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

is globally controlled by the GCV. We assumed that the GCVs, like other speech parameters, can be modeled by means of a quasi-stationary random process (Rabiner and Schafer, 1978; Rabiner and Juang, 1993). Therefore the hidden Markov model (HMM) was adopted for description of the target model involving the GCVs.

Objective and subjective tests were performed to evaluate the efficiency of the proposed method. For the objective tests, both the distance reduction ratio and the global variance were used to evaluate performance of the transformation. Subjective evaluations of the quality of the converted speech and of its similarity to the target speech were conducted in formal listening test.

This paper is organized as follows. Section 2 provides an overview of the proposed VT method; including both the training and online transformation procedures. Section 3 describes both the modeling and transformation of the features. The experimental results and the discussion associated with practical implementation are presented in Section 4. Finally, concluding remarks are summarized in Section 5.

2. Overview of the voice transformation system

A block diagram of the proposed voice personality transformation system is shown in Fig. 1.

In the training stage, voices from both source and target speakers were recorded. These speech samples were then analyzed for determination of the feature parameters to be transformed. In this work, the LPCC, ΔLPCC, spectrum of the LP-residual magnitude (LPRM) and the pitch were used as the feature parameters. A time domain pitch synchronous analysis framework was adopted in the proposed voice transformation system. 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. In practice, even if two speakers utter the same words, given their different speaking rates, it is unlikely that a synchronized set of LPCC sequences would result. To time-align these sequences, dynamic time warping (DTW) (White and Neely, 1976) was applied in a preprocessing step. The resulting time-aligned LPCC, ΔLPCC, LPRM, and pitch sequences were used to build conversion rules for each feature parameter. The prediction rules for LP-residual phase (LPRP) were also constructed in the training stage. Note that pitch is valid only for the voiced frames. Hence, the unvoiced frames were eliminated from the time-aligned pitch sequences. The main steps in parameterizing the LP-residual magnitude/phase spectra for each voiced frame are as follows.

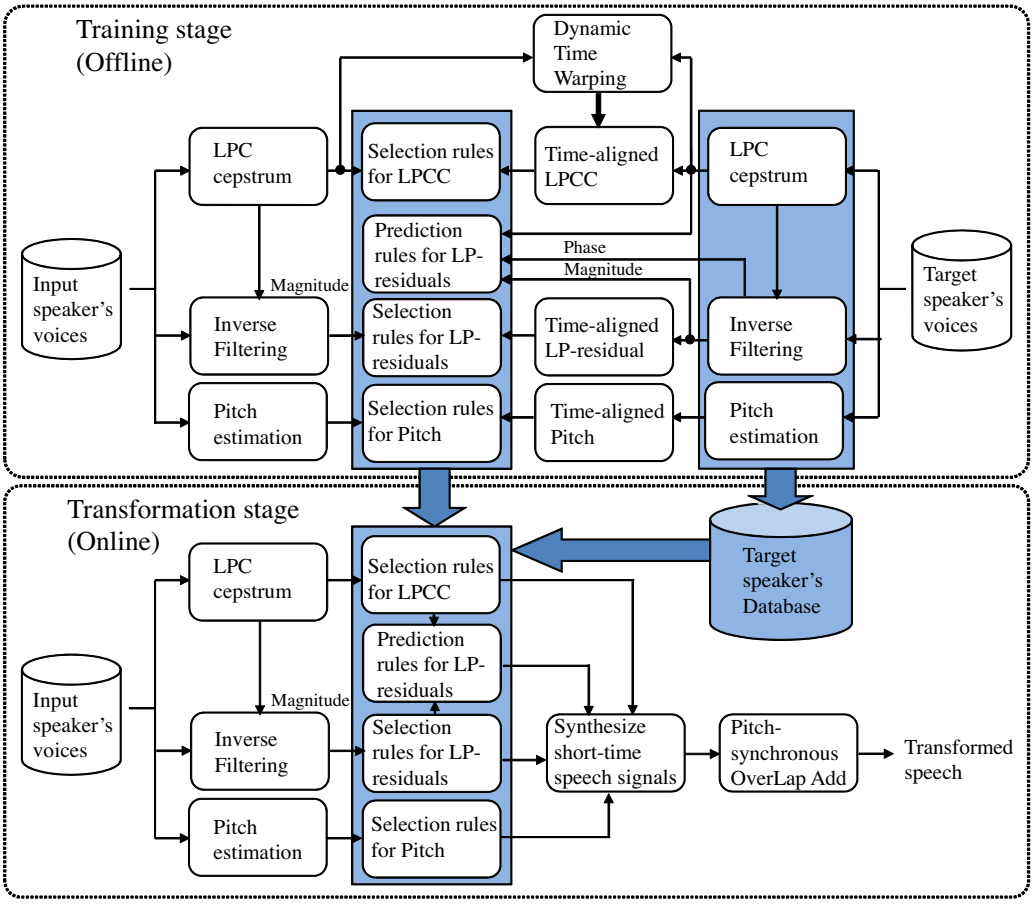

Fig. 1. Block diagram of the proposed voice transformation system.

(1) The amplitudes/phases of the harmonics were determined by the pitch synchronous sinusoidal model to represent the spectrum of the LP-residual. Note that the total number of the harmonics is determined by the fundamental frequency.
(2) Resample the spectrum so that all frames have the same number (in this study, 100) of the harmonics. The cubic spline interpolation and nearest neighbor interpolation are adopted to resample the magnitude and phase spectra, respectively.
(3) Principle component analysis (PCA) is carried out on the resampled magnitude spectrum. The number of the principle components was determined so that the cumulative relative dispersion of eigenvalue was more than 0.95. The resultant number of the principle components typically ranges from 50 to 80.

In the online stage, the features extracted during the training stage were derived from the incoming speech signals. The features were then replaced with those selected from a target database using the conversion rules constructed during the training stage. The LPRP were predicted using the estimated target LPCC, LPRM and pitch. The short-time speech signals were synthesized from the estimated parameters. Finally, continuous waveforms were obtained by concatenating the short-time speech signals. This procedure used the Pitch synchronous Over-Lap and Add (PSOLA) (Moulines and Charpentier, 1990) algorithm to align each short-time speech signal. Note that the interval between the two neighboring frames is given by the modified pitch period.

Each part of the proposed system is described in greater detail in the following sections.

3. Transformation rules

3.1. Overall transformation rules

In this work, transformation is performed on a sequence of features during speaking spurts. Let $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ and $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{T}$ be the source and the target sequences, respectively, where the features of a sequence are assumed to be time-aligned. Note that $\mathbf{x}_t$ and $\mathbf{y}_t$ include all three features selected from transformation-LPCC, LPRM and pitch i.e.

$$\mathbf{x}_t = [\mathbf{x}_{l,t}\Delta\mathbf{x}_{l,t}\mathbf{x}_{r,t}x_{p,t}]^T$$

$$\mathbf{y}_t = [\mathbf{y}_{l,t}\Delta\mathbf{y}_{l,t}\mathbf{y}_{r,t}y_{p,t}]^T$$

where the terms "l", "r" and "p" denote the LPCC, the LPRM and the pitch, respectively. The LPRM is given by the principal components of the magnitude spectrum of the LPR, which is actually obtained by Karhunen–Loéve transformation (KLT).

In the present study, the optimal transformed sequence $\hat{\mathbf{Y}}^*$ for a given source sequence $\mathbf{X}$ is given by

$$\hat{\mathbf{Y}}^* = \arg\max_{\hat{\mathbf{Y}}\in S_Y} f_{\mathrm{Y|X,\Lambda}}(\mathbf{Y}|\mathbf{X},\Lambda_\mathrm{Y}) \tag{1}$$

where $f_{\mathrm{Y|X,\Lambda}}(\mathbf{Y}|\mathbf{X},\Lambda_\mathrm{Y})$ is the likelihood function of $\mathbf{Y}$ given $\mathbf{X}$ and Λ_Y. Λ_Y is a model that describes the target features, which are represented in the context of the HMM. In (1) S_Y is a set of features obtained from the target speaker's utterances that were recorded in the training stage. The transformation rule in this work indicates that a transformed sequence for a given source sequence $\mathbf{X}$ is composed of the selected features from a target database, where the likelihood of the selected features is maximized with respect to both the given source sequence $\mathbf{X}$ and the target model. The objective function in (1) can be written as follows:

$$\begin{aligned} f_{\mathrm{Y|X,\Lambda}}(\mathbf{Y}|\mathbf{X},\Lambda_\mathrm{Y}) &= \frac{f_{\mathrm{X,Y}}(\mathbf{X},\mathbf{Y})}{f_\mathrm{X}(\mathbf{X})f_\mathrm{Y}(\mathbf{Y})} f_{\mathrm{Y|\Lambda_Y}}(\mathbf{Y}|\Lambda_\mathrm{Y}) \\ &= \rho(\mathbf{X},\mathbf{Y}) f_{\mathrm{Y|\Lambda_Y}}(\mathbf{Y}|\Lambda_\mathrm{Y}) \end{aligned} \tag{2}$$

where $\rho(\mathbf{X},\mathbf{Y})$ is the joint PDF between $\mathbf{X}$ and $\mathbf{Y}$. Note that the two functions $\rho(\mathbf{X},\mathbf{Y})$ and $f_{\mathrm{Y|\Lambda_Y}}(\mathbf{Y}|\Lambda_\mathrm{Y})$ in proposed transformation rule (2) are associated with the *inter*-and *intra*-speaker models, respectively. A more detailed description of each model is explained in the following subsection.

