Comprehensive Source-Target Speaker Voice Conversion Analysis

He Pan, Yangjie Wei, and Nan Guan
Northeastern University
Shenyang, China
panhe_sy@126.com, weiyangjie@ise.neu.edu.cn,
guannan@ise.neu.edu.cn

Yi Wang
Uppsala University
Uppsala, Sweden
wangyi@ise.neu.edu.cn

***Abstract*— Voice conversion system modifies a speaker's voice to be perceived as another speaker uttered, and now it is widely used in many real applications. However, most research only focuses on one aspect performance of voice conversion system, rare theoretical analysis and experimental comparison on the whole source-target speaker voice conversion process has been introduced. Therefore, in this paper, a comprehensive analysis on source-target speaker voice conversion is conducted based on three key steps, including acoustic features selection and extraction, voice conversion model construction, and target speech synthesis, and a complete and optimal source-target speaker voice conversion is proposed. First, a simple and direct serial feature fusion form consisting of prosodic feature, spectrum parameter and spectral envelope characteristic, is proposed. Then, to void the discontinuity and spectrum distortion of a converted speech, D_GMM (Dynamic Gaussian Mixture Model) considering dynamic information between frames is presented. Subsequently, for speech synthesis, STRAIGHT algorithm synthesizer with feature combination is modified. Finally, the objective contrast experiment shows that our new source-target voice conversion process achieves better performance than the conventional methods. In addition, both objective evaluation (speaker recognition system) and subjective evaluation are used to evaluate the quality of converted speech, and experimental result shows that the converted speech has higher target speaker individuality and speech quality.**

Keywords-Voice Conversion, Serial Feature Fusion, D_GMM, STRAIGHT Synthesis, Speaker Recognition.

I. INTRODUCTION

Source-target speaker voice conversion is a technique that modifies a source speaker's speech to make it sound like that uttered by a target speaker without changing the speech content [1]. In the last two decades, much attention has been attracted to it due to its wide potential application areas, such as dubbing in films, restoring damaged voice and disguising personal voice [2].

Normally, source-target speaker voice conversion consists of three key steps: (1) selection and extraction of representative acoustic features; (2) construction of voice conversion model; (3) synthesis of the target speech. Firstly, acoustic features selection and extraction is to select acoustic features those can represent a speaker's individual identity, and extract them correctly. Many researchers have proved that prosodic features, formant frequency and spectral parameters are the most important features used in real applications [3-4]. However, most study only focus on one or two of them to represent a speaker's individual identity [5-6], that causing the covered acoustic features to be incomprehensive and the final synthesized speech to be inaccurate.

Secondly, a voice conversion model consists of a sequence of mapping rules between the source speaker and the target speaker. Gaussian mixture model, which converts acoustic parameters based on the minimum mean square error, is one of the most widely used models in voice conversion research [5-7] because of its statistical framework and stable performance. But GMM's usefulness also has some limitations. For example, GMM is on the base of single conversion frame, where the sequential information between frames is ignored. So it may lead to spectrum distortion and speech quality deterioration.

Finally, there are many methods of speech synthesis [8], among which speech transformation and representation using adaptive interpolation of weighted spectrum, or STRAIGHT, is a comparatively mature algorithm [9]. STRAIGHT can divide speech signal into independent spectral parameters and F0 parameters, and in synthesis process, it can modify duration, F0, and spectral parameters flexibly. In addition, it will not cause obvious deterioration of speech quality.

Although source-target speaker voice conversion methods are widely used in real application, most research focuses on the performance of one aspect, such as conversion model, rare theoretical analysis and experimental comparison on the whole source-target speaker voice conversion process has been introduced. Therefore, in this paper, a comprehensive analysis on source-target speaker voice conversion is conducted based on three key steps above, a sequence of evaluation results based on theoretical analysis and experiments are attained, and a complete and optimal source-target speaker voice conversion is proposed.

This paper is organized as follows. In Section II, the basic principle of voice conversion including feature extraction, GMM model and speech synthesis, is

introduced. In Section III, a comprehensive and improved source-target speaker voice conversion is proposed, as well as voice conversion evaluation criterion. In Section IV, experimental results and evaluations are presented. Section V is conclusion.

II. BASIC VOICE CONVERSION PROCESS

A. Acoustic Feature Extraction

Psychophysical studies have shown that human perception of the sound frequency contents for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, in Hz, a subjective pitch is measured on a scale called 'Mel' scale.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Mel-cepstra analysis, based on the human hearing mechanism has better recognition performance and anti-noise property in voice conversion, and its extraction process is illustrated in Figure 1 [10, 20].

