Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 24 (2010) 474–494
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Voice conversion by mapping the speaker-specific features using pitch synchronous approach

K. Sreenivasa Rao *

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721 302, India

Received 11 September 2008; received in revised form 17 January 2009; accepted 11 March 2009
Available online 24 March 2009

Abstract

The basic goal of the voice conversion system is to modify the speaker-specific characteristics, keeping the message and the environmental information contained in the speech signal intact. Speaker characteristics reflect in speech at different levels, such as, the shape of the glottal pulse (excitation source characteristics), the shape of the vocal tract (vocal tract system characteristics) and the long-term features (suprasegmental or prosodic characteristics). In this paper, we are proposing neural network models for developing mapping functions at each level. The features used for developing the mapping functions are extracted using pitch synchronous analysis. Pitch synchronous analysis provides the estimation of accurate vocal tract parameters, by analyzing the speech signal independently in each pitch period without influenced by the adjacent pitch cycles. In this work, the instants of significant excitation are used as pitch markers to perform the pitch synchronous analysis. The instants of significant excitation correspond to the instants of glottal closure (epochs) in the case of voiced speech, and to some random excitations like onset of burst in the case of nonvoiced speech. Instants of significant excitation are computed from the linear prediction (LP) residual of speech signals by using the property of average group-delay of minimum phase signals. In this paper, line spectral frequencies (LSFs) are used for representing the vocal tract characteristics, and for developing its associated mapping function. LP residual of the speech signal is viewed as excitation source, and the residual samples around the instant of glottal closure are used for mapping. Prosodic parameters at syllable and phrase levels are used for deriving the mapping function. Source and system level mapping functions are derived pitch synchronously, and the incorporation of target prosodic parameters is performed pitch synchronously using instants of significant excitation. The performance of the voice conversion system is evaluated using listening tests. The prediction accuracy of the mapping functions (neural network models) used at different levels in the proposed voice conversion system is further evaluated using objective measures such as deviation (D_i), root mean square error (μ_{RMSE}) and correlation coefficient ($\gamma_{X,Y}$). The proposed approach (i.e., mapping and modification of parameters using pitch synchronous approach) used for voice conversion is shown to be performed better compared to the earlier method (mapping the vocal tract parameters using block processing) proposed by the author.

Keywords: Mapping function; Feedforward neural network (FFNN); Pitch contour; Excitation source; LP residual; Instants of significant excitation (epochs); Prosody characteristics; Duration and energy patterns; Glottal closure; Voice conversion; Objective measures; Mean opinion score (MOS); ABX test

* Tel.: +91 3222 282336.
E-mail addresses: ksrao@iitkgp.ac.in, ksrao@sit.iitkgp.ernet.in, ksrao1969@gmail.com

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2009.03.003

1. Introduction

The basic goal of voice conversion is to transform the characteristics of an input (source) speech signal such that the output (transformed) signal is perceived to be produced by another (target) speaker. Its applications include customization of text-to-speech systems (e.g., to speak with a desired voice or to read out email in the sender's voice) as well as entertainment, story telling and security applications (Kain and Macon, 2001; Sundermann, 2005; Turk, 2007). In film industry, voice conversion systems can be employed for dubbing and translation to a different language by preserving speaker characteristics. Personification of synthesized speech is another important application as many automated systems use synthesized speech as a computer interaction tool (Turk and Arslan, 2002).

Voice transformation is generally performed in two steps. In the first step, the training stage, a set of speech feature parameters of both the source and target speakers are extracted, and appropriate mapping rules that transform the parameters of the source speaker onto those of the target speaker are generated. In the second step, the transformation stage, the features of the source signal are transformed using mapping rules developed in the training stage so that the synthesized speech possesses the personality of the target speaker (Lee, 2007).

To implement voice transformation, two problems need to be considered: what features are extracted from the underlying speech signals, and how to modify these features in such a way so that the transformed speech signals mimic target speakers voices. The first problem can be solved by identifying the speaker-specific features from the given speech signals. It is known that the shape of the glottal pulse, vocal tract transfer function and the prosodic features are uniquely characterize the speaker (Yegnanarayana et al., 2001). Feature parameters representing the vocal tract transfer function have been widely used in voice transformation. They include formant frequencies, linear prediction coefficients (LPCs), cepstrum and line spectral frequencies (LSFs) (Narendranadh et al., 1995; Arslan, 1999; Lee et al., 1996). In our previous work, we carried out the voice conversion by modifying the formant frequencies, pitch contour, duration patterns and energy profile by fixed scale factors (Rao and Yegnanarayana, 2006). As a result, the desired speaker characteristics are not much perceived in the synthesized speech. Later, we have upgraded the VC system by using LPCs for representing vocal tract system characteristics, and syllable level pitch contours for representing the intonation patterns (Rao and Koolagudi, 2007; Rao et al., 2007).

For mapping the speaker-specific features between source and target speakers, various models have been explored in the literature. These models are specific to the kind of features used for mapping. For instance, Gaussian mixture model (GMM) and vector quantization (VQ) are widely used for mapping the vocal tract characteristics (Lee, 2007; Toda et al., 2001; Abe et al., 1998). Scatter plots, GMMs and linear models are used for mapping the prosodic features (Inanoglu, 2003; Stylianou et al., 1998). In VQ, the entire speaker space was partitioned into several clusters, the mapping rules for each partition are then estimated in the form of a histogram or minimum mean square error criterion. The underlying assumption is that each cell implicitly corresponds to each phoneme. Hence, the mapping rules reflect phonetic variation. These methods, however, reveal problems, due to the hard clustering property of VQ-based classification, and leads to a discontinuity problem in transition regions (Lee, 2007). In the case of GMM, the nature of the transformed parameters depends on number of mixtures. The parameters of the target speaker will be over-smoothed in the case of less number of mixtures used for developing the GMM. Otherwise, the target speaker parameters will be discontinuous (crisp) and noisy, if the number of mixtures used for estimating the parameters is more. Choosing the optimal number of mixtures is difficult, in general for GMM. Another short coming with GMM as the mapping function is that its underlying assumption of distribution of data is the linear combination of component Gaussians (Yegnanarayana and Kishore, 2002).

In this work, we used neural network models for mapping the vocal tract characteristics, excitation source characteristics and prosodic characteristics. The main reason for exploring neural network models for developing the mapping functions at different levels (source, system and prosodic levels) is that, they are good at capturing the nonlinear relations present in the feature patterns of source and target speakers.

This paper is focused mainly on three objectives: (1) Parameterization of vocal tract characteristics (spectral characteristics) using pitch synchronous approach and analyzing its effect on the performance of voice conversion system. (2) Exploring neural network models for capturing the complex nonlinear relations between source and target speaker features at different levels. (3) Mapping of prosodic features at gross level (phrase