3.2. Inter-speaker model

The model proposed in our previous study (Lee, 2007), in which inter-speaker variability was described by an inter probabilistic model, was used in the present study. In this model, it is assumed that the probability of random source β_j contributing to the target feature $\mathbf{y}_t$, while random source α_i contributes to the source feature $\mathbf{x}_t$, is defined by a cross-correlational probability as follows:

$$f_{x,y,\alpha,\beta}(\mathbf{x}_t,\mathbf{y}_t,\alpha_i,\beta_j) = f_{\mathrm{x|\alpha}}(\mathbf{x}_t|\alpha_i) f_{\beta|\alpha}(\beta_j|\alpha_i) f_{y|\beta}(\mathbf{y}_t|\beta_j) f_\alpha(\alpha_i) \tag{3}$$

where $f_{\beta|\alpha}(\beta_j|\alpha_i)$ is the cross correlation probability between the ith random source of the source feature and the jth random source of the target feature (Cheng et al., 1994). This term describes the dependencies of the two random vector sets. Because the random sources $\{\alpha_i,\beta_j\}$ are assumed to be Gaussian,

$$\begin{aligned} f_{\mathrm{x|\alpha}}(\mathbf{x}_t|\alpha_i) &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_{x,i}|^{1/2}} \\ &\quad\times \exp\left\{-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_{x,i})'\mathbf{\Sigma}_{x,i}^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_{x,i})\right\} \end{aligned} \tag{4}$$

$$\begin{aligned} f_{\mathrm{y|\beta}}(\mathbf{y}_t|\beta_j) &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_{y,j}|^{1/2}} \\ &\quad\times \exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_{y,j})'\mathbf{\Sigma}_{y,j}^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_{y,j})\right\} \end{aligned} \tag{5}$$

where $\mathbf{\Sigma}_{x,i}$ and $\boldsymbol{\mu}_{x,i}$ are the covariance matrix and mean vector of the ith random source for the source feature, respectively. Similarly, $\mathbf{\Sigma}_{y,j}$ and $\boldsymbol{\mu}_{y,j}$ are the covariance matrix and

mean vector, respectively, of the jth random source for target feature. D is the order of the features. The method for estimating $f_{\beta|\alpha}(\beta_j|\alpha_i)$, $f_{\alpha}(\alpha_i)$ and parameters describing $f_{x|\alpha}(\mathbf{x}_t|\alpha_i)$ and $f_{y|\beta}(\mathbf{y}_t|\beta_j)$ from a given training corpus is based on a maximum likelihood criterion, as described in Lee (2007). Using the adopted inter-speaker model, the cross-correlation PDF $\rho(\mathbf{X},\mathbf{Y})$ is given by

$$\rho(\mathbf{X},\mathbf{Y}) = \prod_{t=1}^{T} \varrho(\mathbf{x}_t,\mathbf{y}_t) \tag{6}$$

where

$$\varrho(\mathbf{x}_t,\mathbf{y}_t) = \frac{f_{\mathrm{x,y}}(\mathbf{x}_t,\mathbf{y}_t)}{f_{\mathrm{x}}(\mathbf{x}_t) f_{\mathrm{y}}(\mathbf{y}_t)} \tag{7}$$

and

$$f_{\mathrm{x,y}}(\mathbf{x}_t,\mathbf{y}_t) = \sum_i \sum_j f_{\mathrm{x,y},\alpha,\beta}(\mathbf{x}_t,\mathbf{y}_t,\alpha_i,\beta_j),$$

$$f_{\mathrm{x}}(\mathbf{x}_t) = \sum_i f_{\mathrm{x}|\alpha}(\mathbf{x}_t|\alpha_i) f_{\alpha}(\alpha_i),$$

$$f_{\mathrm{y}}(\mathbf{y}_t) = \sum_j \left[f_{\mathrm{y}|\beta}(\mathbf{y}_t|\beta_j) \sum_i f_{\beta|\alpha}(\beta_j|\alpha_i) f_{\alpha}(\alpha_i) \right]$$

Note that observations of both source and target features are independent in different time frames t. We assumed that the cross-correlation PDFs for each type of feature are also independent. Hence the cross-correlation PDF at time t, $\varrho(\mathbf{x}_t,\mathbf{y}_t)$ is given by

$$\varrho(\mathbf{x}_t,\mathbf{y}_t) = \varrho_l(\mathbf{x}_{l,t},\mathbf{y}_{l,t}) \varrho_{\Delta l}(\Delta\mathbf{x}_{l,t},\Delta\mathbf{y}_{l,t}) \varrho_r(\mathbf{x}_{r,t},\mathbf{y}_{r,t}) \varrho_p(x_{p,t},y_{p,t}) \tag{8}$$

where $\varrho_l(\mathbf{x}_{l,t},\mathbf{y}_{l,t})$, $\varrho_l(\Delta\mathbf{x}_{l,t},\Delta\mathbf{y}_{l,t})$, $\varrho_r(\mathbf{x}_{r,t},\mathbf{y}_{r,t})$ and $\varrho_p(x_{p,t},\mathbf{y}_{p,t})$ are the cross correlation PDFs for LPCC, ΔLPCC, LPRM, and pitch, respectively.

3.3. Intra-speaker model

As noted above, the target model Λ_{Y} is represented in the HMM context. Hence, the target model includes the following HMM parameters.

$$\Lambda_{\mathrm{Y}} = \{\mathbf{A}_{\mathrm{Y}},\mathbf{B}_{\mathrm{Y}},\boldsymbol{\pi}_{\mathrm{Y}}\} = \{a_{ij},b_i,\pi_i,1 \leqslant i,j \leqslant N_s\} \tag{9}$$

where a_{ij} is the transient PDF from states i and the state j, b_i is the state observation PDF for state-i and π_i is the initial PDF of state-i. N_s is the number of states. In this work, we focused on representation of the state observation PDF b_i, which models the relationship between features.

A model of state observation density was previously proposed when multi-channel observation sequences were given (Summerfield, 1992; Manabe and Zhang, 2004). This model was primarily used to represent the relationship between multi-channel observations. In the present study, this model was adopted for representation of inter-feature relationships. There are several methods available for integration of individual features to represent the relationships among them. The models can be categorized as either early integration (EI) or late integration (Li) models (Summerfield, 1992). In the EI model, integration is performed in the feature space to form a composite feature vector that represents multiple features of each channel. Hence, the state observation density is given by the probability of this composite feature vector. In practice, the order of the composite feature vector resulted from the EI model can be large, causing inadequate modeling in (9) due to the curse of dimensionality and insufficient data. This problem can be alleviated by applying an linear discriminant analysis (LDA) projection on the composite vector (Potamianos et al., 2003). In the Li model, a density function is defined for each feature, and the state observation density is obtained by integrating individual density functions. This paper focuses on the Li model.

A simple way of implementing the Li model is based on the assumption that all of the individual density functions are statistically independent. In this case, the state observation density is given by

$$b(\mathbf{y}_t) = f_L(\mathbf{y}_{L,t}) f_r(\mathbf{y}_{r,t}) f_p(y_{p,t}) \tag{10}$$

where f_L, f_r and f_p denote the density functions for LPCC, LPRM and pitch, respectively. In (10), the state index i is omitted for simplicity. Note that $\mathbf{y}_{L,t} = [\mathbf{y}_{l,t}\Delta\mathbf{y}_{l,t}]^T$, i.e., a composite LPCC vector that includes the ΔLPCC vector. When the Gaussian mixture model is adopted, an individual density function is given by

$$\begin{aligned} f_L(\mathbf{y}_{L,t}) &= \sum_{i=1}^{N_L} f_{\lambda_L}(\lambda_{L,i}) f_{\mathrm{y}|\lambda_L}(\mathbf{y}_{L,t}|\lambda_{L,i}) \\ f_r(\mathbf{y}_{r,t}) &= \sum_{i=1}^{N_r} f_{\lambda_r}(\lambda_{r,i}) f_{\mathrm{y}|\lambda_r}(\mathbf{y}_{r,t}|\lambda_{r,i}) \\ f_p(y_{p,t}) &= \sum_{i=1}^{N_p} f_{\lambda_p}(\lambda_{p,i}) f_{\mathrm{y}|\lambda_p}(y_{p,t}|\lambda_{p,i}) \end{aligned} \tag{11}$$

where N_L, N_r and N_p are the number of Gaussian components for LPCC, LPRM and pitch, respectively, and $f_{\lambda_L}(\lambda_{L,i})$, $f_{\lambda_r}(\lambda_{r,i})$ and $f_{\lambda_p}(\lambda_{p,i})$ are the mixture weights of each feature of the ith Gaussian component. $f_{\mathrm{y}|\lambda_L}(\mathbf{y}_{L,t}|\lambda_{L,i})$, $f_{\mathrm{y}|\lambda_r}(\mathbf{y}_{r,t}|\lambda_{r,i})$ and $f_{\mathrm{y}|\lambda_p}(y_{p,t}|\lambda_{p,i})$ are the ith Gaussian component for each feature. Using (11), the state observation density is given by

