ELSEVIER

Speech Communication 28 (1999) 211–226

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Speaker Transformation Algorithm using Segmental Codebooks (STASC) [1]

Levent M. Arslan [2]

Electrical and Electronics Department, Boğaziçi University, Bebek, 80815 Istanbul, Turkey

Received 27 February 1998; received in revised form 17 December 1998; accepted 8 February 1999

Abstract

This paper presents a new voice conversion algorithm which 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). A novel method is proposed which finds accurate alignments between source and target speaker utterances. Using the alignments, source speaker acoustic characteristics are mapped to target speaker acoustic characteristics. The acoustic parameters included in the mapping are vocal tract, excitation, intonation, energy, and duration characteristics. 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 simple 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 speaker model. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Voice conversion; Speaker transformation; Codebook; Line spectral frequencies; Hidden Markov models; Time-varying filter; Overlap-add analysis

1. Introduction

There has been a considerable amount of research effort directed at the problem of voice transformation recently (Abe et al., 1988; Baudoin and Stylianou, 1996; Childers, 1995; Iwahashi and Sagisaka, 1995; Kuwabara and Sagisaka, 1995; Narendranath et al., 1995; Stylianou et al., 1995). 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 the target speaker. The acoustic space can be characterized by a number of possible acoustic features which have been studied extensively in the literature. The most popular features used for voice transformation include formant frequencies (Abe et al., 1988; Narendranath et al., 1995), and LPC cepstrum coefficients (Lee et al., 1996). The transformation is in general based on codebook mapping (Abe et al., 1988; Acero, 1993; Baudoin and Stylianou, 1996; Lee et al., 1996).

[1] Speech files available. See www.elsevier.nl/locate/specom.

[2] Tel.: +90 212 2631540/1421; fax: +90 212 2872465; e-mail: arslanle@boun.edu.tr. The author was with Entropic Research Laboratory, Washington, DC.

PII: S0167-6393(99)00015-1

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 excitation characteristics in addition to the vocal tract characteristics. Several studies proposed solutions to address this issue recently (Childers, 1995; Lee et al., 1996). 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 Speaker Transformation Algorithm using Segmental Codebooks (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

For the representation of the vocal tract characteristics of source and target speakers line spectral frequencies (LSF) are selected. The reason for selecting LSFs is that these parameters relate closely to formant frequencies (Crosmer, 1985), but in contrast to formant frequencies they can be estimated quite reliably. They have been used for a number of applications successfully in the literature (Hansen and Clements, 1991; Arslan et al., 1995; Arslan and Talkin, 1997; Crosmer, 1985; Laroia et al., 1991; Itakura, 1975; Pellom and Hansen, 1997). They have good interpolation properties and they are stable (Paliwal, 1995). In addition, they have a fixed dynamic range which makes them attractive for real-time DSP implementation. LSFs can be estimated by modifying the LPC polynomial, $A(z)$, in two ways: $P(z)$ and $Q(z)$ are obtained by augmenting $A(z)$'s PARCOR sequence with $a+1$ and -1, respectively. This results in the following two polynomials which have all their roots on the unit circle:

$$\begin{aligned} P(z) &= (1 - z^{-1}) \\ &\quad \times \prod_{k=1,3,5,\ldots}^{P-1} (1 - 2\cos w_k (z^{-1} + z^{-2})), \\ Q(z) &= (1 + z^{-1}) \\ &\quad \times \prod_{k=2,4,6,\ldots}^{P-1} (1 - 2\cos w_k (z^{-1} + z^{-2})), \end{aligned} \tag{1}$$

where P is the LPC analysis order, and the angles of the roots, w_k, are LSFs. In STASC algorithm, codebooks of LSFs are used to represent the vocal tract characteristics of individual speakers. The codebooks can be generated in two ways.

The first method assumes that the orthographic transcription is available along with the training data. 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. The segmentation algorithm uses Mel-cepstrum coefficients and delta coefficients within an HMM framework and is described in detail in (Wightman and Talkin, 1994). The LSFs for source and target speaker utterances are calculated on a frame-by-frame basis and each LSFs vector is labeled using the phonetic segmenter. Next, a centroid LSFs vector for each phoneme is estimated for both source and target speaker codebooks by averaging across all the corresponding speech frames. The estimated codebook spectra for an example male source speaker and female target speaker combination from the database is shown in Fig. 1 when monophones are selected as speech units. A one-to-one mapping is established between the source and target codebooks to accomplish the voice transformation.

