Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 60 (2014) 30–43

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A unit selection approach for voice transformation

Ki-Seung Lee *

Department of Electronic Engineering, Konkuk University, 1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Republic of Korea

Received 9 August 2013; received in revised form 12 February 2014; accepted 21 February 2014
Available online 4 March 2014

Abstract

A voice transformation (VT) method that can make the utterance of a source speaker mimic that of a target speaker is described. Speaker individuality transformation is achieved by altering four feature parameters, which include the linear prediction coefficients cepstrum (LPCC), ΔLPCC, LP-residual and pitch period. The main objective of this study involves construction of an optimal sequence of features selected from a target speaker's database, to maximize both the correlation probabilities between the transformed and the source features and the likelihood of the transformed features with respect to the target model. A set of two-pass conversion rules is proposed, where the feature parameters are first selected from a database then the optimal sequence of the feature parameters is then constructed in the second pass. The conversion rules were developed using a statistical approach that employed a maximum likelihood criterion. In constructing an optimal sequence of the features, a hidden Markov model (HMM) with global control variables (GCV) was employed to find the most likely combination of the features with respect to the target speaker's model.

The effectiveness of the proposed transformation method was evaluated using objective tests and formal listening tests. We confirmed that the proposed method leads to perceptually more preferred results, compared with the conventional methods.

Keywords: Voice conversion; Unit selection; Hidden Markov model

1. Introduction

Voice transformation (VT) is a process of changing the features derived from the speech signals, so that one voice is made to sound like another. If the features of one speaker (*source speaker*) are modified so that the features are close to those of another specific speaker (*target speaker*), the resultant speech signals sound as if it was spoken by target speaker. This technique is referred to as voice personality transformation. Voice personality transformation has numerous applications in a variety of areas such as personification of text-to-speech synthesis systems, preprocessing for speech recognition (Cox and Bridle, 1989), enhancing the intelligibility of abnormal speech (Bi and Qi, 1997), and foreign language training systems (Moulines and Charpentier, 1990).

Voice personality transformation is generally performed in three steps. In the first step, the analysis stage, a set of speech feature parameters of both the source and target speakers are extracted. The major issue associated with the analysis stage is to determine which features should be extracted from the underlying speech signals. The vocal-tract transfer function (VTF) is a primary identifier of speaker individuality (Childers et al., 1985). For this reason, feature parameters that represent the VTF including formant frequencies (Mizuno and Abe, 1995; Narendranath et al., 1995), the linear prediction coefficient cepsturm (LPCC) (Savic and Nam, 1991; Lee et al., 1996, 2002), Mel-frequency cepstral coefficients (MFCCs) (Stylianou et al., 1998; Toda et al., 2007; Helander et al., 2012; Huang et al., 2013; Erro et al., 2013), and Line

* Tel.: +82 196985145.
E-mail address: kseung@konkuk.ac.kr

http://dx.doi.org/10.1016/j.specom.2014.02.002
0167-6393/

Spectrum Pair (LSP) coefficients (Arslan, 1999; Rao, 2010), have been widely used in voice personality transformation. In the presented study, LPCC was used as a feature parameter that represents the VTF. Signal details beyond the LPC envelope contribute to the naturalness of speech and may also contain vital speaker information (Kain and Macon, 2001). To address this, the linear prediction residual (LPR) and the fundamental frequency (F0) were also used in the proposed voice transformation method.