$$b(\mathbf{y}_t) = \sum_{i=1}^{N_L} \sum_{j=1}^{N_r} \sum_{k=1}^{N_p} f_{\mathbf{y},\boldsymbol{\lambda}_y}(\mathbf{y}_{L,t},\mathbf{y}_{r,t},y_{p,t},\lambda_{L,i},\lambda_{r,j},\lambda_{p,k}) \tag{12}$$

where $f_{\mathbf{y},\boldsymbol{\lambda}_y}(\mathbf{y}_{L,t},\mathbf{y}_{r,t},y_{p,t},\lambda_{L,i},\lambda_{r,j},\lambda_{p,k})$ is the joint probability function of the set of the observation features $\mathbf{y}_{L,t}$, $\mathbf{y}_{r,t}$ and $y_{p,t}$ and the set of the Gaussian random sources $\lambda_{L,i}$, $\lambda_{r,j}$ and $\lambda_{p,k}$, which is given by

$$\begin{aligned} &f_{\mathbf{y},\boldsymbol{\lambda}_y}(\mathbf{y}_{L,t},\mathbf{y}_{r,t},y_{p,t},\lambda_{L,i},\lambda_{r,j},\lambda_{p,k}) \\ &\quad = f_{\lambda_L}(\lambda_{L,i}) f_{\mathrm{y}|\lambda_L}(\mathbf{y}_{L,t}|\lambda_{L,i}) \times f_{\lambda_r}(\lambda_{r,j}) f_{\mathrm{y}|\lambda_r}(\mathbf{y}_{r,t}|\lambda_{r,j}) \\ &\qquad \times f_{\lambda_p}(\lambda_{p,k}) f_{\mathrm{y}|\lambda_p}(y_{p,t}|\lambda_{p,k}) \end{aligned} \tag{13}$$

This model does not account for inter-feature dependence. We defined a state observation density function that reflects inter-feature dependence as follows.

$$\begin{aligned} & f_{\mathbf{y},\boldsymbol{\lambda}_y,\theta}(\mathbf{y}_{L,t},\mathbf{y}_{r,t},y_{p,t},\lambda_{L,i},\lambda_{r,j},\lambda_{p,k},\theta_m) \\ &= f_{\lambda_L|\theta}(\lambda_{L,i}|\theta_m) f_{\mathrm{y}|\lambda_L}(\mathbf{y}_{L,t}|\lambda_{L,i}) \\ &\quad \times f_{\lambda_r|\theta}(\lambda_{r,j}|\theta_m) f_{\mathrm{y}|\lambda_r}(\mathbf{y}_{r,t}|\lambda_{r,j}) \\ &\quad \times f_{\lambda_p|\theta}(\lambda_{p,k}|\theta_m) f_{\mathrm{y}|\lambda_p}(y_{p,t}|\lambda_{p,k}) \times f_\theta(\theta_m) \end{aligned} \tag{14}$$

where

$$1 \leqslant i \leqslant N_L,\ 1 \leqslant j \leqslant N_r,\ 1 \leqslant k \leqslant N_p,\ 1 \leqslant m \leqslant N_G$$

The mixture weights $f_{\lambda_L}(\lambda_{L,i})$, $f_{\lambda_r}(\lambda_{r,j})$ and $f_{\lambda_p}(\lambda_{p,k})$ of the independent model (13) are replaced with the conditional probabilities $f_{\lambda_L|\theta}(\lambda_{L,i}|\theta_m), f_{\lambda_r|\theta}(\lambda_{r,j}|\theta_m)$ and $f_{\lambda_p|\theta}(\lambda_{p,k}|\theta_m)$, respectively. Therefore, the occurrence of each observation feature is controlled, in part, by the random source θ_m, as shown in Fig. 2. Because the random source typically contributes to each mixture weight, we referred to it as the global control variable (GCV). In other words, relationships between the multi-feature observations are described by use of the correlation probabilities with the common (or shared) random variables. Note that N_G is the number of GCVs in (14).

The state observation density involved with GCV is given by

$$b(\mathbf{y}_t) = \sum_{i=1}^{N_L}\sum_{j=1}^{N_r}\sum_{k=1}^{N_p}\sum_{m=1}^{N_G} f_{\mathbf{y},\boldsymbol{\lambda}_y,\theta}(\mathbf{y}_{L,t},\mathbf{y}_{r,t},y_{p,t},\lambda_{L,i},\lambda_{r,j},\lambda_{p,k},\theta_m) \tag{15}$$

Note that the state observation densities (12) and (15) are identical when $f_{\lambda_L,\theta}(\lambda_{L,i},\theta_m) = f_{\lambda_L}(\lambda_{L_i}) f_\theta(\theta_m)$, $f_{\lambda_r,\theta}(\lambda_{r,j},\theta_m) = f_{\lambda_r}(\lambda_{r_j}) f_\theta(\theta_m)$ and $f_{\lambda_p,\theta}(\lambda_{p,k},\theta_m) = f_{\lambda_p}(\lambda_{p,k}) f_\theta(\theta_m)$. Thus, the state observation density in the absence of inter-feature dependence is a special case of the GCV model, where the Gaussian random source λ_{i_m} and GCV θ_k are statistically independent for $1 \leqslant i \leqslant N_L$, $1 \leqslant j \leqslant N_r$, $1 \leqslant k \leqslant N_p$ and $1 \leqslant m \leqslant N_G$. Hence, the GCV model yields a more generalized form of the state observation density function compared with the independent model. The method used to estimate model parameters for the HMM with the GCV described by Lee (2008) can be used to estimate the model parameters for the intra-speaker model.

3.4. Feature selection

For the approach proposed in the present study, the use of a large database is critical for the transformation of high-quality speech signal, because high-quality speech synthesis requires a sufficient variety of waveforms to cover various manifestations of each feature. In practice, most corpus-based TTS systems involve a large database that is constructed from a speech corpus that exceeds 1 h in length (Beutnagel et al., 1999). However, selection of the optimal features from a large database is not a trivial undertaking.

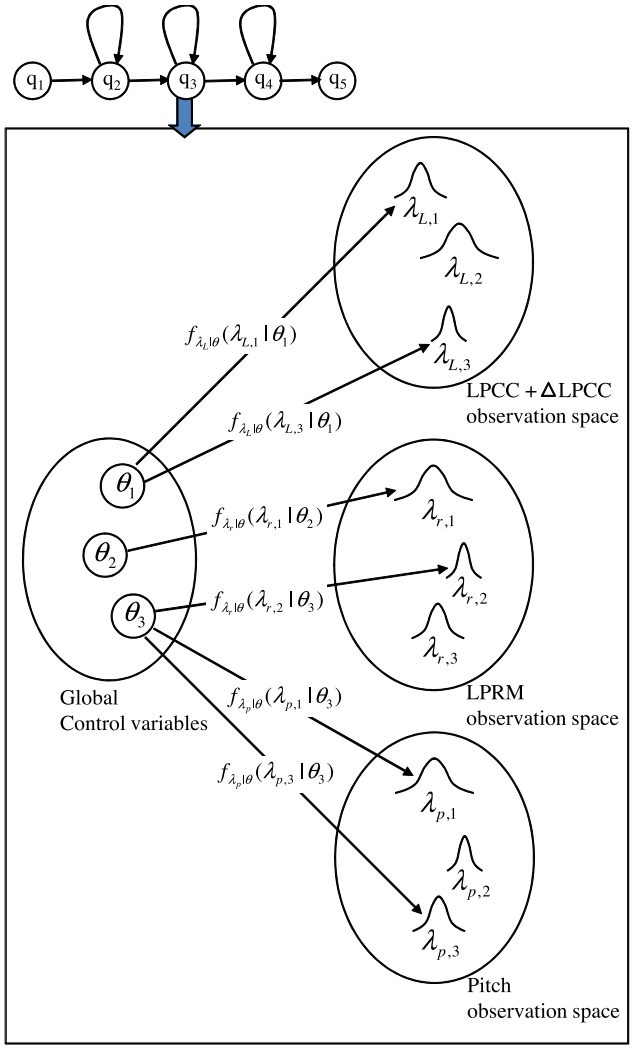

Fig. 2. Graphical depiction of the GCV model for representation of the state observation density where the state index $= 3$, $N_G = 3$, and $N_c = N_p = N_g = 3$).

As the size of the target database increases, it takes more times to select the optimal features. Hence, it would be highly desirable to determine the required number of candidates *a prior*, rather than computing the likelihood, given by (1) for all features in the database. A set of the candidates $S_C(\mathbf{x})$ was constructed individually for each feature using the predicted target feature $\hat{f}_{\mathbf{y}}(\mathbf{x})$ as follows.

$$\begin{aligned} S_{C,l}(\mathbf{x}_{l,t}) &= \{\mathbf{y}_l | \mathbf{y}_l \in S_{Y_l}, ||\mathbf{y}_l - \hat{f}_{\mathbf{y}_l}(\mathbf{x}_{l,t})||^2 \leqslant \epsilon_{l,th}\} \\ S_{C,r}(\mathbf{x}_{r,t}) &= \{\mathbf{y}_r | \mathbf{y}_r \in S_{Y_r}, ||\mathbf{y}_r - \hat{f}_{\mathbf{y}_r}(\mathbf{x}_{r,t})||^2 \leqslant \epsilon_{r,th}\} \\ S_{C,p}(x_{p,t}) &= \{y_p | y_p \in S_{Y_p}, ||y_p - \hat{f}_{y_p}(x_{p,t})||^2 \leqslant \epsilon_{p,th}\} \end{aligned} \tag{16}$$

where $\epsilon_{\{l,r,p\},th}$ are the thresholds for each feature, which can be adjusted so that the number of candidates is 50∼100. In the present study, the GMM-based transformation method (Stylianou et al., 1998) was employed to get the predicted target feature for LPCC and LPRM. For pitch, the predicted target value was obtained by scaling, so that the average of the scaled values is identical to that of the target values. It was observed that in case when the larger number

of candidates was adopted for pitch, the resulting synthesized speech had highly fluctuating pitch and an unnatural sound. Hence, a smaller number of candidates (i.e. $\leqslant 10$, even 1) for pitch was desirable to produce speech that sound more natural.

