum match is evaluated with the perceptual LSF distance metric in [2]. The duration match shows the average difference between corresponding state durations of target and test utterances. Finally, RMS energy distance represents the mean distance between corresponding state RMS energies. For the training, both methods used approximately five minutes of speech from source and target talkers. Table 1 summarizes the results obtained on one minute of speech that was set aside for testing. The results show that the new STASC algorithm showed improvement in terms of mimicing both vocal tract and prosodic characteristics.

Objective Test Evaluation of STASC			
Test condition	LSF	RMS (dB)	Duration (sec)
Source speech	0.97	5.0	0.024
previous STASC	0.38	4.9	0.022
new STASC	0.32	4.7	0.017

Table 1: The objective scores for LSF perceptual distance, RMS energy distance, and duration distance between target speaker utterances and source speech under three processing conditions.

While informal listening tests showed that the transformation of speaker characteristics was successful, we wanted to test whether the transformation process introduced a degradation in intelligibility. This was necessary, since the most important application (i.e., text to speech personification) relies heavily on the level of intelligibility. The test material was 150 short nonsense sentences. One example of the sentences used in the test was "Shipping gray paint hands even". The main purpose of using nonsense sentences was to limit the ability of the listener to anticipate words based on context. Two conditions, transformed speech and natural speech, were presented to the listeners with random order. We used three inexperienced listeners to transcribe the words of the test material. Listeners were allowed to listen to each sentence up to three times. The standard NIST scoring algorithm was then used to compare the utterance and transcribed phone strings. The phone sequences were determined by dictionary look-up. The transformation tested in this experiment was from a male speaker to another male speaker. The result of the experiment was surprising. The phone accuracy for the transformed speech (93.8%) was slightly higher than it was for natural speech (93.4%). The reason for the slight increase in intelligibility might be due to measurement noise. Another possible reason might be that the target speaker was more intelligible than the source speaker, and the transformation algorithm took advantage of that. Of course, the transformation between different speaker combinations may reveal different results. When the acoustic characteristics of two speakers are extremely different (e.g., male to female transformation), we may expect degradation in intelligibility. Our future plans include testing other speaker combinations.

4. CONCLUSION

In this study, several improvements to our previous voice conversion system are described. First a new concept, sentence HMM, is introduced to refine the alignments between source and target speaker utterances. Sentence HMMs can provide more robust and finer detail alignments when compared to traditional methods such as DTW or phonetic alignments. In addition they have the advantage over phonetic alignment method that we used in our previous system of being vocabulary independent.

In terms of prosodic characteristics, previous version of STASC was only adjusting mean pitch level and speaking rate. Now, in addition to mean pitch level the pitch range is adjusted to match the target talker intonation. Moreover, codebook based duration and energy modifications are performed to capture context dependent prosodic characteristics. These enhancements to STASC resulted in better characterization of the target speaker speech. Finally, subjective tests verified that additional processing did not introduce degradation in intelligibility scores for the transformed speech.

5. 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 and D. Talkin. "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum". In *Proc. EUROSPEECH*, volume 3, pages 1347–1350, Rhodes, Greece, September 1997.

[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] H. Kuwabara and Y. Sagisaka. "Acoustic characteristics of speaker individuality: Control and conversion". *Speech Communication*, 16(2):165–173, February 1995.

[7] B.L. Pellom and J.H.L. Hansen. "Spectral Normalization employing Hidden Markov Modeling of Line Spectrum Pair Frequencies". In *Proc. IEEE ICASSP*, volume 2, pages 943–946, Munich, Germany, 1997.