한국음향학회지 제31권 제1호 pp. 39~50 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.1 (2012)

Feature Selection-based Voice Transformation

Ki-Seung Lee

Department of Electronic Engineering, Konkuk University
(Received November 29, 2011; accepted December 23, 2011)

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 three feature parameters, which include the LPC cepstrum, pitch period and gain. 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) 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 informal listening tests. We confirmed that the proposed method leads to perceptually more preferred results, compared with the conventional methods.
Key words: Voice conversion, Unit selection, Hidden markov model.
ASK subject classification: Speech Signal Processing (2)

I. Introduction

Voice personality transformation [1-16] is a process by which voice personality is altered, so that one voice is made to sound like another. The process has numerous applications in a variety of areas such as personification of text-to-speech synthesis systems, preprocessing for speech recognition [17], and enhancing the intelligibility of abnormal speech [8].

Voice personality 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.

To implement voice personality transformation, the first problem is to determine which features should be extracted from the underlying speech signals and how to modify these features in a way so that the transformed speech signals mimic target speaker's voice. The vocal-tract transfer function (VTF) is a primary identifier of speaker individuality [18]. For this reason, feature parameters that represent the VTF including formant frequencies [4,5], the linear prediction coefficient cepsturm (LPCC) [2,10,11] and LSP (Line Spectrum Pair) coefficients [9], have been widely used in voice personality transformation. In the presented study, the LPCC was used as a feature

***Corresponding author:** Ki-Seung Lee
(kseung@konkuk.ac.kr)
Department of Electronic Engineering, Konkuk University,
1 Hwayang-dong, Gwangjin-gu, Seoul, 143-701, Korea
(Tel: 82-2-450-3489; Fax: 82-2-3437-5235)

parameter that represents the VTF. Prosody is another discriminator of speaker individuality [18]. Speaking style is highly correlated with prosody [2]. Hence, prosody modification is highly desirable for acquisition of transformed speech signals that are perceptually closer to a target voice. In the proposed method, prosody modification is accomplished by replacement of both the pitch and the gain.

The second problem can be described as finding acceptable mapping rules from the source speaker's feature parameters to those of the target speaker. In previous studies, the entire speaker space was partitioned into several clusters using vector quantization (VQ) [19], the mapping rules for each partition are then estimated using either a histogram [1] or minimum mean square error criterion [3,10]. The underlying assumption is that each cell corresponds to a phoneme. 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 [7], VQ-based classification causes discontinuity in transition regions. Hence, for voice conversion, the use of a soft-clustering approach is desirable [7,11,12]. Recently, a unit-selection based approach, which was originally devised for implementing the corpus-based concatenative text-to-speech (TTS) systems [20] was used to both alter the VTF parameters [13,14,16] and predict the target LP-residuals [15].

This paper is an extension of our previous work on voice transformation [12] 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 [1], GMM-based [7] and MMSE-based [12]).

To alleviate this problem, a feature-selection based approach was employed in the present study, where the sequence of the transformed features is given by the sequence of the features selected from the target speaker's database. We propose 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. Objective and subjective tests were performed to evaluate the efficiency of the proposed method. For the objective tests, both the distance reduction ratio and likelihood ratio for each feature were used to evaluate performance of the transformation. ABX tests using several phonetically balanced sentences were performed to subjectively evaluate performance. In addition, a preference test was administered to evaluate improvement in quality.

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 are presented in Section 4, and concluding remarks are summarized in Section 5.

II. 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, pitch and gain were used as the feature parameters. 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