The optimal transformed sequence is constructed from the features selected from the set of candidates. For the source feature sequence $\mathbf{X} = \{\mathbf{x}\}_{t=1}^{T}$, the target model $\Lambda_Y = \{a_{ij}, b_i, \pi_i, 1 \leqslant i, j \leqslant N_s\}$ and the arbitrarily selected HMM state sequence $\Omega = \{\omega_t\}_{t=1}^{T}$, the log-likelihood of the target feature sequence $\mathbf{Y}\{\mathbf{y}\}_{t=1}^{T}$ is given by

$$\log\left[f_{Y|X,\Lambda,\Omega}(\mathbf{Y}|\mathbf{X},\Lambda_Y,\Omega)\right] = \sum_{t=1}^{T}\left[\rho'(\mathbf{x}_t,\mathbf{y}_t) + b'_{\omega_t}(\mathbf{y}_t) + C_{t-1,t}\right] \tag{17}$$

where

$$\rho'(\mathbf{x}_t,\mathbf{y}_t) = \log\rho(\mathbf{x}_t,\mathbf{y}_t)$$

$$b'_{\omega_t}(\mathbf{y}_t) = \log b_{\omega_t}(\mathbf{y}_t)$$

$$C_{t-1,t} = \begin{cases} \log\pi_{\omega_t} & \text{if } t = 1. \\ \log a_{\omega_{t-1},\omega_t} & \text{otherwise} \end{cases}$$

The optimal transformed sequence $\hat{\mathbf{Y}}^*$ is then given by

$$\hat{\mathbf{Y}}^* = \arg\max_{\mathbf{Y}\in\mathbf{S}_C}\left\{\max_{\Omega\in\Omega_T}\log\left[f_{Y|X,\Lambda,\Omega}(\mathbf{Y}|\mathbf{X},\Lambda_Y,\Omega)\right]\right\} \tag{18}$$

where $\mathbf{S}_C = \{S_{C,l}(\mathbf{x}_{l,t}) \cup S_{C,r}(\mathbf{x}_{r,t}) \cup S_{C,p}(x_{p,t})\}_{t=1}^{T}$ is the set of candidates for $1 \leqslant t \leqslant T$ and Ω_T denotes the set of all possible HMM state sequences.

Eq. (18) can be maximized using a dynamic programming technique, such a Viterbi-trellis search. In the trellis, each node corresponds to combination of HMM states, LPCC candidates, LPRM candidates, and pitch candidates. Hence, M_t, the number of nodes at time t is given by $N_s M_{l,t} M_{r,t} M_{p,t}$, where $M_{l,t}$, $M_{r,t}$, and $M_{p,t}$ are the number of candidates for LPCC, LPRM, and pitch at time t, respectively.

Let $\mathfrak{Q}_t(n)$ be the accumulated likelihood for the nth node at time t, the forward recursion is as follows:

$$\begin{aligned}\mathfrak{Q}_t(n) &= \max_{1\leqslant m\leqslant M_{t-1}}\{\mathfrak{Q}_{t-1}(m) + \log a_{\omega(m)\omega(n)} + b'(n,m)\} + \rho'(n) \\ \Psi_t(n) &= \arg\max_{1\leqslant m\leqslant M_{t-1}}\{\mathfrak{Q}_{t-1}(m) + \log a_{\omega(m)\omega(n)} + b'(n,m)\}\end{aligned} \tag{19}$$

where $1 \leqslant n \leqslant M_t$ and $\omega(m)$ is the HMM-state index of the mth node, $\Psi_t(n)$ is the back-tracking pointer for the nth node at time t. Let $\mathbf{y}_{l,t}^{(n_l)}$, $\mathbf{y}_{r,t}^{(n_r)}$, and $y_{p,t}^{(n_p)}$ are the n_lth LPCC candidate, n_rth LPRM candidate, and n_pth pitch candidate at time t, respectively, $b'(n,m)$ and $\rho'(n)$ are given by

$$\begin{aligned}b'(n,m) &= b'_{\omega(n)}([\mathbf{y}_{l,t}^{(n_l)}\Delta\mathbf{y}_{l,t}^{(n_l,m_l)}\mathbf{y}_{r,t}^{(n_r)}y_{p,t}^{(n_p)}]^T) \\ \rho'(n) &= \log\{\varrho_l(\mathbf{x}_{l,t},\mathbf{y}_{l,t}^{(n_l)})\varrho_{\Delta l}(\Delta\mathbf{x}_{l,t},\Delta\mathbf{y}_{l,t}^{(n_l,m_l)})\} \\ &\quad + \log\{\varrho_r(\mathbf{x}_{r,t},\mathbf{y}_{r,t}^{(n_r)})\varrho_p(x_{p,t},y^{(n_p)})_{p,t}\}\end{aligned} \tag{20}$$

where $1 \leqslant n_l \leqslant M_{l,t}, 1 \leqslant m_l \leqslant M_{l,t-1}$, $1 \leqslant n_r \leqslant M_{r,t}$, and $1 \leqslant n_p \leqslant M_{p,t}$, and $\Delta\mathbf{y}_{l,t}^{(n_l,m_l)} = \mathbf{y}_{l,t}^{(n_l)} - \mathbf{y}_{l,t-1}^{(m_l)}$, i.e., the Δ LPCC vector $\Delta\mathbf{y}_{l,t}^{(n_l,m_l)}$ is given by the difference between the m_lth candidate LPCC at time $t-1$ and the n_lth candidate LPCC at time t. Note that $S_{C,l}(\mathbf{x}_{l,t-1}) = \{\mathbf{y}_{l,t}^{(m_l)}\}_{m_l=1}^{M_{l,t-1}}$, $S_{C,l}(\mathbf{x}_{l,t}) = \{\mathbf{y}_{l,t}^{(n_l)}\}_{n_l=1}^{M_{l,t}}$, $S_{C,r}(\mathbf{x}_{r,t}) = \{\mathbf{y}_{r,t}^{(n_r)}\}_{n_r=1}^{M_{r,t}}$, and $S_{C,p}(x_{p,t}) = \{y_{p,t}^{(n_p)}\}_{n_p=1}^{M_{p,t}}$.

After the final accumulated likelihoods $\mathfrak{Q}_T(n)$ for all n have been computed, the best node sequence $\mathbf{Q}^* = q_1^*, q_2^*, \ldots, q_T^*$ is obtained using the following backward recursion:

$$\begin{aligned}q_T^* &= \arg\max_{1\leqslant n\leqslant M_t}\mathfrak{Q}_T(n) \\ q_t^* &= \Psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \ldots, 1.\end{aligned} \tag{21}$$

The optimal transformed sequence $\hat{\mathbf{Y}}^* = \{\mathbf{y}_t^{(q_t^*)}\}_{t=1}^{T}$ is constructed by selecting the candidate LPCC, LPRM, and pitch corresponding to the best node.

Although the number of the possible features is limited by candidate selection, the Viterbi decoding process still requires lots of computation. (e.g. the number of paths between $t-1$ and t is 12500×12500 when $N_s = 5$, $N_{l,t} = N_{l,t-1} = N_{r,t} = N_{r,t-1} = 50$, and $N_{p,t} = N_{p,t-1} = 1$). Therefore, pruning is highly desirable to reduce the computational time. In this study, the two-step pruning procedures are employed. In the first step, combinations of the three features with relatively low state observation density are removed. Our results showed that pruning of feature combinations with lower 50% state observation density did not seriously degraded the performance of the VT system. In the second step, the paths with low state transition probabilities (e.g. $a_{\omega(m)\omega(n)} \leqslant 0.1$) are removed. By employing the two-step pruning procedures, the number of paths can be reduced by about ten times.

3.5. LPRP prediction

It is not very easy to convert the magnitude and phase spectra of the LP-residual both via a single unified transformation (Ye and Young, 2006). As is well known, the spectral magnitude and phase of human speech are highly correlated. By using this property, several methods have been proposed to predict the LPRP using the estimated target magnitude spectrum (Ye and Young, 2006; Kain and Macon, 2001; Sundermann et al., 2005; Dutoit et al., 2007). In this study, an unit-selection framework was also adopted for prediction of the LPRP where the LPRP was selected from the training database. Since the selected residual sequence already contains less artifacts, this approach has an advantage of improving the quality of the converted speech (Sundermann et al., 2005).