An N-point DFT of the discrete input signal $y(n)$ is defined as

$$Y(k) = \sum_{n=1}^{M} y(n) e^{\left(-\frac{j\,2\pi nk}{M}\right)} \quad 1 \le k \le M \quad (2)$$

Next, the filter bank which has linearly spaced filters in the Mel scale, is imposed on the spectrum, and the filter response $\phi_i(k)$ of the ith filter in the bank is defined as

$$\phi_i(k) = \begin{cases} 0 & k < k_{b_{i-1}} \\ k - k_{b_{i-1}} \big/ k_{b_i} - k_{b_{i-1}} & k_{b_{i-1}} \le k < k_{b_i} \\ k_{b_{i+1}} - k \big/ k_{b_{i+1}} - k_{b_i} & k_{b_i} \le k < k_{b_{i+1}} \\ 0 & k \ge k_{b_{i+1}} \end{cases} \quad (3)$$

Subsequently, the boundary points for each filter i (i=1,2,...,Q) are calculated as equally spaced points in the Mel scale using Eq. (4).

$$K_{b_i} = \left(M \big/ f_s\right) f_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\left\{f_{mel}(f_{high}) - f_{mel}(f_{low})\right\}}{Q+1}\right] \quad (4)$$

where f_s is the sampling frequency in Hz, f_{low} and f_{high} are the low and high frequency boundaries of the filter bank, respectively. Q denotes the number of filters in the filter bank. f_{mel}^{-1} is the inverse of the transformation shown in Eq. (1) and is defined as

$$f_{mel}^{-1}(f_{mel}) = 700\left[10^{\frac{f_{mel}}{2595}} - 1\right] \quad (5)$$

In the next step, the output energies $e(i)$ (i=1,2,...,Q) of the Mel-scaled bank-pass filters are calculated as a sum of the signal energies $|Y(k)|^2$ falling into a given Mel frequency band weighted by the corresponding frequency response $\phi_i(k)$.

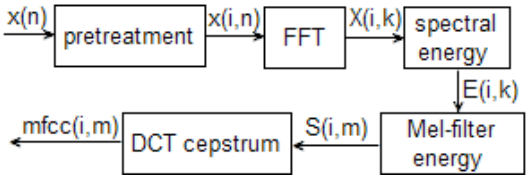


Figure 1. Extracting Mel-cepstra parameters process.

$$e(i) = \sum_{k}^{M} |y(k)|^2 \phi_i(k) \quad (6)$$

Finally, the DCT-II is applied to the log filter bank energies for decorrelating the energies and the final Mel-cepstra coefficients C_m are provided in Eq. (7).

$$c_m = \sqrt{\frac{2}{N}} \sum_{l=0}^{Q-1} \log\left[e(l+1)\right] \cos\left[m\left(\frac{2l+1}{2}\right)\frac{\pi}{Q}\right] \quad (7)$$

where m=0,1,2,...,R-1, and R is the desired number of Mel-cepstra.

But normal Mel-cepstra only captures static characteristics of speech parameters mentioned above, rather than dynamic characteristics those ears are more sensitive to. So its Delta and Delta-Delta parameters can eliminate the relevance between frames and approximate speech dynamic characteristics well. In real implementation, Delta feature calculation is simplified as in Eq. (8).

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta\left(c_{i+\theta} - c_{i-\theta}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \quad (8)$$

where d_t is the Delta coefficient at frame t, computed in terms of the corresponding static coefficients $c_{i+\theta}$ to $c_{i-\theta}$ and Θ is the size of a delta window.

B. GMM Voice Conversion Model

GMMs have been proved to be very useful in research on voice conversion [11]. According to feature extraction methods introduced before, we can get source feature vectors and target feature vectors, respectively. Let x_t and y_t be source and target feature vectors at frame t, respectively. The form of GMM model is as follows,

$$P(z_t) = \sum_{m=1}^{M} \omega_m N(z_t; \mu_m^{(z)}; \Sigma_m^{(z)}) \quad (9)$$

where z_t is vector pairs $[x_t^{\mathrm{T}}, y_t^{\mathrm{T}}]^{\mathrm{T}}$, $(.)^{\mathrm{T}}$ denotes transposition of the vector, M is the total number of mixture components, ω_m is the weight of the m-th mixture component, $N(z_t;\mu_m^{(z)};\Sigma_m^{(z)})$ is the m-th mixture component, $\mu_m^{(z)}$, $\Sigma_m^{(z)}$ are the mean vector and variance matrix of z separately, written as,

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \ \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \quad (10)$$

where, $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vector of the mth mixture component for the source and that for the target, respectively. The matrices $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the covariance matrix of the mth mixture component for the source and that for the target, respectively. The matrices $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the cross-covariance matrix of the mth mixture component for the source and that for the target, respectively. These covariance, $\Sigma_m^{(xx)}$, $\Sigma_m^{(xy)}$, $\Sigma_m^{(yx)}$, $\Sigma_m^{(yy)}$ are diagonal in this paper.

There are now several techniques available to estimate the parameters of a GMM. So far, the most popular and well-established method is maximum likelihood (ML) estimation. The main of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. It is often completed by Expectation-Maximization algorithm (EM) [12].

The GMM is trained with the EM algorithm using the joint vectors, which are automatically aligned by dynamic time warping (DTW), in a training set. This training method provides estimates of the model parameters robustly compared with the least-squares estimation [7], particularly when the amount of training data is small [19].