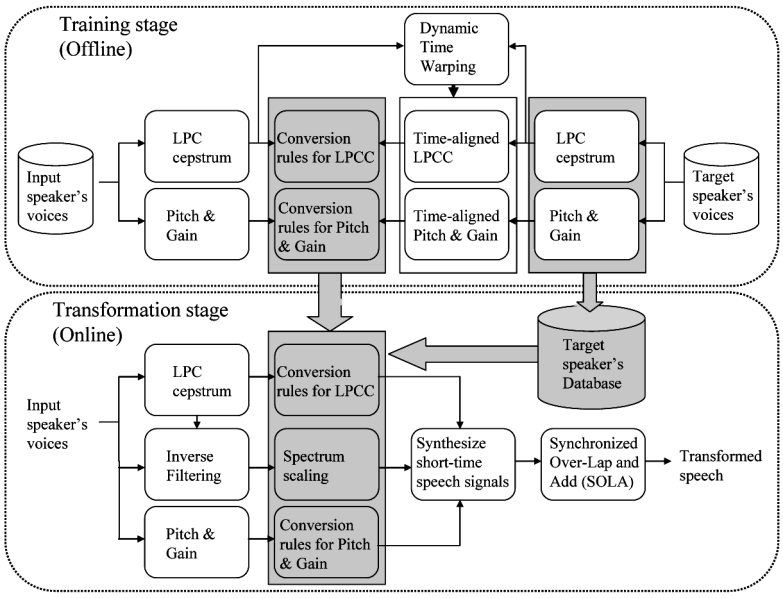

Fig. 1. Block diagram of the proposed voice transformation method.

would result. To time-align these sequences, dynamic time warping (DTW) [22] was applied in a preprocessing step. The resulting time-aligned LPCC, pitch and gain sequences were used to build conversion rules for each feature parameter. Note that pitch is valid only for the voiced frames. Hence, the unvoiced frames were eliminated from the time-aligned pitch sequences.

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 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 Synchronized Over-Lap and Add (SOLA) [23] algorithm to align each short-time speech signal. Note that the LP-residual was not included in the list of features to be transformed. Rather, the LP-residual from the source speaker was scaled in the frequency domain so that the fundamental frequency of the scaled LP-residual was identical to the target fundamental frequency (or equivalently, the inverse of the target pitch).

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

III. Transformation Rules

3.1 Overall transformation rules

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

$$x_t = [x_{c,t}\ x_{p,t}\ x_{g,t}]^T,\quad y_t = [y_{c,t}\ y_{p,t}\ y_{g,t}]^T$$

where the terms "C", "p" and "g" denote the LPCC, the pitch and the gain, respectively.

In the present study, the optimal transformed sequence Y^* for a given source sequence X is given by

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

where $f_{\mathrm{Y}|X,\Lambda}(Y \mid X, \Lambda_{\mathrm{Y}})$ is the likelihood function of Y given 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 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 X and the target model. The objective function in (1) can be written as follows:

$$f_{\mathrm{Y}|X,\Lambda}(Y \mid X, \Lambda_{\mathrm{Y}}) = \frac{f_{\mathrm{X},Y}(X,Y)}{f_{\mathrm{X}}(X) f_{\mathrm{Y}}(Y)} f_{\mathrm{Y}|\Lambda_Y}(Y \mid \Lambda_{\mathrm{Y}}) = \rho(X,Y) f_{\mathrm{Y}|\Lambda_Y}(Y \mid \Lambda_{\mathrm{Y}}) \quad (2)$$

where $\rho(X,Y)$ is the cross-correlation probability density function (PDF) between **X** and **Y**. Note that the two functions $\rho(X,Y)$ and $f_{\mathrm{Y}|\Lambda_Y}(Y \mid \Lambda_{\mathrm{Y}})$ 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 [12], in which inter-speaker variability was described by an inter probabilistic model, was used in the present study. According to this model, the joint probability of the source feature x_t, the target feature y_t, source speaker's i-th random source α_i and target speaker's j-th random source β_j is given by

$$f_{x,y,\alpha,\beta}(x_t, y_t, \alpha_i, \beta_j) = f_{x|\alpha}(x_t \mid \alpha_i) f_{\beta|\alpha}(\beta_j \mid \alpha_i) f_{y|\beta}(y_t \mid \beta_j) f_{\alpha}(\alpha_i) \quad (3)$$

where $f_{\beta|\alpha}(\beta_j \mid \alpha_i)$ is the cross correlation probability between the i-th random source of the source feature and the j-th random source of the target feature. This term describes the dependencies of the two random vector sets. Because the random sources $\{\alpha_i, \beta_j\}$ are assumed to be Gaussian,

$$f_{\mathrm{x}|\alpha}(x_t \mid \alpha_i) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma}_{x,i} \mid^{1/2}} \times \exp\left\{ -\frac{1}{2} (x_t - \boldsymbol{\mu}_{x,i})' \boldsymbol{\Sigma}_{x,i}^{-1} (x_t - \boldsymbol{\mu}_{x,i}) \right\} \quad (4)$$