The searched LPRP sequence $\tilde{\mathbf{\Phi}}_K = \{\tilde{\phi}_k\}_{k=1}^{K}$ for the given target feature sequence $\mathbf{Y}_K = \{\mathbf{y}_k\}_{k=1}^{K}$ is determined by minimizing the sum of the target and concatenation costs, as follows

$$\tilde{\Phi}_K = \arg\min_{\Phi_K}\left\{\sum_{k=1}^{K} C^T(\phi_k, \mathbf{y}_k) + \sum_{k=2}^{K} C^C(\phi_k, \phi_{k-1})\right\} \quad (22)$$

The target cost $C^T(\phi_k, \mathbf{y}_k)$ has a role in estimating the difference between the phase spectrum and that selected by means of feature vector $\mathbf{y}_k$. Note that since the real target feature sequence $\mathbf{Y}_K$ is not available in the transformation stage, the estimated target feature sequence $\hat{\mathbf{Y}}^*$ given by (18) is used in the target cost. The concatenation cost $C^C(\phi_k, \phi_{k-1})$ is given by the distance between the two unwrapped phase spectra ϕ_k, ϕ_{k-1}. There is a special condition for the concatenation cost. $C^C(\phi_k, \phi_{k-1})$ is defined to be zero, if ϕ_k and ϕ_{k-1} are consecutive in the database.

In this study, the target cost is given by the weighted sum of the LPCC distance, the LPRM distance and the pitch distance.

$$C^T(\phi_k, \mathbf{y}_k) = w_l||\mathbf{y}_{l,\phi_k} - \mathbf{y}_{l,k}||^2 + w_r||\mathbf{y}_{r,\phi_k} - \mathbf{y}_{r,k}||^2 + w_p||\mathbf{y}_{p,\phi_k} - \mathbf{y}_{p,k}||^2 \quad (23)$$

where w_l, w_r and w_p are the weights for the distances of LPCC, LPRM and pitch, respectively. $\mathbf{y}_{\phi_k}$ denotes the feature vector whose phase spectrum is given by ϕ_k. Determination of the weights is very crucial as reliable estimation of phase difference is more or less affected by the weight values. In this study, the weights are determined by minimizing overall square errors between the distortions of the unwrapped phase spectra and those of the feature vectors, given the training database.

$$w_l^*, w_r^*, w_p^* = \arg\min_{w_l, w_r, w_p} \sum_{n \neq m}\left\{||\phi_n - \phi_m||^2 - C^T(\phi_n, \mathbf{y}_{\phi_m})\right\}^2 \quad (24)$$

The experimental results showed that the LPCC weight w_l was relatively large and the pitch weight w_p was very small. Accordingly, the pitch distance was omitted in computing the target cost.

4. Experimental results

The speech data used in all experiments is from the CMU ARCTIC database (Kominek and Black, 2004) for US English and is sampled at 16 kHz. Two male speakers, bdl and rms, and two female speakers, clb and slt were used. The database used to obtain the conversion rules consisted of 200 utterances (from a0001 to a0200). An additional 100 utterances (from a0201 to a0300) spoken by the same individuals were prepared for both objective and subjective evaluation. The order of the LPCC coefficients was 20. The speech data was analyzed pitch-synchronously, at the manually labelled pitch marks. The clipped autocorrelation method (Rabiner and Schafer, 1978) was employed to obtain initial pitch marks. A variable length Hanning window was used to both compute and extract the LPC parameters. Each Gaussian component was constrained to a diagonal covariance matrix. Variance limiting (Reynolds and Rose, 1995) was also used to estimate each component of the covariance matrices. In our experiments, four different voice conversion tasks were investigated including male-to-male (rms → bdl), male-to-female (bdl → slt), female-to-female (slt → clb) and female-to-male (clb → rms) conversion.

4.1. Comparison of the independent feature model, the dependent feature model (GCV model)

Prior to evaluation of model performance in terms of transformation quality, we evaluated how well each model matched the real features of an individual speaker. The average log likelihoods $\sum_{\mathbf{Y}} \log f_{\mathbf{Y}|\Lambda_Y}(\mathbf{Y}|\Lambda_Y)$ of the given test data relative to the underlying models were computed and compared. Both the inter-feature (14) and independent (13) models were evaluated using this method. The average log likelihoods were computed from the features of the target speakers (rms, bdl, clb and slt). The lengths of the training corpus for the each speaker were 97789, 104662, 182381 and 155708 frames, respectively. Because the inter-feature model requires valid pitch values, model training was performed on only the voiced frames. An ergodic HMM was employed for both models. For the inter-feature model, the number of GCVs was set equal to the number of Gaussians. The number of states was set to five for both the independent and inter-feature models. In our experiment, the overall likelihood converged after 40–50 iterations for both models. For model evaluation, the HMM constructed from the training corpus was applied to the test corpus.

The results for the four speakers are shown in Fig. 3. In these figures, the average log likelihoods are shown for $2 \leqslant N_l,\ N_r,\ N_p \leqslant 5$. For all speakers, it was commonly observed that the average log likelihoods were increased as the number of Gaussians. The average log likelihoods of the inter-feature model were always higher than those of the independent model except when the number of Gaussians was three for speaker rms. The maximum of the differences between the two models was 1.11 in case when the number of Gaussians was five for speaker clb. There are two possible reasons for the higher average likelihood of the inter-feature model, compared with the independent model. First, more parameters were used to describe the inter-feature model than were employed for the independent model, because a number of GCVs were added to the parameters of the inter-feature model. Since GCVs are estimated by maximizing the overall likelihood, these additional parameters increase the overall likelihood of the test corpus as well as that of the training corpus.

Another reason for the higher likelihood of the inter-feature model is inter-dependence of the features. Hence, it can be expected that the naturalness of transformed speech is improved when the inter-feature model is adopted for voice transformation.

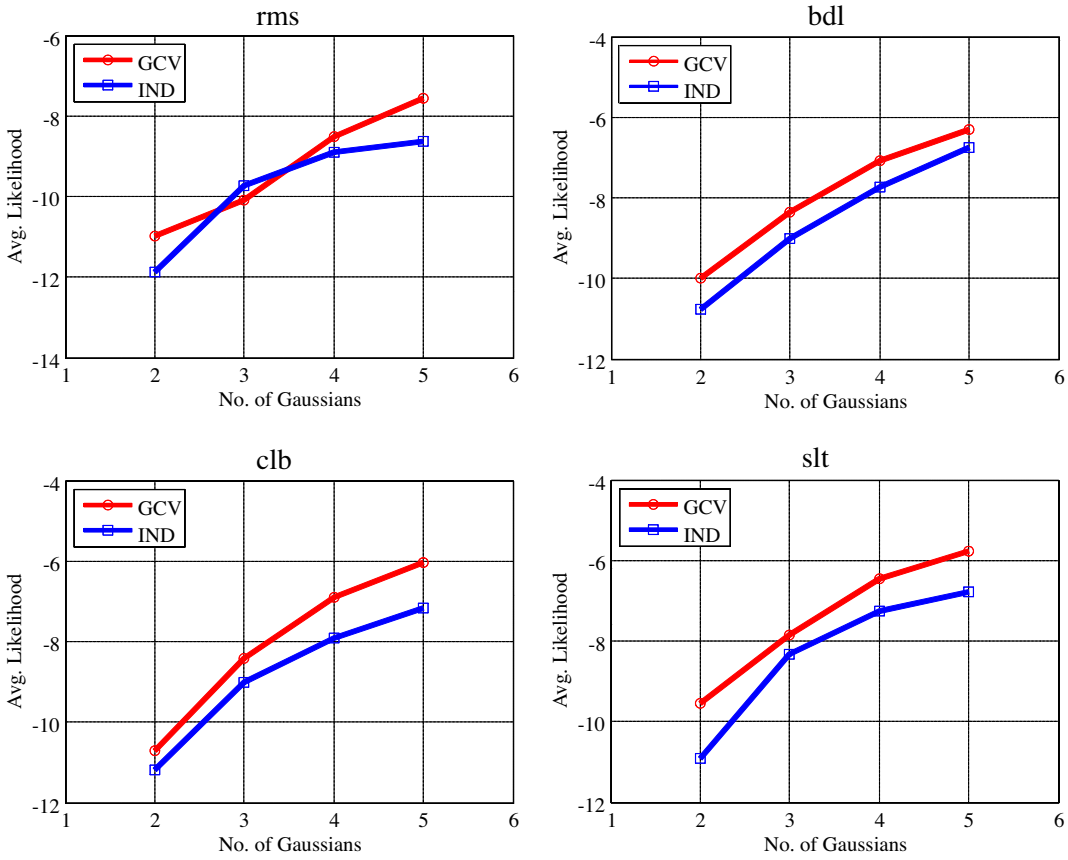

Fig. 3. Average log likelihood of each model, (squares: independent model, circles: inter-feature model (GCV model)) for the speakers rms (top-left), bdl (top-right), clb (bottom-left) and slt (bottom-right), according to the number of Gaussians.