The conditional probability density of y_t, given x_t, is also represented as a GMM, as shown in Eq. (11).

$$P(\mathrm{y}_t \left| \mathrm{x}_t \right.) = \sum_{m=1}^{M} P(\mathrm{m} \left| \mathrm{x}_t \right.) P(\mathrm{y}_t \left| \mathrm{x}_t \right., m) \quad (11)$$

where

$$P(\mathrm{m} \left| \mathrm{x}_t \right.) = \frac{\omega_m N(\mathrm{x}_t;\ \mu_m^{(x)};\ \Sigma_m^{(xx)})}{\sum_{m=1}^{M} \omega_m N(\mathrm{x}_t;\ \mu_m^{(x)};\ \Sigma_m^{(xx)})} \quad (12)$$

$$P(\mathrm{y}_t \left| \mathrm{x}_t \right., m) = N(\mathrm{y}_t;\ \mathrm{E}_{m,t}^{(y)};\ \mathrm{D}_m^{(y)}) \quad (13)$$

The mean vector $E_{m,t}^{(y)}$ and the covariance matrix $D_m^{(y)}$ of the m-th conditional probability distribution are written as

$$\begin{aligned} \mathrm{E}_{m,t}^{(y)} &= \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (\mathrm{x}^t - \mu_m^{(x)}) \\ \mathrm{D}_m^{(y)} &= \Sigma_m^{(yy)} - \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} \Sigma_m^{(xy)} \end{aligned} \quad (14)$$

The conversion is performed on the basis of the minimum mean-square error as follows

$$\begin{aligned} y_t^{'} &= \int P(\mathrm{y}_t \left| \mathrm{x}_t \right.) y_t dy_t \\ &= \int \sum_{m=1}^{M} P(\mathrm{m} \left| \mathrm{x}_t \right.) N(\mathrm{y}_t;\ \mathrm{E}_{m,t}^{(y)};\ \mathrm{D}_m^{(y)}) y_t dy_t \\ &= \sum_{m=1}^{M} P(\mathrm{m} \left| \mathrm{x}_t \right.) \mathrm{E}_{m,t}^{(y)} \end{aligned} \quad (15)$$

where $y_t^{'}$ is the converted target feature vector.

C. STRAIGHT Voice Synthesis Algorithm

STRAIGHT algorithm analyses model parameters based on signal processing and auditory processing relevant algorithms, makes synthetic speech closer to original speech. STRAIGHT analysis and synthesis algorithm consists of following three modules: (1) spectral estimation removing cycle impact; (2) smoothing and reliable pitch contour extraction; (3) synthesizer. Step (1) and (2) are processed in the train phase, two input parameters, F0, and two-dimensional spectral envelope, are attained. After that, combination of PSOLA and minimum phase impulse response are conducted to reconstruct speech, and specific realization is introduced in [13].

III. IMPROVED VOICE CONVERSION PROCESS

Section II introduces the whole process of source-target speaker voice conversion. However, it is limited to build preliminary frame, and deep analysis of each step implementation has rarely conducted. So in this section, an improved voice conversion process is proposed. Firstly, in acoustic feature selection, considering the pronunciation mechanism [14], we select F0 of prosodic feature and formant frequency presenting spectral envelope characteristic to improve the quality of traditional voice synthesis. Secondly, an improvement is introduced into traditional GMM voice conversion model. Finally, with F0 and formant frequency, we modify the basic STRAIGHT synthesizer. Our complete voice conversion process is illustrated in Figure 2.

A. Additional Feature Extraction

1) Formant Frequency

Vocal tract can be taken as a vocal pipe with non-uniform sections, and it plays a role of resonator when pronouncing. In glottis, it causes resonance when periodic pulse excitation comes into the vocal tract and brings a group of resonance frequency, called formant frequency [14].

Common formant estimation methods are spectrum analysis and LPC (linear prediction coding), etc. [15] LPC is a fast and accurate method which is often used in estimating basic speech parameters. The most direct method is LPC root method.

In LPC model, speech signal sample $s(n)$ can be shown as following difference equation,

$$s(\mathrm{n}) = \sum_{i=1}^{p} a_i s(\mathrm{n} - i) + Gu(\mathrm{n}) \quad (16)$$

where $u(n)$ is excitation function, G is gain factor, $\{a_i;\ i = 1, 2, \ldots, p\}$ is coefficients of LPC. The corresponding digital filter transformation function $H(z)$ is measured by,