$$f_{y|\beta}(y_t \mid \beta_j) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma}_{y,j} \mid^{1/2}} \times \exp\left\{ -\frac{1}{2} (y_t - \boldsymbol{\mu}_{y,j})' \boldsymbol{\Sigma}_{y,j}^{-1} (y_t - \boldsymbol{\mu}_{y,j}) \right\} \quad (5)$$

where $\boldsymbol{\Sigma}_{x,i}$ and $\boldsymbol{\mu}_{x,i}$ are the covariance matrix and mean vector of the i-th random source for the source feature, respectively. Similarly, $\boldsymbol{\Sigma}_{y,j}$ and $\boldsymbol{\mu}_{y,j}$ are the covariance matrix and mean vector, respectively, of the j-th random source for target feature. D is the order of the features. The method for estimating $f_{\beta|\alpha}(\beta_j \mid \alpha_i)$, $f_{\alpha}(\alpha_i)$ and parameters describing $f_{x|\alpha}(x_t \mid \alpha_i)$ and $f_{y|\beta}(y_t \mid \beta_j)$ from a given training corpus is based on a maximum likelihood criterion, as described in [12]. Using the adopted inter-speaker model, the cross-correlation PDF $\rho(X,Y)$ is given by

$$\rho(X,Y) = \prod_{t=1}^{T} \rho(x_t, y_t) \quad (6)$$

where

$$\rho(x_t, y_t) = \frac{f_{\mathrm{x},y}(x_t, y_t)}{f_{\mathrm{x}}(x_t) f_{\mathrm{y}}(y_t)} \quad (7)$$

and

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

$$f_{\mathrm{x}}(x_t) = \sum_i f_{\mathrm{x}|\alpha}(x_t \mid \alpha_i) f_{\alpha}(\alpha_i),$$

$$f_{\mathrm{y}}(y_t) = \sum_j [f_{y|\beta}(y_t \mid \beta_j) \sum_i f_{\beta|\alpha}(\beta_j \mid \alpha_i) f_{\alpha}(\alpha_i)]$$

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, $\rho(x_t, y_t)$ is given by

$$\rho(x_t, y_t) = \rho_c(x_{c,t}, y_{c,t})\rho_p(x_{p,t}, y_{p,t})\rho_g(x_{g,t}, y_{g,t}) \quad (8)$$

where $\rho_c(x_{c,t}, y_{c,t})$, $\rho_p(x_{p,t}, y_{p,t})$ and $\rho_g(x_{g,t}, y_{g,t})$ are the cross correlation PDFs for LPCC, pitch and gain, 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.

$$\Lambda_Y = \{A_Y, B_Y, \pi_Y\} = \{a_{ij}, b_i, \pi_i, 1 \le i, j \le N_s\} \quad (9)$$

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 when multi-channel observation sequences are given has been proposed previously [24]. This model was primarily used for representation of the relationship between multi-channel observations. In the present study, this model is adopted for representation of inter-feature relationships. There are several methods available for integration of individual features to represent the relationship among them. The models can be categorized as either early integration (EI) or late integration (LI) models [24]. 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 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

$$b(y_t) = f_c(y_{c,t}) f_p(y_{p,t}) f_g(y_{g,t}) \quad (10)$$

where f_c, f_p and f_g denote the density functions for LPCC, pitch and gain, respectively. In (10), the state index i is omitted for simplicity. When the Gaussian mixture model is adopted, an individual density function is given by

$$f_k(y_{k,t}) = \sum_{i=1}^{N_k} f_{\lambda_k}(\lambda_{k,i}) f_{y|\lambda_k}(y_{k,t} \mid \lambda_{k,i}),\ k = \{c, p, g\} \quad (11)$$

where $N_{k=\{c,p,g\}}$ are the number of Gaussian components for LPCC, pitch and gain, respectively, and $f_{\lambda_k}(\lambda_{k,i})$ are the mixture weights of each feature of the i-th Gaussian component. $f_{y|\lambda_k}(y_{k,t} \mid \lambda_{k,i})$, are the i-th Gaussian component for each feature. Using (11), the state observation density is given by