4.2. Objective evaluation

To evaluate the performance of the proposed voice transformation method, two objective measurements were adopted. First, the following distance reduction ratio (Iwahashi and Sagisaka, 1995) was used

$$D_{ratio} = \left\{1 - \frac{D(\hat{\mathbf{Y}}, \mathbf{Y})}{D(\mathbf{X}, \mathbf{Y})}\right\} \times 100(\%) \tag{25}$$

where $\mathbf{X}, \mathbf{Y}$ and $\hat{\mathbf{Y}}$ are the feature sequences for the source speaker, the target speaker and the transformation, respectively, $D(\mathbf{X}, \mathbf{Y})$ denotes the averaged spectral distance between the magnitude spectra derived from the feature vectors $\mathbf{X}$ and $\mathbf{Y}$. A large reduction ratio indicates increased similarity between the transformed and target features.

For comparison, three types of conversion methods were adopted in this experiment; the GMM-based spectral parameter conversion method proposed by Toda et al. (2007), the unit-selection approach based on a maximum likelihood (ML) criterion without inter-feature model and the ML-based unit-selection approach with inter-feature model. These methods are referred to as "GMM-Traj", "UnitSel-IND", and "UnitSel-GCV", respectively.

For GMM-Traj, UnitSel-IND, the conversion rules for each feature (LPCC, LPRM and pitch) were constructed separately. Note that in these two methods, the features were modified independently, hence inter-feature dependence was not considered. For UnitSel-GCV, five states, Gaussians and GCVs were used in the inter-feature HMM model. The number of Gaussians was set to 64 for the GMM-Traj method, which was identical to that of the random sources for the unit selection methods (UnitSel-IND/GCV). The size of the database used in the unit selection methods ranged from 5 K to 40 K voiced units (the length of each voiced unit is two pitch periods). Fig. 4 presents the D_{ratio} obtained for the test corpus using the four methods. Except for the case of slt → clb conversion, the overall D_{ratio} of the UnitSel-GCV method was higher than those of the GMM-Traj method when the size of the database exceeded 20 K. In most cases, D_{ratio} has an increasing trend as the database size. It is interesting to note that when transformation was carried out on the male voices (rms → bdl and bdl → slt conversions), the UnitSel-GCV method with even very small size database (5 K) yielded higher D_{ratio} compared with the GMM-Traj method. Comparing the UnitSel-IND method and the UnitSel-GCV method, the UnitSel-GCV method yielded superior performance in terms of D_{ratio} to the UnitSel-

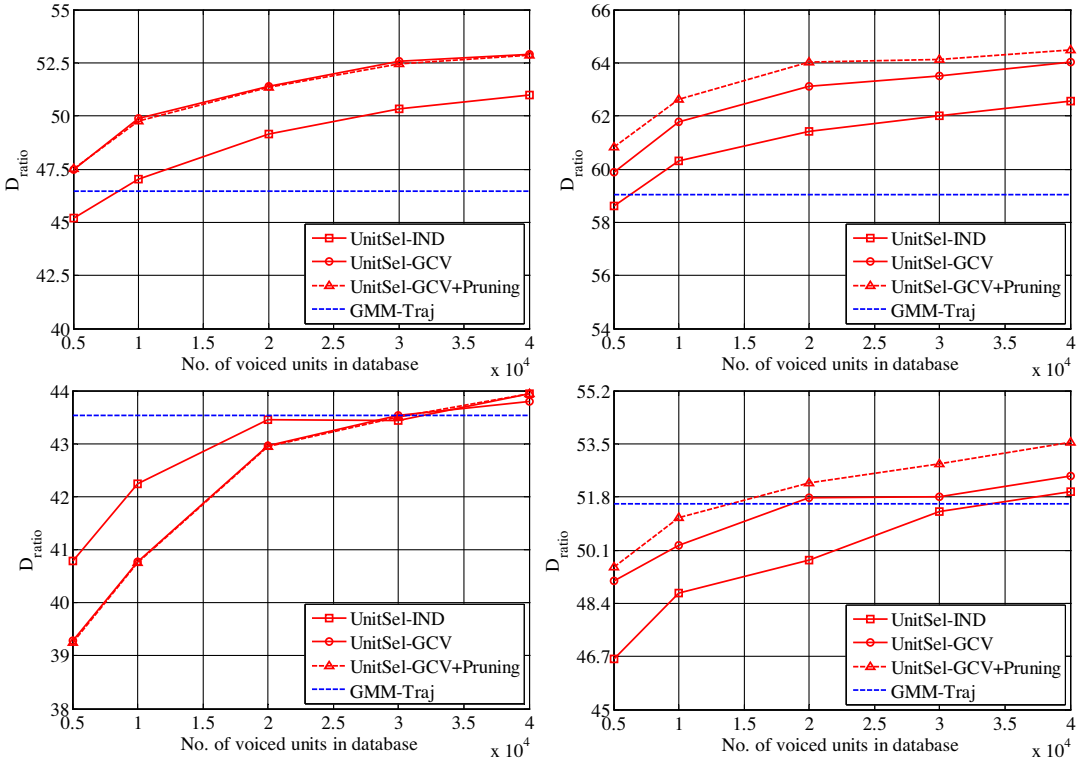

Fig. 4. Objective performance of the four voice conversion methods, (broken lines: GMM-Traj, solid-square lines: UnitSel-IND, solid-circle lines: UnitSel-GCV without pruning, and triangle-broken lines: UnitSel-GCV with pruning) for rms→bdl conversion (top-left), bdl→slt conversion (top-right), slt→clb conversion (bottom-left), clb→rms conversion (bottom-right). Note that since no database is necessary for GMM-based method, No. of the voiced units is valid only for the two unit selection methods. (Unit Sel-IND and Unit Sel-GCV).

IND method except for slt → clb conversion. For slt → clb conversion, D_{ratio} of the UnitSel-GCV method was higher than that of the UnitSel-IND method when the size of the database exceeded 30 K.

We also investigated the contribution of pruning in feature selection to voice transformation performance by comparing the overall D_{ratio} from the UnitSel-GCV methods with/without pruning, respectively. For same-gender conversions (rms → bdl and slt → clb), the differences in D_{ratio} between with and without pruning were not observed. Whereas, it appears that voice transformation performance was improved by employing feature selection with pruning in case of cross-gender conversions (bdl → slt and clb → rms).

To consider the oversmoothing problem of the converted spectra, the global variance (GV) of the converted spectra over a time was also investigated. Fig. 5 shows the GV of several LPCC sequences: the converted LPCC using the GMM-Traj method, that using the UnitSel-GCV method and the raw LPCC of the target speech. It can be seen that the GV of the low order LPCCs converted by the UnitSel-GCV method and the GMM-Traj method was very close to that of the raw LPCC of the target speech. For higher order LPCCs, the GV of the converted LPCC is slightly smaller than that of the raw LPCC. The frame-by-frame basis conversion (GMM) revealed smaller GV compared with other methods. It appears that the GV of the converted LPCC using the UnitSel-GCV method is similar to that of the GMM-Traj method. This results indicate that the UnitSel-GCV method keeps the GV of the converted LPCC to the same level as the GMM-Traj method.

Consequently, compared with the conventional method, the proposed unit selection VT method produced the spectral sequences which are closer to those of target, while the same degree of reduction of oversmoothing was achieved.

4.3. Subjective evaluation

In addition to the objective evaluation, two subjective listening tests were conducted. The first one was designed to evaluate the conversion of speaker individuality using the ABX test. For this test, 10 utterances were selected from 100 test utterances and each sentence was presented to 15 subjects. The first and second stimuli, A and B, were either the source speaker's utterances or the target speaker's utterances, while the last stimuli X was the transformed speech. Then, the subjects were asked to select either A or B as the original source of X. This test was performed in a silent room. The subjects were presented the stimuli via equalized headphones. Each listener was allowed to listen to the stimuli as many times as needed before determining which utterance (A or B) sounded more like the transformed utterance.