$$H(\mathrm{z}) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \quad (17)$$

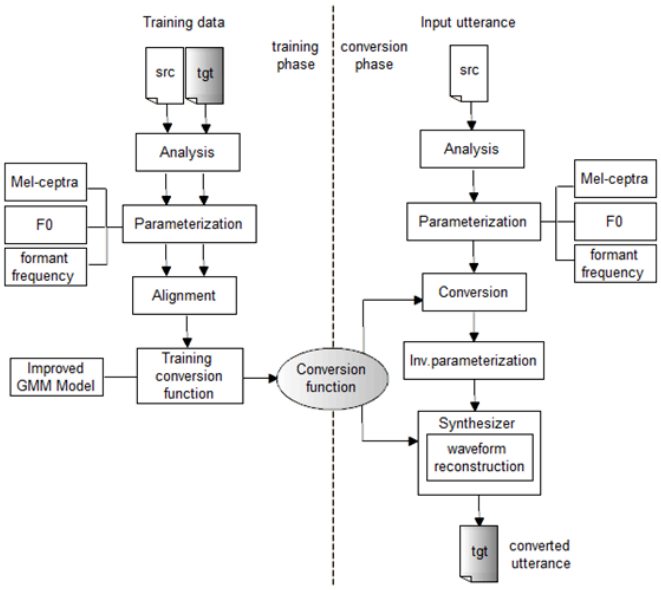


Figure 2. Complete and improved voice conversion process.

where prediction error filter

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{18}$$

where the polynomial coefficient decomposition of $A(z)$ can get formant center frequency and bandwidth accurately. The detailed process of finding polynomial complex root is following.

Setting $z_i = r_i e^{j\theta_i}$ is arbitrary complex root value, so its conjugate value $z_i^* = r_i e^{-j\theta_i}$ is also a root. The formant frequency corresponding to z_i is F_i, 3dB bandwidth is B_i, so the relation of F_i, B_i and z_i is shown as

$$\begin{aligned} 2\pi T F_i &= \theta_i \\ e^{-B_i \pi T} &= r_i \end{aligned} \tag{19}$$

where T is sampling period. Therefore,

$$F_i = \theta_i / 2\pi T \tag{20}$$

Because the prediction error filter order p is given in advance, the quantity of the complex conjugate pair is not more than $2/p$. Thus, it is not complex to judge whether an extreme point belongs to formant. In addition, it is easier to remove extra extreme points which are not formant, since the bandwidth of those extra extreme points is much bigger than the formant.

2) F0 Contour

When we pronounce sonant, vocal cords vibrate and generate period excitation pulse. The vibration period of excitation pulse is a pitch period. The reciprocal of a pitch period is F0. To make synthesized target speech with high quality, reasonable F0 model is essential.

There are various algorithms for modeling and converting F0 contours, and a common view of these methods is that the F0 conversion is directly performed on F0 contour itself. However, there is serious uncertainty in prosody even when a speaker utters the same transcription in different time. In other words, the purpose of F0 conversion should not only focus on estimating target F0 contour. According to it, we argue that the emphasis of F0 conversion should convert the underlying F0 pattern dependent on the speaker rather than F0 contour itself. Xu and Wang proposed a theoretical framework for estimating F0 contours in Mandarin [16].

In the pitch target model, variations in surface F0 contours result not only from the underlying pitch units (syllables for Mandarin), but also from the articulatory constraints. Pitch target are defined as the smallest operable units associated with linguistically functional pitch units, and these targets may be static or dynamic. Among these models, the features of the pitch target model are quite suitable for prosody conversion.

In this model, the surface F0 contour is used as the asymptotic approximation of the underlying pitch target, which is defined as the smallest unit. The host unit of a pitch target is assumed to be syllable. So a continuous speech is divided into syllables with a known syllable boundary (beginning and end), which is labeled by "double-threshold" method in this paper [15]. First, we make preliminary annotation with "double-threshold" method, and then modify it by the repeated artificial perception with Praat tool.

Let the syllable boundary be [0, D]. The Pitch Target model is denoted as the following equations,

$$T(t) = at + b \tag{21}$$

$$\begin{aligned} y(t) &= \beta \exp(-\lambda t) + at + b \\ 0 &\le t \le D, \lambda \ge 0 \end{aligned} \tag{22}$$

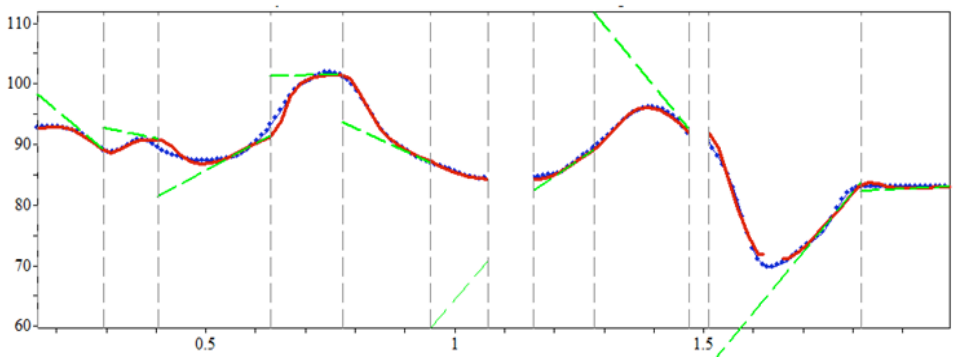


Figure 3. F0 parameters extraction of an example speech with the pitch target model.