$$b(y_t) = \sum_{i=1}^{N_c}\sum_{j=1}^{N_p}\sum_{k=1}^{N_g} f_{y,\lambda_y}(y_{c,t}, y_{p,t}, y_{g,t}, \lambda_{c,i}, \lambda_{p,j}, \lambda_{g,k}) \quad (12)$$

where $f_{y,\lambda_y}(y_{c,t}, y_{p,t}, y_{g,t}, \lambda_{c,i}, \lambda_{p,j}, \lambda_{g,k})$ is the joint probability function of the set of the observation features $y_{c,t}$, $y_{p,t}$ and $y_{g,t}$ and the set of the Gaussian random sources $\lambda_{c,i}$, $\lambda_{p,j}$ and $\lambda_{g,k}$, which is given by

$$f_{y,\lambda_y}(y_{c,t}, y_{p,t}, y_{g,t}, \lambda_{c,i}, \lambda_{p,j}, \lambda_{g,k}) = \prod_{k=\{c,p,g\}} f_{\lambda_k}(\lambda_{k,i}) f_{y|\lambda_k}(y_{k,t} \mid \lambda_{k,i}) \quad (13)$$

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 hour in length [20]. However, selection of the optimal features from a large database is not a trivial undertaking. 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 $S_C(x)$ was constructed using the predicted target feature $\hat{f}_y(x)$ as follows.

$$S_C(x_t) = \{y \mid y \in S_Y, \| y - \hat{f}_y(x) \|^2 \le \varepsilon_{th}\} \quad (14)$$

where ε_{th} is the threshold, which can be adjusted so that the number of candidates is 20~30. In the present study, the minimum mean square error (MMSE) -based transformation method [12] was employed to get the predicted target feature for LPCC. For pitch and gain, the predicted target value was obtained by scaling, so that the average of the scaled values is identical to that of the target values.

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

$$\log[f_{Y|X,\Lambda,\Omega}(Y \mid X, \Lambda_Y, \Omega)] = \sum_{t=1}^{T}[\rho'(x_t, y_t) + b'_{\omega_t}(y_t) + C_{t-1,t}] \quad (15)$$

where

$$\rho'(x_t, y_t) = \log \rho(x_t, y_t)$$

$$b'_{\omega_t}(y_t) = \log b_{\omega_t}(y_t)$$

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

The optimal transformed sequence Y^* is then given by

$$Y^* = \arg \max_{Y \in S_C} \{\max_{\Omega \in \Omega_T} \log[f_{Y|X,\Lambda,\Omega}(Y \mid X, \Lambda_Y, \Omega)]\} \quad (16)$$

where $S_C = \{S_C(x_t)\}_{t=1}^{T}$ is the set of candidates for $1 \le t \le T$ and Ω_T denotes the set of all possible HMM state sequences.

Equation (16) can be maximized using a dynamic programming technique, such a Viterbi-trellis search.

In some sense, the objective of the log likelihood function (15) is similar to that of the cost function employed in corpus-based concatenative TTS systems [20]. In this type of TTS system, the optimal unit sequence is obtained by minimizing the total cost function which includes a target cost and a concatenation cost. The objective of a target cost is to maximize the similarities (or, equivalently, minimizing the differences) between the selected units and the targets. In TTS systems, the targets are specified based on the context information to be synthesized. Whereas, in VT, the targets are characterized by the target speaker's speech signals and the source speaker model. Accordingly, the term $\rho'(x_t, y_t) + b'_{\omega_t}(y_t)$ in (15) is referred to as the target likelihood function in this study. The objective of a concatenation cost in TTS systems is to build the unit sequence so that spectral trajectory of the selected unit sequence possess some degree of smoothness. For a quasi-stationary random process (e.g. speech signal), the transient PDFs between the same states are higher than those between the different states. Thus, the maximum likelihood criterion (16) tends to select the feature sequence so that the state indices of the neighboring features are same. Since the the features belonging to the same states are close to each other, maximizing $C_{t-1,t}$ in (15) leads to smoothly evolving spectral trajectories over time. This means that the role of $C_{t-1,t}$ is similar to that of the concatenation cost function in TTS systems. Accordingly, $C_{t-1,t}$ is referred to as the concatenation likelihood function in this study.

