JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 30, 1149-1166 (2014)

Frequency Warping for Speaker Adaptation in HMM-based Speech Synthesis

WEIXUN GAO[1] AND QIYING CAO[1,2]
[1]*School of Information Science and Technology*
[2]*College of Computer Science and Technology*
Donghua University
Shanghai, 200051 P.R. China

Speaker adaptation in speech synthesis transforms a source utterance to a target utterance that differs from the source in terms of voice characteristics. In this paper, we employ vocal tract length normalization, which is generally used in speech recognition to remove individual speaker characteristics, to speaker adaptation in speech synthesis. We propose a frequency warping approach based on a time-varying bilinear function to reduce the weighted spectral distance between the source speaker and the target speaker. The warped spectra of the source speaker are then converted to line spectrum pairs to train hidden Markov models (HMM). HMMs are further adapted by algorithms based on maximum likelihood linear regression with the target speaker's data. The experimental results show that our frequency warping approach can make the warped spectra of the source speaker closer to the target speaker, and the resultant adapted HMMs perform better than the HMMs trained by unwrapped spectra in terms of synthesized speech naturalness and speaker similarity.

Keywords: frequency warping, VTLN, speaker adaptation, HMM-based speech synthesis, MLLR

1. INTRODUCTION

Nowadays more and more human-computer interaction applications and services require the text-to-speech (TTS) synthesis system to produce high-quality synthesized speech with various personalized voices. Generally, the state-of-the-art TTS synthesis system can deliver only a single voice speech, since the recording speech database is collected from one person. The database is usually large and recorded with a well-controlled recording environment, speaking style, and recording text. Building such a database is quite a time-consuming and high-cost process. Voice transformation is a technique that can convert the voice of speaker A (source speaker) to that of speaker B (target speaker) with limited recording speech of the target speaker. It modifies the voice characteristics of the source speaker's speech so that it can be perceived as if it were uttered by the target speaker without losing any information or message delivered by the source speaker's speech. It is used as one of the key technologies in many applications, such as customizing a voice needed for a smart home system, simulating dialogues between various characters in a language learning system, personalizing output in a speech-to-speech translation system, and converting electrolaryngeal speech to normal speech for laryngectomees in a speaking-aid system, among others.

Received March 1, 2012; revised May 13 & August 7, 2012; accepted September 4, 2012.
Communicated by Hsin-Min Wang.

Voice transformation systems are roughly based on two major approaches: voice conversion and speaker adaptation. Voice conversion transforms the spectrum and the pitch of the source speaker's speech to match those of the target speaker. The statistical approach in voice conversion has significantly improved the quality of converted speech in the past few decades. Voice conversion first use parallel utterance pairs of the source and target speakers to train a Gaussian mixture model (GMM), then convert the source speaker's parameters into those of the target speaker in a minimum-mean-square error or maximum likelihood (ML) sense, and finally generate speech from converted parameters by a source-filter speech product model [1, 2]. Voice conversion is generally used to convert the speech synthesized by a unit selection–based TTS system [3]. Speaker adaptation adjusts the parameters of the source speaker's model to be close to the target speaker's model so that the generated speech sounds like the target speaker. Since speech synthesis based on hidden Markov models (HMM) [4] can deliver high-quality synthesized speech, the speaker adaptation techniques that are originally developed for HMM-based speech recognition, *e.g.*, maximum a posterior, maximum likelihood linear regression (MLLR), constrained MLLR (CMLLR), and speaker adaptive training, are extended to adapt HMMs in speech synthesis [5].

It is well known that the variability in inter-speaker voice characteristics is mainly caused by vocal tract length, mouth dimension, nasal cavity, accent, and speaking rate. Statistical approaches to voice transformation try to convert all aspects of inter-speaker difference. Vocal tract length normalization (VTLN) aims to compensate for the vocal tracts of different sizes. It is generally implemented by warping the frequency axis of one spectrum by expanding or compressing in different regions to minimize the distance to another spectrum [6]. There are many frequency warping approaches applied in speech recognition and voice transformation [6-16, 29, 30]. The warping functions can be linear, piecewise linear, bilinear, or nonlinear. A bilinear warping function is used and the warping factor is obtained by a grid search based on ML criterion [13] or by minimizing log spectral distance [15]. A piecewise linear frequency warping function generated by mapping formants of the source and target speakers is employed to voice conversion [8] and speaker adaptation [14]. To solve the over-smoothing problem that results in the degradation of converted speech quality, dynamic frequency warping, which minimizes the distance between the source and the target spectra by a dynamic programming technique, and an alternative method named weighted frequency warping are combined with the GMM-based approach for voice conversion [9-11]. However, these methods generally achieve high performance on speech naturalness but low performance on speaker similarity.