where parameter *a* and *b* are the slope and intercept of the underlying pitch target, respectively, and they describe an intended intentional goal of the speaker; parameter β presents the distance between F0 contour and the underlying pitch target at $t = 0$; Parameter λ is a positive number representing the rate of decay of the exponential part. In other words, it describes how fast the underlying pitch target is approached. The greater the value of λ is, the faster the speed. $T(t)$ is the underlying pitch target, and $y(t)$ is the surface F0 contour. As mentioned above, the F0 contour of a speech is presented by a series of (a, b, β, λ), where the horizontal axis denotes time (s), the vertical axis denotes the F0 parameter (Hz). The dotted curve is the real F0 contour, the solid one is fitting F0 contour in pitch target model (22) and the dashed one is approximation result (21). In (22), with the increase of time *t*, exponential term $\beta \exp(-\lambda t)$ gradually approaches to zero, and it is approximately equivalent to (21). Namely, in theory, the end of dotted curve and solid curve are nearly overlapping of each syllable. Indeed, the result in Figure 3 shows that feature extraction result is comparably ideal.

3) Feature Fusion

Serial fusion: It splices multi-feature parameters vectors into a single vector simply. If two feature vectors are α and β, the vector after fusion is denoted as follows.

$$\gamma = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{23}$$

Because it only splices vectors, the vector dimension after fusion is very high and it includes some interference and redundancy information. So some dimension reduction methods are proposed to reduce computation complexity.

Parallel fusion: It splices different features in complex form, as follows [20].

$$\gamma = \alpha + \beta i \tag{24}$$

Kernel function fusion: It maps original linear inseparable feature space R^n to high dimensional linear separable space H. Kernel function is always denoted as inner product form,

$$K(\alpha_i, \alpha_j) = < \varphi(\alpha_i), \varphi(\alpha_j) > \tag{25}$$

The kernel function after the feature fusion,

$$\gamma = \mu_1 K(\alpha_i, \alpha_j) + \mu_2 K(\beta_i, \beta_j) \tag{26}$$

where μ_1 and μ_2 are consolidation coefficients. In general, principal component analysis (PCA) and linear discriminant analysis (LDA) are used to reduce dimension.

This paper focuses on comparing voice conversion performance between only extracting Mel-cepstra and feature combination. So the simple and direct serial fusion form is enough. The fusion of Mel-cepstra, F0 and formant frequency is applied to voice conversion. In further work, we can apply another two feature fusion forms in our system.

B. GMM Considering Dynamic Feature

In [5], Toda proposed a method of combining static and dynamic features to include inter-frame correlations into model training and parameter generation algorithm, and it is called as D-GMM. In this paper, we select D-GMM as our voice conversion model based on a sequence of experiments.

Let $x = [x_1^T, x_2^T, ..., x_T^T]^T$ and $y = [y_1^T, y_2^T, ..., y_T^T]^T$ be the acoustic feature vectors characterizing the speech produced by source speaker and target speaker, respectively. $x_t = [x_{St}^T, \Delta x_{St}^T]^T$ and $y_t = [y_{St}^T, \Delta y_{St}^T]^T$ are the acoustic vectors at frame *t*, where x_{St} and y_{St} are static features, Δx_{St} and Δy_{St} are dynamic features. The speech parameter sequence *x*, *y* can be calculated from static parameter sequence of source speaker $x_S = [x_{S1}^T, x_{S2}^T, ..., x_{ST}^T]^T$ and target speaker $y_S = [y_{S1}^T, y_{S2}^T, ..., y_{ST}^T]^T$ by $x = Wx_S$ and $y = Wy_S$, where *W* is a matrix determined by the calculation function of dynamic features [5]. The training of mapping function is similar to that of the conventional GMM. The specific training process and *W* are introduced in [5].

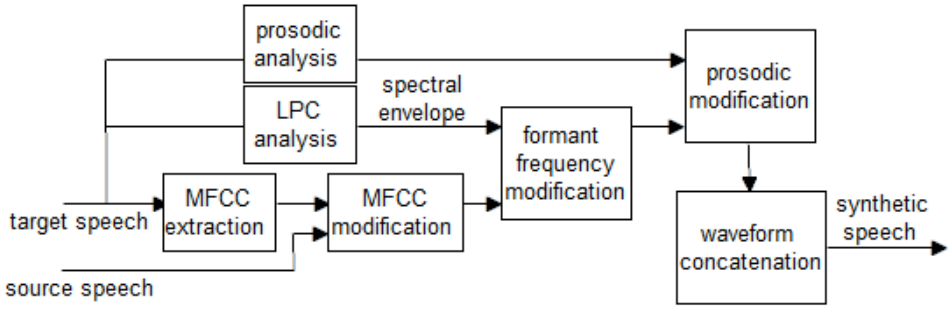


Figure 4. The speech synthesis process modifying Mel-cepstra, prosodic feature and formant frequency.