IV. Experimental Results

The database used to obtain the conversion rules consisted of 200 utterances spoken in Korean by three men and one woman whom we refer to as M1, M2, M3 and F, respectively. M1, M2 and F were professional voice actors. An additional 100 utterances spoken by the same individuals were prepared for both objective and subjective evaluation. Speech signals were digitized at a rate of 16 kHz. The orders of the LPC coefficients and the LPC cepstrum were 20 and 30, respectively. A 25-ms Hanning window was used to both compute and extract the LPC parameters at 10 ms intervals. The pitch period was estimated by applying the clipped autocorrelation method [21]. Each Gaussian component was constrained to a diagonal covariance matrix. Variance limiting [25] was also used to estimate each component of the covariance matrices. Results of two VT experiments are presented. The first experiment involved male-to-male conversion (M3→M1), while the second tested male-to-female conversion (M2→F).

4.1 Objective evaluation

To evaluate the performance of the proposed voice transformation method, two objective measurements were adopted. First, the following distance reduction ratio [6] was used

$$D_{ratio} = \left\{1 - \frac{D(\hat{Y},Y)}{D(X,Y)}\right\} \times 100(\%) \tag{17}$$

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

Another objective measure is the following log likelihood ratio [9].

$$L_{ratio} = \log\left(\frac{p(\hat{Y} \mid \Lambda_Y)}{p(Y \mid \Lambda_X)}\right) = \log p(\hat{Y} \mid \Lambda_Y) - \log p(Y \mid \Lambda_X) \tag{18}$$

where Λ_X and Λ_Y are probabilistic models estimated from the source speaker's training corpus and the target speaker's training corpus, respectively. In this work, the HMMs were employed to represent each speaker's probabilistic model. These models used five states and five Gaussians. According to the above equation, L_{ratio} is typically less than zero when Y is similar to the source speaker's feature sequence. By contrast, if Y approximates the target speaker's feature sequence, L_{ratio} is greater than zero. Hence, large positive values of the log likelihood ratio indicate that the transformed features are statistically similar to the target features.

For comparison, three types of conversion methods were adopted in this experiment; the VQ-based approach proposed by Abe *et al*. [1], the GMM-based approach proposed by Stylianou *et al*. [7] and the ML-based statistical approach proposed by author [12]. For each method, the conversion rules for each feature (LPCC, pitch and gain) were constructed separately. Table 1 presents the D_{ratio} obtained for the test corpus using the three methods. The number of source/target random sources, centroids (VQ-based) or Gaussians (GMM-based method) ranged from 4 to 128. When the number of random sources exceeded 128, it was impossible to evaluate the performance of all three methods due to over-estimation. In most cases, D_{ratio} increased with the number of random sources. For the GMM method, the D_{ratio} was nearly saturated, when the number of Gaussians exceeded 32. The overall D_{ratio} of the proposed method was not greater than the values reported previously, because the proposed method employed the maximum likelihood criterion. For the GMM-based approach, the mapping rules were set so as to minimize the overall distance between the transformed and target

Table 1. LPCC distance reduction ratios for each method.

Conversion	M3 → M1				M2 → F			
# random sources	VQ-based	GMM	ML-based	Proposed	VQ-based	GMM	ML-based	Proposed
4	40.3	56.0	40.0	20.1	51.7	67.7	52.0	50.3
8	45.5	56.6	46.7	33.3	57.4	68.6	58.3	52.7
16	48.7	56.8	49.1	39.3	60.9	69.1	62.2	56.7
32	50.8	56.2	51.7	41.6	63.6	69.4	64.9	58.6
64	52.2	54.7	53.8	42.8	65.5	69.0	66.7	60.3
128	53.3	52.0	55.0	44.0	66.8	67.8	68.0	61.6

Table 2. Log likelihood ratios for each methods.