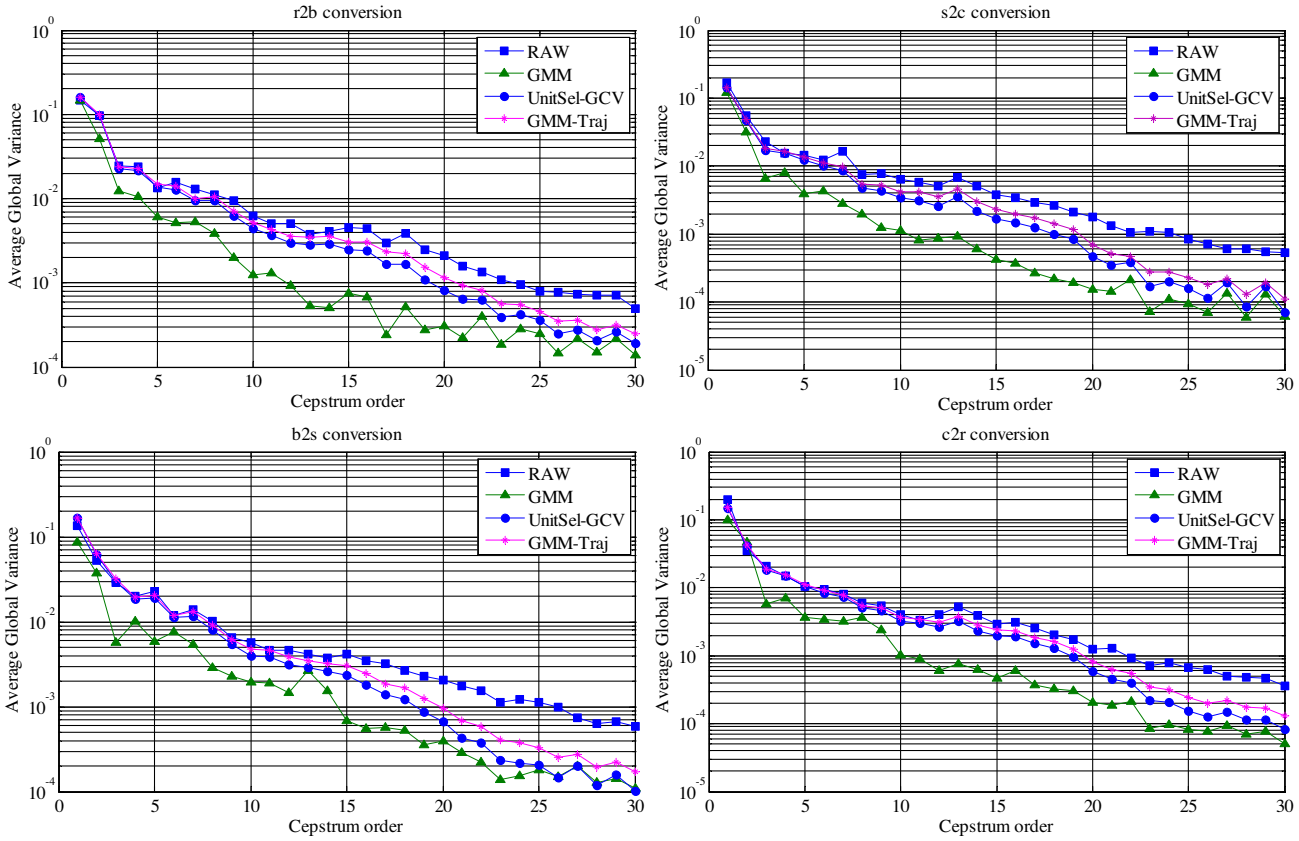

Fig. 5. Global variances (GVs) of the three voice conversion methods for rms→bdl conversion (top-left), slt→clb conversion (top-right), bdl→slt conversion (bottom-left), clb→rms conversion (bottom-right).

The source/target utterances presented in random order. We also compared the results from the three methods, GMM-Traj, UnitSel-IND, and UnitSel-GCV. For GMM-Traj, the number of Gaussians was set to 64, which was identical to that of random sources for UnitSel-IND and UnitSel-GCV. For the two unit selection-based VT schemes (UnitSel-IND and UnitSel-GCV) the number of the voiced units included in the database for each speaker was 20 K, and the number of HMM states was set to 5. The number of GCVs for UnitSel-GCV was set to 64. The same number of Gaussians was used to represent the state observation densities of the all the states.

Fig. 6 shows the correct identification ratios for the three different conversion methods. For all conversion pairs, the UnitSel-GCV method revealed higher correct identification ratios compared with the other methods. Considering that relatively high D_{ratio}s of the UnitSel-GCV method were often observed in the objective evaluation, this result is somewhat expected.

Although the objective performance of the UnitSel-GCV method in terms of D_{ratio} was inferior to that of the GMM-based approach in case of clb → rms conversion, the subjective performance of the U-GCV method was slightly higher than that of the GMM-based approach. In case when conversion was carried out on the male speech signals, the UnitSel-GCV method showed remarkably

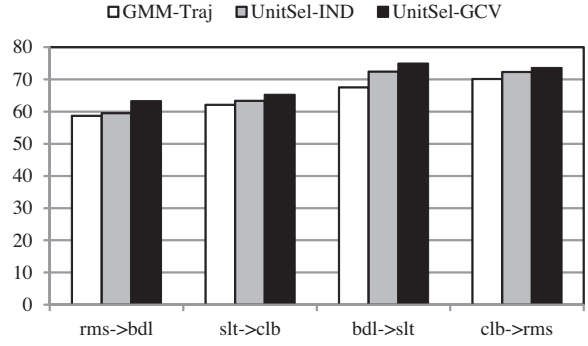

Fig. 6. ABX test results.

higher identification ratios even though the average D_{ratio} of the UnitSel-GCV method was not very higher than that of the UnitSel-IND method. For female speech conversion, the overall differences in correct identification ratio among the three methods were not as high as those in case of male speech conversion. However, the average correct identification ratios of the UnitSel-GCV method were higher than other methods in this case. The significant test (two-way ANOVA) showed that the conversion method was the major factor to affect correct identification ratio. ($p = 0.0045$) This result indicates that the utterances converted by the UnitSel-GCV method were perceptually closer to a target voice compared with conventional

GMM-based method and unit-selection-based conversion methods without inter-feature model. A possible explanation for this ABX test result is that for the UnitSel-GCV method, the speech signals are synthesized using the most likely combination of the features with respect to the target speaker's model. By contrast, for other methods, the transformed features were independently obtained.

For cross-gender conversions (bdl → slt and clb → rms), the subjects clearly perceived differences between the speakers, and even smaller modifications affected the perceived voice characteristics. The subjects tended to choose the target source when the transformed utterances sounded more or less different from the source speaker's utterances, regardless of the perceptual similarities with the target voices. As a result, the overall identification ratio for the cross-gender conversion was increased for all four methods. The highest correct identification ratio was obtained for bdl → slt conversion. This was confirmed by the fact that D_{ratio} of that conversion was remarkably higher than other conversions.

In the second test, the quality of the transformed speech signals was evaluated. In this test, subjects were asked to indicate which was more preferred. The utterances used in the first test were also used in the second test. As presented in Fig. 7, the UnitSel-GCV method performed remarkably better than the other methods in case of same-gender conversions (rms → bdl and slt → clb). According to the significant test result, the p-value of the conversion method factor is 0.0078. Whereas the p-value of 1.0 was observed when the factor was given by conversion pair. This indicates that the major factor to affect the quality of the transformed speech signals is the conversion method.

In addition to the preference test, a mean opinion score (MOS) test was conducted in which the same subjects participated in the ABX test listened to two randomly selected converted utterances per method and conversion pair and compared them with a reference target utterance. The subjects were asked to rate both the similarity between the converted and target voices and the quality of the converted utterance on a 1-to-5 scale. As usual, '5' is the best possible score in both performance dimension.

Fig. 8 shows the MOSs achieved by the three methods in separate conversion pair and also the global average MOS.

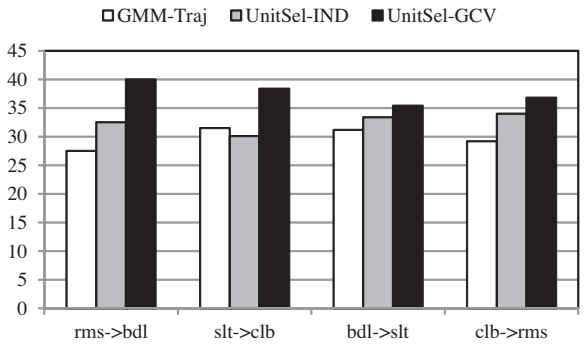

Fig. 7. Preference scores of each method.

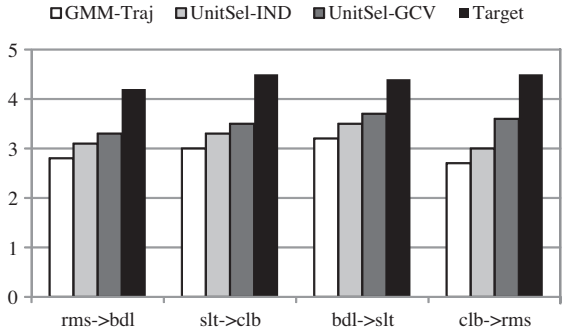

Fig. 8. Results of most opinion score (MOS) test.

For comparison, the average MOS for raw target voices is also presented. The result is close to that of the preference test in which the UnitSel-GCV method reveals higher opinion scores compared with other methods. The significant test showed that the p-value of the conversion method was almost zero (5.46×10^{-7}) and that of the conversion pair was about 0.026. This indicated that the conversion method was the major factor to affect MOS.

The subjects indicated that the clarify of the speech signals synthesized using the unit-selection approach was superior to that of the GMM approach. This difference in quality most likely resulted from bandwidth widening problem caused by the averaging of the transformed speech signals in the GMM-based method. Compared with the unit selection methods without inter feature model, the superiority of the UnitSel-GCV method in sound quality results from the selection process, in which the three types of features are jointly selected to maximize the likelihood functions involved with GCVs. Minimization of inappropriate combinations of LPCC, LP-residual magnitude spectrum and pitch increases the natural quality of the transformed speech signals. Moreover, pop and click sounds, which were sometimes perceived in the speech signals transformed using the unit-selection methods were significantly reduced by the proposed unit selection methods. These annoying artifacts might result from abrupt changes in the sequence of the selected feature and could be reduced by adopting the state transition probability in unit selection. ($C_{t,t-1}$ in (17)). However, correlation between the neighboring features was not sufficiently explained only by the state transition probability. In the proposed method, time difference between the neighboring LPCC vectors is taken into consideration in unit selection. The sequence of LPCC vectors are selected to maximize the overall likelihoods of both static features and dynamic feature (ΔLPCC) with respect to the target speaker's HMM. This means that the dynamic characteristics of the selected LPCC sequence is statistically similar to those of the target speaker. Hence, inappropriate combinations of neighboring LPCC vectors, which will potentially produce the pop and click noises, are rarely selected. This significantly reduced discontinuities at concatenation boundaries and improved the overall quality of the transformed speech signals.