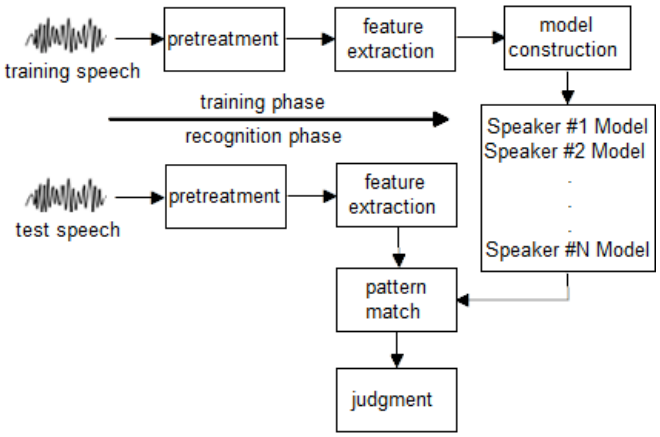


Figure 5. Speaker recognition implementation process.

C. *Voice Synthesis*

In this paper, we use the synthesizer of STRAIGHT algorithm. Since besides Mel-cepstra, we also select F0 and formant frequency, we modify the synthesizer process as shown in Figure 4.

D. *Voice Conversion Evaluation Criterion*

In order to validate the performance of our voice conversion algorithm, we evaluate it from two perspectives: objective and subjective evaluation. The latter includes some subjective factors and requires a certain amount of listeners, and comparatively the objective numerical comparison is believed to be more convincing. For different cases, we introduce two objective evaluation methods as follows.

1) *Mel-cepstral Distortion*

The Mel-cepstral distortion between the target and converted Mel-cepstra in the evaluation set given by the following equation is used as the objective evaluation measure,

$$Mel - CD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(mc_d - mc'_d)^2} \quad (27)$$

where mc_d and mc_d' are the d-th cofficients of the target and converted Mel-cepstra, respectively.

The source-target speaker voice conversion is a complete process and the final aim is to synthesize a speech with high identification and articulation. Mel-CD is an evaluation criterion focuses on 2nd step, voice conversion model. So we introduce speaker recognition system to evaluate synthetic speech, judging whether it has acoustic feature of target speaker.

2) *Speaker recognition*

Speaker recognition system can identify the speaker from a registered finite set of speakers through a speech [17,18]. Figure 5 is the schematic diagram of speaker recognition. As shown in Figure 5, speaker recognition is a pattern identification process that is divided into train and identification phase. In the train phase, we build a given model (GMM model in section II) for each user by collecting speech data and related algorithm processing. In identification phase, extracting acoustic feature (Mel-cepstra and its Deltas in section II) of test speech, and comparing it with model parameters in model library. The minimum distance or maximum probability score one is the recognition result according to similarity criterion or probabilistic likelihood comparison.

TABLE I. RECOGNITION RATE OF TWO ACOUSTIC FEATURE SELECTION IN SPEAKER RECOGNITION SYSTEM (MALE-MALE).

	Only Mel-cepstra	Feature Combination
group 1	80%	88%
group 2	76%	86%
group 3	78%	92%

TABLE II. RECOGNITION RATE OF TWO ACOUSTIC FEATURE SELECTION IN SPEAKER RECOGNITION SYSTEM (MALE-FEMALE).

	Only Mel-cepstra	Feature Combination
group 1	92%	96%
group 2	90%	96%
group 3	94%	98%

In this paper, we input the converted voice to the designed speaker recognition system and find out that the voice is identified as the source or target speaker. In this way, we can know the similarity of converted voice and target voice.

IV. EXPERIMENTS AND EVALUATIONS

We used a Mandarin speech database to evaluate the performance of voice conversion. Two females and two males parallel speech are chosen as our corpus. A sample frequency 16KHz and sample rate 16Bits are used. The duration of training speech is about 3-4s. Since the acoustic features combination and voice model may affect synthetic speech quality directly, we test these two aspects as follows first, and then test the conversion performance of the whole voice conversion process proposed in this paper.

A. Acoustic Feature Selection Comparison

First, we compare the performance that only Mel-cepstra is considered with the result that the feature combination extraction is conducted. For each case, we make 3 group experiences, and each one contains 50 test utterances. In this experiment, the input synthesized speech (test speech) is input into a speaker recognition system, and matches with source speech and target speech, respectively. The final aim is to judge the converted speech is closer to which one. The conversion is effective if converted speech is closer to the target than the source. The within-gender recognition success rate (male to male) of two cases is shown in Table I. And the cross-gender recognition success rate (male to female) of two cases is shown in Table II. Numerically, feature combination form can get higher recognition rate, namely synthetic speech have the similar voice individuality of target one. Mel-cepstra cannot present speech personalized feature comprehensively. The combination form contains prosodic feature, spectrum parameter and spectral envelope characteristic, almost covering all acoustic features. It can distinguish different speakers well.