Conversion	M3 → M1				M2 → F			
# random sources	VQ-based	GMM	ML-based	Proposed	VQ-based	GMM	ML-based	Proposed
4	1.39	3.85	1.57	0.95	7.09	9.79	6.45	5.95
8	1.19	4.23	1.62	0.90	7.06	10.56	7.62	7.71
16	1.12	4.46	1.83	2.24	7.04	11.38	8.34	8.45
32	1.41	4.71	2.37	2.92	7.73	11.92	8.72	8.94
64	1.65	4.99	3.42	3.52	8.53	12.52	10.33	10.34
128	2.15	5.28	4.08	4.00	8.56	12.93	12.01	12.00

features. Similarly, in VQ-based and ML-based, the transformed feature was given by a linear combination of the code vectors, for which the linear combination weights were obtained using the minimum mean square error (MMSE) criterion. In the proposed method, the optimal combinations of the three features were selected to maximize the likelihood with respect to the target model. The results indicate that the selected features that have the maximum likelihood do not necessarily correspond to minimal distortion. The overall D_{ratio} of the voice transformation between genders (M2→F) was greater than that within gender (M3→M1). The difference between inter- and intra-gender transformation was more remarkable for pitch transformation. This result most likely occurred because the difference in feature distributions of men and women is larger than the difference in feature between two speakers of the same gender. In this case, the larger denominator $D(X,Y)$ in (17) leads to an increase in D_{ratio}.

In terms of the likelihood ratio, the performance of the proposed method was superior to that of VQ-based approach and close to that of the GMM approach, as shown in Table 2. Moreover, the reults of the proposed method was slighlty better than the ML-based approach. Considering that the maximum likelihood criterion was employed in the proposed method, this result is somewhat expected. It was also observed that the likelihood ratio increased with the number of candidates. This is because the probability of finding combinations of the three features (LPCC, pitch and gain) that more closely approximate the target speaker's model increases when there are more candidates. The average log likelihood for the transformed features relative to the target model was -10.21 and -11.01 for the speakers F and M1, respectively. These values are close to those of the target features (-9.81 for speaker F and -10.45 for speaker M1). This result indicates that the proposed method yields transformed features that are statistically close to the target.

4.2 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. The subjects were presented the stimuli via headphones. Each listener was allowed to listen to the stimuli as many times as needed before determination. Audio examples can be listened on the web site: http://home.konkuk.ac.kr/~kseung/VT/ demo_page.htm.

Fig. 2 shows the correct identification ratios versus the number of target/source random sources. Although the objective performance of the proposed method was inferior to that of the GMM-based approach, the subjective performance of the proposed method was nearly identical to that of the GMM-based approach. Compared with the VQ-based method, the proposed method showed the higher identification ratios even thought average D_{ratio} of the proposed method was not higher than that of the VQ-based method. A possible explanation for this ABX test result is that for the proposed method, the transformed features were given by the features derived from the target speaker. By contrast, for both the VQ-based and GMM-based methods, the transformed features were artificially obtained.

For inter-gender transformation (M2→F), 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 M2→F conversion was increased for all three methods. For the M2→F conversion, the number of random sources was highly correlated with the correct identification (The correlation coefficients for VQ-based, the GMM-based and the proposed method were 0.8279, 0.7242 and 0.7075, respectively).

In the second test, we compared the quality of speech synthesized using our method with that obtained using either VQ-based or the GMM-based method. 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 summarized in Table 3, the proposed method performed slightly better than the conventional approaches. The subjects indicated that the clarify of the speech signals synthesized using the proposed method was superior to that of the other methods. This difference in quality most likely resulted from bandwidth widening problem caused by the averaging of the transformed speech signals in VQ-mapping and the GMM-based method. An example of the spectrograms for the target signal, the transformed signal using the GMM-based method and the transformed signal using the proposed method is shown in Fig. 3. This example clearly explains the reasons for the perceptual superiority of the proposed method, compared with the GMM-based method. In this example, more clear formant trajectories are observed in the transformed signal using the proposed method, especially in the high frequency regions. For the GMM-based method, the formants at the higher frequency regions are weaker and sometimes lost. This is mainly due to the averaging effects of the GMM-based method. However, pop and click sounds were sometimes perceived in the speech signals transformed using the proposed method. These sounds might result from abrupt changes in the sequence of the transformed features. Increased emphasis on