The second method does not require the phonetic translation of the orthographic transcription for the training utterances, however it assumes that both source and target speakers are speaking the same sentences during the training session. This method is a new method and it is referred to as "Sentence HMM" method. The method is as fol-

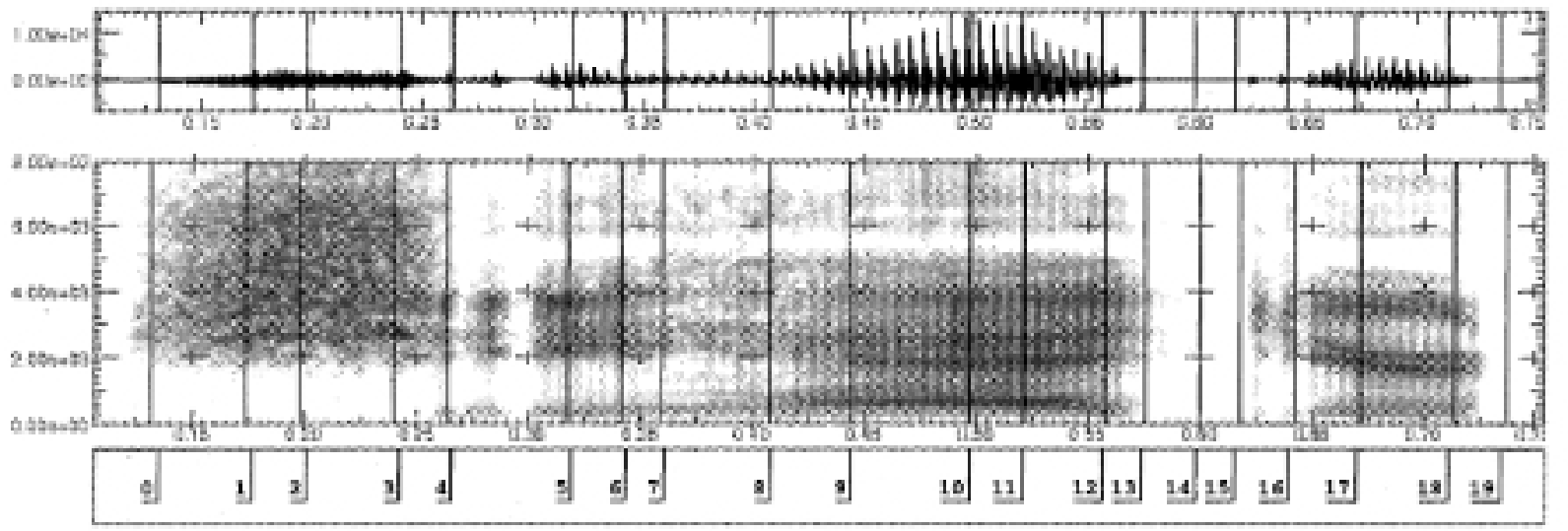

Fig. 1. The state alignments for source and target speaker utterances "She had your".

lows. First, template sentences are selected which are phonetically balanced to be uttered by the source and target speakers. After the training data are collected, silence regions at the beginning and end of each utterance are removed. Each utterance is normalized in terms of its RMS energy to account for differences in the recording gain level. Next, cepstrum coefficients are extracted along with log-energy and zero-crossing for each analysis frame in each utterance. Zero-mean normalization 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 template sentence using data from the source speaker. The number of states for each sentence HMM is set proportional to the duration of the utterance. The training is done using a segmental k-means algorithm followed by the Baum–Welch algorithm. The initial covariance matrix is estimated over the complete training dataset, and is not updated during the training since the amount of data corresponding 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 the Viterbi algorithm. The average LSFs vector for each state is calculated for both source and target speakers using frame vectors corresponding to that state index. Finally, these average LSFs vectors for each sentence are collected to build the source and target speaker codebooks. In Fig. 2, the alignments to the state indices are shown for the utterance "She had your" both for source and target speaker utterances. From the figure, it can be observed that detailed acoustic alignment is achieved quite accurately using sentence HMMs. The transformation process will be explained in detail later in this section.

Another factor that influences speaker individuality is excitation characteristic. The LPC residual can be a reasonable approximation to the excitation signal. It is well known that the residual can be very different for different phonemes (e.g., periodic