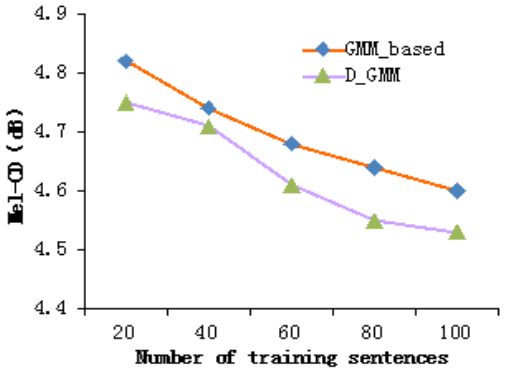
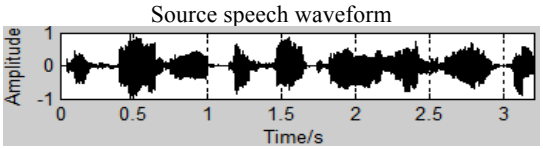
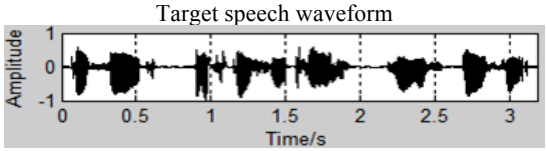


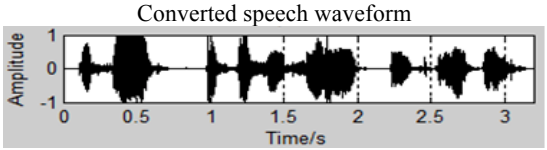
Figure 6. Mel-cepstral distortion as a fuction of the number of training sentences with GMM and D_GMM. The average Mel-CD between source and target is 8.29dB.



(a)



(b)



(c)

Figure 7. The source, target and converted speech waveform of an example speech.

B. Two GMMs Model Performance Contrast

The average Mel-CD between the converted and target feature sequence is used to evaluate the performance of voice conversion objectively. Mel-cepstra extraction is introduced above. The number of GMM components is set to be 64. The dimension of complete feature vectors of source and target speakers for GMM training is 48 including 24 mel-cepstrums and their Deltas. Figure 6 shows Mel-cepstral distortions in the evaluation set as a function of the number of training sentences. In Figure 6, the horizontal axis denotes the number of training sentences, the vertical axis denotes Mel-CD. The D_GMM significantly outperforms the conventional GMM. This is because D_GMM realizes an appropriate parameter trajectory by considering the interframe correlation that is ignored in the conventional method.

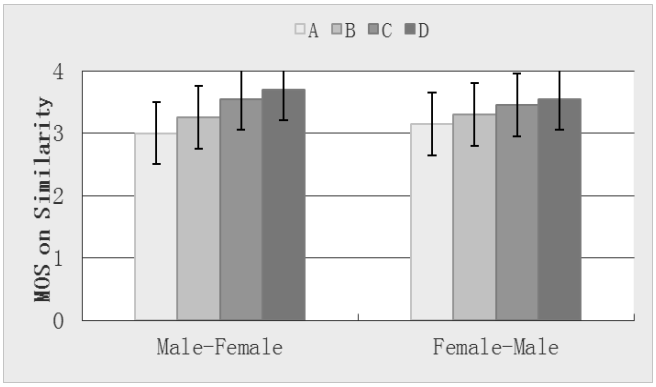


Figure 8. Mean opinion scores on similarity with 95% confidence intervals for the four systems above.

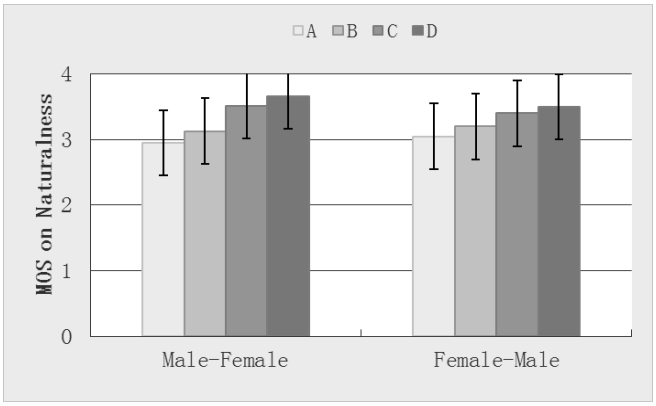


Figure 9. Mean opinion scores on naturalness with 95% confidence intervals for the four systems above.

The above two experiments A and B, proving that three features combination has better performance than only Mel-cepstra extraction in voice synthesis, and D_GMM outperforms traditional GMM.

C. Voice Conversion Performance

Finally, we converted the test speech into a target speech with the proposed voice target conversion process scheme in this paper, its source speech, target speech and converted speech waveform are presented in Figure 7 (a), (b), and (c), respectively. Intuitively, converted speech waveform is much closer to target speech waveform rather than the source one. With feature combination extraction and D_GMM model, recognition rate reaches nearly 90%, indicating our work is feasible.

D. Subjective Evaluation

A subjective listening test was conducted to evaluate the similarity and naturalness of converted speech using different methods:

***A*:** GMM trained only with Mel-cepstra;
***B*:** GMM trained with feature combination;
***C*:** D_GMM trained only with Mel-cepstra;
***D*:** D_GMM trained with feature combination;

Fifty sentences from the test set were converted by the four systems list above. Six listeners, all of whom were graduate students specializing in speech synthesis, took part in the test. They were asked to give a 5-scale opinion score (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) on the similarity and naturalness of each converted sentence they heard. All the sentences were played in random order.