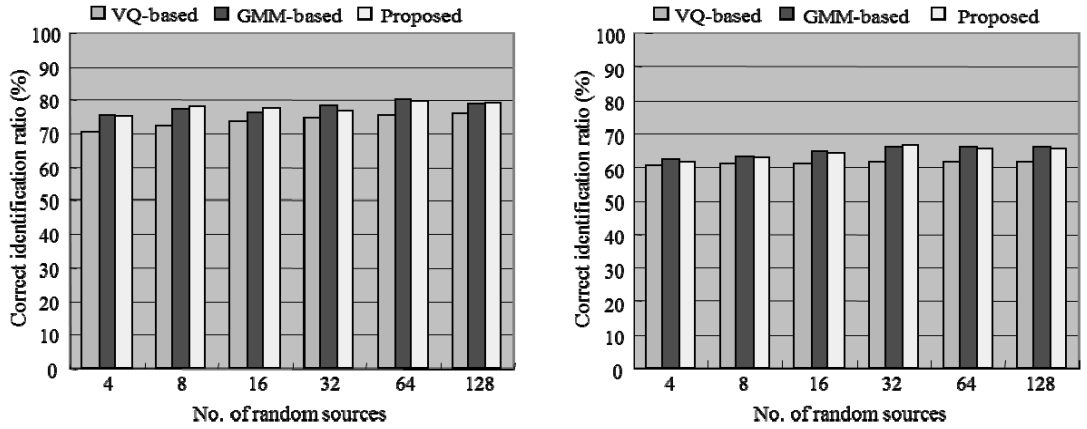

Fig. 2. ABX test results for each methods (Left: M2→F conversion, Right: M3→M1 conversion).

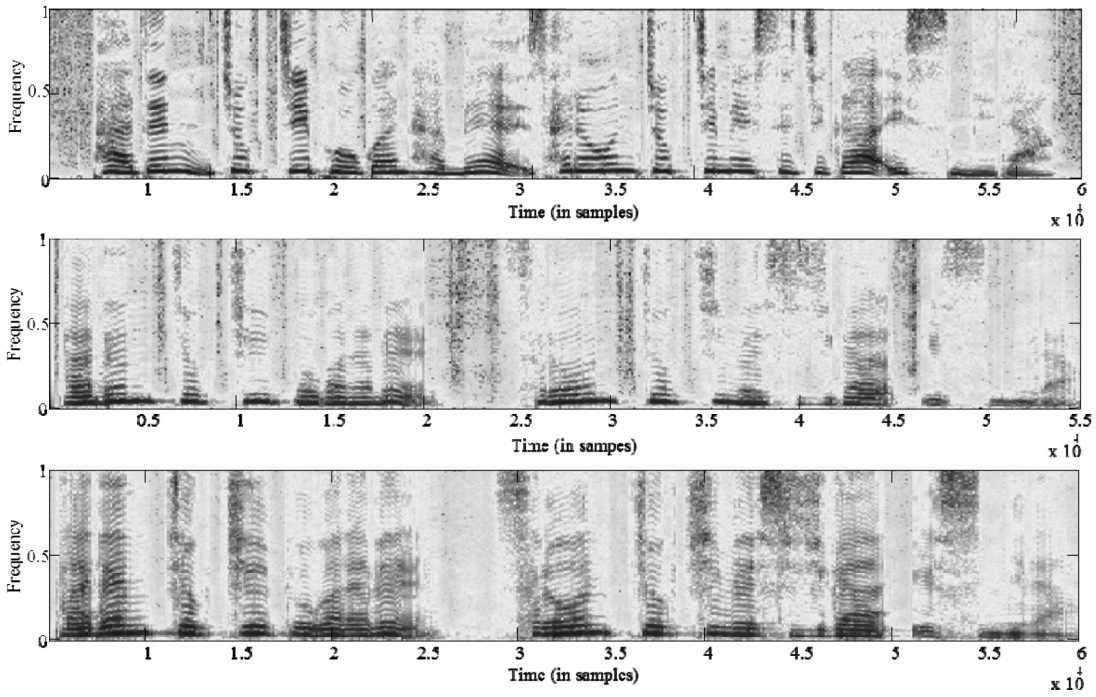

Fig. 3. An example of the spectrograms; Top: Target speech, Middle: Transformed speech using the GMM-based method, Bottom: Transformed speech using the proposed unit-selection method (The frequency axis for each spectrogram is normalized by Nyquist frequency (=16000 Hz)).

the concatenation likelihood function in (15) might reduce these sounds, as the transitional probabilities between identical states tend to have a larger value. However, excessive emphasis on the concatenation likelihood function might degrade voice transformation performance if the target likelihood function is not given adequate weight. Hence the quality of the synthesized speech signals can be further improved by careful weighting of the likelihood functions in (15).