In the second step, the training stage, appropriate mapping rules that transform the parameters of the source speaker onto those of the target speaker are generated. In previous studies, the entire speaker space was partitioned into several clusters using vector quantization (VQ) (Linde et al., 1980), the mapping rules for each partition are then estimated using either a histogram (Abe et al., 1988) or minimum mean square error (MMSE) criterion (Valbret et al., 1992; Lee et al., 1996). The underlying assumption is that each cell corresponds to a phoneme category. Hence these mapping rules reflect phonetic variation. However, mapping rules based on VQ present problems that result from hard clustering of VQ-based classification. According to Stylianou's study (Stylianou et al., 1998), VQ-based classification causes discontinuity in transition regions. Hence, for voice conversion, the use of a soft-clustering approach is desirable (Stylianou et al., 1998; Lee et al., 2002, 2007). In this approach, the conversion rules were built based on a MMSE criterion (Stylianou et al., 1998; Lee et al., 2002, 2007; Erro et al., 2013) or a maximum likelihood criterion (Kain and Macon, 1998; Toda et al., 2007; Saito et al., 2012). Recently, a unit-selection based approach, which was originally devised for implementing the corpus-based concatenative text-to-speech (TTS) systems (Beutnagel et al., 1999) was used to both alter the VTF parameters (Shuang et al., 2008; Jian and Zhen, 2007; Sundermann et al., 2006; Dutoit et al., 2007) and predict the target LP-residuals (Sundermann et al., 2005).

The last step of voice personality transformation is the transformation stage where 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. The pitch-synchronous overlap and add (PSOLA) method (Valbret et al., 1992), the harmonic pulse noise model (HNM) (Stylianou et al., 1998), and STRAIGHT (Kawahara et al., 1999; Helander et al., 2012; Huang et al., 2013) were often adopted to synthesize the transformed speech signals.

This paper is an extension of our previous work on voice transformation (Lee, 2007) based on a statistical approach. The listeners indicated that transformed utterances converted by the previous method sounded "ambiguous" and "unclear." This is mainly due to the bandwidth widening problem caused by the averaging effects. The artifacts caused by the averaging effects cannot be avoided in the voice transformation methods where the transformed feature vector is given by the weighted sum of the mean vectors (e.g. codebook mapping (Abe et al., 1988), Gaussian Mixture Model (GMM)-based (Stylianou et al., 1998) and Minimum Mean Squared Error (MMSE)-based (Lee, 2007).

To alleviate this problem, a conversion method based on the maximum-likelihood estimation of a spectral parameter trajectory was proposed by Toda et al. (2007). In the present study, an approach based on feature-selection was employed, where the sequence of the transformed features is given by the sequence of the features selected from the target speaker's database. Such an approach was first proposed by Dutoit et al. (2007) where pre-selection of group of frames followed by frame selection was employed. In this method, the stream of frames was built so as to minimize the weighted linear combination of the target cost and the concatenation cost. This method considered only similarity with respect to the targets which are estimated by GMM-based transformation. We propose herein selection of the features that optimize the overall similarities between the transformed and the target features by maximizing two likelihood functions: the correlation probability between the transformed and the source parameters and the likelihood of the transformed parameters with respect to the target model. A similar approach was proposed by Saito et al. (2012), where speaker GMM of the target and probability densities of joint vectors of a source and a target speakers were taken into consideration in the conversion rules. Our method is based on the assumption that because LPCC, LP-residual, and pitch originated from one source (speaker), these variables may be related. Thus, the natural quality of synthetic speech improved when these relationships were considered during the selection process. The relationship between the VTF and the source-related features has been investigated in several studies. The interaction between the VTF and the glottal source was experimentally proven by Childers and Wong (1994) who found that the first formant in voiced speech was related to characteristics of the glottal pulse. The interaction between formant frequency and pitch was investigated to judge voice category (Erickson, 2003). Component grouping of pitch and spectral information was also proposed for implementation of voice transformation (Ma and Liu, 2005). To integrate this relationship into the mapping rules, we first defined a model that describes the relationship among the employed features (LPCC, ΔLPCC, LP-residuals, and pitch). This model was then integrated with the underlying transformation rules. In the present study, the occurrence of each feature was assumed to be controlled by both *intra*/*inter* probabilistic models. The term *intra model* is one that describes intra-feature variability, whereas an *inter model* describes inter-feature variability. In the proposed method, *intra*/*inter* probabilistic models are achieved by replacing the mixture weights in GMM with the cross correlation probabilities for each feature. The cross correlation probability density functions (PDFs) for each feature commonly include a shared random variable, which is referred to as the global control variable (GCV) (Lee, 2008). Thus, the occurrence of each feature