4.4. Practical issue

Compared with the baseline VT method proposed by Toda et al. (2007), there are several practical issues associated with implementation of the proposed VT scheme. In terms of computational complexity, the proposed VT method is dependent on several factors including the size of database and the number of the candidates, whereas in the baseline VT, the computational complexity is determined by the number of feature vectors in all frames over a time sequence. In the baseline VT method, the GV was calculated utterance by utterance. This means that one time sequence corresponds to one utterance. When VT is applied to the relatively long utterance, the number of the features in one utterance becomes large. This leads to increasing the size of the matrices used in transformation. In this case, the required memory size and the computation times would be increased. As for the proposed VT method, feature selection is carried out on a voiced stream in which the consecutive frames are all voiced. Normally, the length of one voiced segments is much shorter than that of utterances. Hence, the required memory size and computations are not as much as those of the baseline VT method. In the proposed method, however, the feature selection process requires lots of computation. To alleviate this problem, the efficient pruning method was employed to reduce the computation cost while maintaining reasonable voice transformation performance in terms of both objective and subjective evaluations. As a result, the overall computation times became comparable to those of the baseline VT.

In the proposed VT method, the overall required memory size was highly affected by the size of the database. Larger size is desirable for high quality speech, since wide varieties of feature combinations can be obtained by the large database. Whereas, in the baseline method, the overall required memory size was determined by the possible number of the frames within the utterances. Under these conditions, we compared the required memory size for each method. According to the results, performance of the feature selection method was comparable with the baseline method, when the size of the database is larger than 20 K voice units. The results showed that the required memory size of the feature selection method with database including 20 K voiced units is smaller than baseline method for all transformation pairs. Even for database including 40 K voiced units, the required memory size of the baseline VT method was still much larger than that of the proposed method when source voice was given by female voice. Consequently, the proposed method has practical advantages over the baseline VT method.

5. Conclusions

A new voice transformation algorithm that is based on feature selection was proposed. The sequences of the transformed features were constructed by selection of the appropriate features from the target speaker's database. During the feature selection process, two probability models were taken into consideration: the inter- and intra-speaker models. For the inter-speaker model, the source/target features were controlled by the random sources shared between the two speakers. The intra-speaker model was subdivided into two models: the inter- and intra-feature models. Relationships between the features were explained by the inter-feature model, in which the occurrence of each feature was controlled by the shared random variables, i.e., the global control variables. The underlying assumption is that preservation of the relationship between the features from a single speaker improves the natural quality of the synthesized speech signals.

Both objective and subjective tests were performed to evaluate the effectiveness of the proposed method. Sets of utterances from four speakers were used in the evaluation. The results of the objective test showed that the performance of the proposed method was mostly inferior to that of the conventional methods. Moreover, the results of the subjective evaluation showed that the proposed method performed better than the conventional methods, because the transformed features were given by the original target features and not by the modified source features. In addition, results of a likelihood test showed that higher likelihood scores were obtained for the proposed model compared with the independent feature model, indicating superior matching of features from real speech signals. It would be interesting to use the inter-feature model in other speech-related applications, such as automatic speech recognition, text-to-speech synthesis and speech coding.

References

Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., 1988. Voice conversion through vector quantization. In: Proc. ICASSP, pp. 565–568.

Arslan, L.M., 1999. Speaker transformation algorithm using segmental codebooks (STASC). Speech Commun. 28, 211–226.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A., 1999. The AT&T next-gen TTS system. In: Proc. Joint Meeting of ASA, EAA, and DAGA.

Bi, N., Qi, Y., 1997. Application of speech conversion to alaryngeal speech enhancement. IEEE Trans. Acoust. Speech Signal Process. 5 (2), 97–105.

Cheng, Y.M., O'Shaughnessy, P., Mermelstein, P., 1994. Statistical recovery of wideband speech from narrowband speech. IEEE Trans. Speech Audio Process. 2 (4), 544–548.

Childers, D.G., Wong, C.F., 1994. Measuring and modeling vocal source-tract interaction. IEEE Trans. Biomed. Eng. 41 (7), 663–671.

Childers, D.G., Yegnanarayana, B., Wu, K., 1985. Voice conversion: factors responsible for quality. In: Proc. ICASSP, pp. 748–751.

Cox, S.J., Bridle, J.S., 1989. Unsupervised speaker adaptation by probabilistic spectrum fitting. In: Proc. ICASSP, pp. 294–297.

Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., Stylianou, Y., 2007. Towards a voice conversion system based on frame selection. In: Proc. ICASSP, pp. 15–20.

Erickson, M.L., 2003. The interaction of formant frequency and pitch in the perception of voice category and jaw opening in female singers. In: The 31st Annual Symposium: Care of the Professional Voice, pp. 24–37.

Erro, D., Navas, E., Hernáez, I., 2013. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio Speech Lang. Process. 21 (3), 556–566.

Helander, E., Silén, H., Virtanen, T., 2012. Voice conversion using dynamic kernel partial least squares regression. IEEE Trans. Audio Speech Lang. Process. 20 (3), 806–817.
Huang, Y.C., Wu, C.H., Chao, Y.T., 2013. Personalized spectral and prosody conversion using frame-based codeword distribution and adaptive CRF. IEEE Trans. Audio Speech Lang. Process. 21 (1), 51–52.
Iwahashi, N., Sagisaka, Y., 1995. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. Speech Commun. 16 (2), 139–152.
Jian, Z.H., Zhen, Y., 2007. Voice conversion using Viterbi algorithm based on Gaussian mixture model. In: Proc. Intelligent Signal Processing and Communication Systems, pp. 32–35.
Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis. In: Proc. ICASSP, Seattle, pp. 285–288.
Kain, A., Macon, M.W., 2001. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In: Proc. ICASSP, pp. 813–816.
Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27 (3-4), 187–207.
Kominek, J., Black, A.W., 2004. The CMU ARCTIC speech databases. In: Proc. Fifth ISCA Speech Synthesis Workshop, pp. 223–224.
Lee, K.S., 2007. Statistical approach for voice personality transformation. IEEE Trans. Audio Speech Lang. Process. 15 (2), 641–651.
Lee, K.S., 2008. EMG-based speech recognition using hidden Markov models with global control variables. IEEE Trans. Biomed. Eng. 55 (3), 930–940.
Lee, K.S., Youn, D.H., Cha, I.W., 1996. A new voice personality transformation based on both linear and nonlinear prediction analysis. In: Proc. ICSLP, pp. 1401–1404.
Lee, K.S., Youn, D.H., Cha, I.W., 2002. Voice conversion using a low dimensional vector mapping. IEICE Trans. Inf. Syst. E85-D (8), 1297–1305.
Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. 28, 84–95.
Ma, J., Liu, W., 2005. Voice conversion based on joint pitch and spectral transformation with component group-GMM. In: Proc. IEEE NLP-KE, pp. 199–203.
Manabe, H., Zhang, A., 2004. Multi-stream HMM for EMG-based speech recognition. In: Proc. the 26th Annual International Conference of the IEEE EMBS, pp. 4389–4392.
Mizuno, H., Abe, M., 1995. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectral tilt. Speech Commun. 16 (2), 153–164.
Moulines, E., Charpentier, F., 1990. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9 (5/6), 453–467.
Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B., 1995. Transformation of formants of voice conversion using artificial neural networks. Speech Commun. 16 (2), 207–216.
Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W., 2003. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE 91 (9), 1306–1326.
Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Englewood Cliffs.
Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signal, Englewood Cliffs.
Rao, K.S., 2010. Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput. Speech Lang. 24, 474–494.
Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Acoust. Speech Signal Process. 3 (1), 72–83.
Saito, D., Watanabe, S., Nakamura, A., Minematsu, N., 2012. Statistical voice conversion based on noisy channel model. IEEE Trans. Acoust. Speech Lang. Process. 20 (6), 1784–1794.
Savic, M., Nam, I.H., 1991. Voice personality transformation. Digital Signal Process. 4, 107–110.
Shuang, Z., Meng, F., Qin, Y., 2008. Voice conversion by combining frequency warping with unit selection. In: Proc. ICASSP, pp. 4661–4664.
Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Acoust. Speech Signal Process. 6 (2), 131–142.
Summerfield, A.Q., 1992. Lipreading and audio-visual speech perception. Philos. Trans. R. Soc. Lond. B 335, 71–78.
Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A.W., 2005. Residual prediction based on unit selection. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 369–374.
Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A.W., Narayanan, S., 2006. Text-independent voice conversion based on unit selection. In: Proc. ICASSP, pp. 14–19.
Toda, T., Black, A.W., Black, Tokuda, K., 2007. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Acoust. Speech Lang. Process. 15 (8), 2222–2235.
Valbret, H., Moulines, E., Tubach, J.P., 1992. Voice transformation using PSOLA technique. Speech Commun. 11, 175–187.
White, G.M., Neely, R.B., 1976. Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming. IEEE Trans. Acoust. Speech Signal Process. 24 (2), 183–188.
Ye, H., Young, S., 2006. Quality-enhanced voice morphing using maximum likelihood transformations. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1301–1312.