Table 3. Preference test results of each method.

Target speaker	VQ-based	GMM-based	Proposed
M1	27.3	34.5	38.2
F	27.1	35.4	37.5

V. 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 represented in the context of HMM, which was built from the training source features.

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 inferior to that of the conventional methods. However, the results of the subjective evaluation show 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.

Acknowledgement

This work was supported by the Konkuk University in 2009.

References

1. M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE ICASSP*, pp. 565-568, 1988.
2. M. Savic and I. H. Nam, "Voice personality transformation," *Digital Signal Processing*, vol. 4, pp. 107-110, 1991.
3. H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2-3, pp. 175-187, 1992.
4. H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectral tilt," *Speech Communication*, vol. 16, no. 2, pp. 153-164, 1995.
5. M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants of voice conversion using artificial neural networks," *Speech Communication*, vol. 16, no. 2, pp. 207- 216, 1995.
6. N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication*, vol. 16, no. 2, pp. 139-152, 1995.
7. Y. Stylianou O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 6, no. 2, pp. 131-142, 1998.
8. N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 5, no. 2, pp. 97-105, 1997.
9. L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, no. 28, pp. 211-226, 1999.
10. K. S. Lee, D. H. Youn and I. W. Cha, "A New voice personality transformation based on both linear and nonlinear prediction analysis," in *Proc. ICSLP*, pp. 1401-1404, 1996.
11. K. S. Lee, D. H. Youn and I. W. Cha, "Voice conversion using a low dimensional vector mapping," *IEICE Trans. on Information and System*, vol-E85D, no. 8, pp. 1297-1305, 2002.
12. K. S. Lee "Statistical approach for voice personality transformation," *IEEE Trans. on Audio, Speech and Language processing*, vol. 15, no. 2, pp. 641-651, 2007.
13. Z.-H. Jian and Y. Zhen, "Voice conversion using Viterbi algorithm based on Gaussian mixture model," in *Proc. Intelligent Signal Processing and Communication Systems*, pp. 32-35, 2007.
14. D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, S. Narayanan, "Text-Independent Voice Conversion Based on Unit Selection," in *Proc. IEEE ICASSP*, pp. 14-19, 2006.
15. D. Sundermann, H. Hoge, A. Bonafonte, H. Ney and A. W. Black, "Residual prediction based on unit

selection," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 369-374, 2005.
16. T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez and Y. Stylianou, "Towards a Voice Conversion System Based on Frame Selection," in *Proc. IEEE ICASSP*, pp. 15-20, 2007.
17. S. J. Cox and J. S. Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting," in *Proc. IEEE ICASSP*, pp. 294-297, 1989.
18. D. G. Childers, B. Yegnanarayana and Ke Wu, "Voice Conversion: Factors responsible for quality," in *Proc. IEEE ICASSP*, pp. 748-751, 1985.
19. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. 28, Issue 1, pp. 84-95, 1980.
20. M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proc. Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, March 1999.
21. L. R. Rabiner and R. W. Schafer, *Digital Processing of speech signals*, Prentice-Hall, 1987.
22. G. M. White and R. B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. ASSP-24, no. 2, pp. 183-188, 1976.
23. S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," in *Proc. ICASSP 85*, pp. 493-469, 1985.
24. A. Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. R. Soc. London B*, vol. 335, pp. 71-78, 1992.
25. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 3, no. 1, pp. 72-83, 1995.

Profile

▸ Ki-Seung Lee

He received the B.S., the M.S. and the Ph.D degrees in electronics engineering from Department of Electronics, Yonsei University, Seoul, Korea, in 1991, 1993 and 1997, respectively. In Feburary 1993, he joined CSPR (Center for Signal Processing Research) in Yonsei University. From October 1997 to September 2000, he had been with AT&T Labs–Research, Florham Park NJ, USA where he was working on very low bit rate speech coding schemes and prosody generation module of Text-to-Speech. From November 2000 to August 2001, he had been with SAIT (Samsung Advanced Institute of Technology), Suwon, Korea, where he was working on a corpus-based concatenative Text-to-Speech. Since August 2001, he has been with the faculty position of Konkuk University, Seoul, Korea. He is now a professorin this university. His interests include bio-signal modeling, speech signal processing, and audio signal processing.