Figure 8 shows the mean opinion scores (MOS) on similarity of converted speech with the target speech. Figure 9 shows the MOS on naturalness of converted speech. We can see that the proposed D_GMM method outperformed the conventional method on both similarity and naturalness. Comparing with GMM, the difference between the performance of D_GMM using only Mel-cepstra and feature combination was less significant. This is consistent with what we found in the objective evaluation as Figure 6 and Table II.

V. CONCLUSION

In this paper, a comprehensive analysis on source-target speaker voice conversion is conducted based on three key steps, including acoustic features selection and extraction, voice conversion model construction, and target speech synthesis, and a complete and optimal source-target speaker voice conversion is proposed. Firstly, in order to overcome feature selection with no representation, Mel-cepstra, F0 and formant frequency as a combination in voice conversion are selected. Secondly, in order to decrease the discontinuity between frames, the dynamic features are implemented into GMM model, and the objective experimental result shows that D_GMM significantly outperforms the conventional GMM. Thirdly, the synthesizer process is modified to fit more feature parameters successively. Finally, the speaker recognition system is applied to evaluate the converted speech, and the system recognition rate proves the effectiveness of our method. Nearly 90% converted speech is recognized as the target speaker.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No.61305025, 61300022), and the Fundamental Research Funds for the Central Universities (N13050411).

REFERENCES

[1] N. Campbell. “Perception of Affect in Speech - Toward an Automatic Processing of Paralinguistic Information in Spoken Conversion,” in *Proc. ICSLP*, 2004, pp.881-884.

[2] Z. Lingli. “Voice Conversion Technology Research based on GMM,” Nanjing University of Posts and Telecommunications, vol. 33, pp.1-2, 2011.

[3] H. Matsumoto and S. Hiki. “Multidimensional Representation of Personal Quality of Vowels and its Acoustical Correlates,” IEEE Transactions on Audio Electroacoustics, vol. 21, no. 4, pp. 428-436, 1973.

[4] Y. Hui and S. Young. "Quality-enhanced Voice Morphing using Maximum Likelihood Transformations," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 1301-1312, 2006.

[5] T. Toda, A.W. Black and K. Tokuda. "Voice Conversion Based on Maximum-likelihood Estimation of Spectral Parameter Trajectory," IEEE Trans. Audio Speech and Language Processing, vol. 15, no. 8, pp. 2222-2235, 2007.

[6] T. Jianhua and K. Yongguo. "Prosody Conversion From Neutral Speech to Emotioanl Speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 56-68, 2006.

[7] Y. Stylianou and E. Moulines. "Continuous Probabilistic Transform for Voice Conversion," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 552-570, 1998.

[8] M. Schroder. "Emotional Speech Synthesis," Proceedings of the 7th European Conference on Speech Communication and Technology Eurospeech, 2001, pp. 561-564.

[9] T. Toda, H. Sruwatari and K. Shikano. "Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT Spectrum," Proc. ICASSP, 2001, pp. 841-844.

[10] Z. Jing, F. Mingand and F. Wenquan. "Speaker Feature Extraction Algorithm Improvement Based on MFCC," Electro Acoustic Technology, pp. 61-64, 2009.

[11] E. Moulines and Y. Sagisaka. "Voice conversion: State of the Art and Perspectives," Speech Commun, vol. 16, no. 2, pp. 125-126, 1995.

[12] A.P. Dempster, N.M. Laird and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, pp. 1-38, 1977.

[13] H. Bin. "Chinese Voice Conversion Research Based on Speech Recognition and Speech Synthesis," YunNan University, pp. 14-17, 2013.

[14] W. Ning. "Voice Emotion Conversion Based on Pitch Target Model and Prosody Feature Modification," SuZhou University, pp. 25-33, 2012.

[15] S. Zhiyong. "Speech Signal Analysis and Synthesis Application in Matlab," BeiHang University Press, pp. 262-266, 2013.

[16] Y. Xu and Q.E. Wang. "Pitch Targets and Their Realization: Evidence from Mandarin Chinese," Speech Commun, vol. 33, pp. 319-337, 2001.

[17] D.A. Reynolds, T.F. Quatieri and R.B. Dunn. "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 33, pp. 19-41, 2000.

[18] M.A. Pathak and B. Raj. "Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models," IEEE Trans. Audio Speech and Language Processing, vol. 28, pp. 39, 2013.

[19] A. Kain and M.W. Macon. "Spectral Voice Conversion for Text-to-Speech Synthesis," in Proc. ICASSP, 1998, pp. 285-288.

[20] J. Yang, J.Y. Yang and D. Zhang. "Feature Fusion: Parallel Strategy vs. Serial Strategy," Pattern Recognition, vol. 36, no. 6, pp.1369-1381, 2003.