This paper presents a study on frequency warping for speaker adaptation in HMM-based speech synthesis. It involves an initialization of speaker adaptation by frequency warping between speakers' spectra and a further integration with MLLR-based speaker adaptation of HMMs. We propose a frequency warping approach based on a time-varying bilinear function to reduce the weighted spectral distance between the source and the target speakers. This approach results in a better initialization for MLLR – based speaker adaptation. The experimental results demonstrate the effectiveness of the approach, *i.e.*, a better performance than the conventional approach in terms of naturalness and speaker similarity.

2. FREQUENCY WARPING FOR SPEAKER ADAPTATION

The speaker adaptation of HMMs is done iteratively. The initialization has a critical impact on the performance of the final synthesized speech. We propose an approach of GMM-based dynamic computation of warping factor to improve the performance of the initialization based on frequency warping. Fig. 1 illustrates our speaker adaptation system, where the frequency warping method is first applied to warp the source speaker's spectra towards the target speaker, then HMMs are retrained with the warped features of the source speaker. Finally, the retrained model is adapted with the target speaker's features by MLLR-based algorithms. Compared with the conventional systems, our system (1) employs a criterion of minimizing weighted log spectral distance between the source speaker and the target speaker, which is perceptually critical for voice characteristics, instead of using ML to transform the source speaker's features; (2) performs a smooth transform over both frequency and time domains by a bilinear warping function with frame-dependent warping factors; and (3) retrains the source speaker's HMMs to get a better initialization for further adaptation.

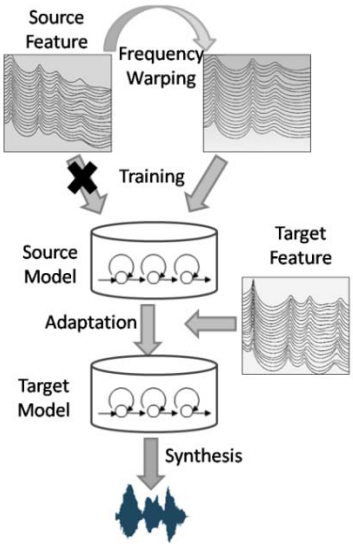


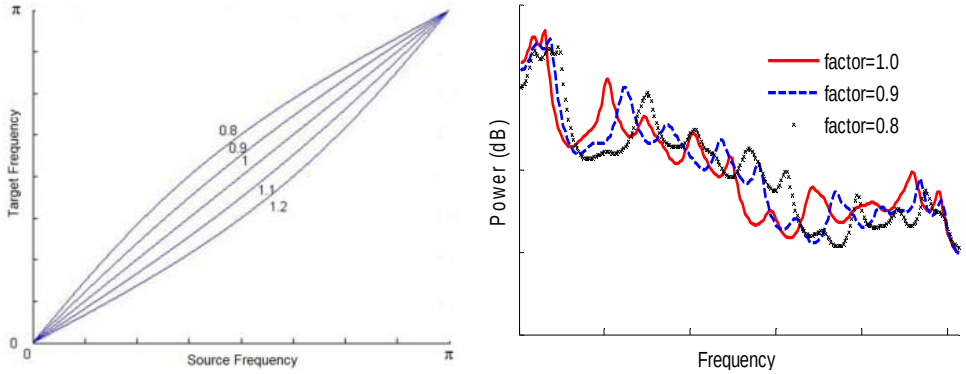
Fig. 1. Diagram of our speaker adaptation system.

2.1 VTLN by Spectral Frequency Warping

VTLN is generally implemented by warping the frequency axis of one spectrum through expansion or compression in different regions to minimize the distance to another spectrum. If the vocal tract is seen as a uniform tube, formant frequency positions are inversely proportional to vocal tract length and hence the linear warping function can be applied. In reality, however, the vocal tract is more complex than a uniform tube [16]. A piecewise linear function, a bilinear function, or more sophisticated functions can be used to warp frequencies for VTLN. Our frequency warping approach is based on a bilinear warping function,

$$\Psi_{\alpha}(\omega)=\omega+2\tan^{-1}\left(\frac{(1-\alpha)\cdot\sin(\omega)}{1-(1-\alpha)\cdot\cos(\omega)}\right) \tag{1}$$

where ω is the spectral frequency for warping and α is the warping factor. Fig. 2 (a) shows a schematic plot of a bilinear warping function performance. Fig. 2 (b) depicts examples of warped spectra with warping factors 0.8, 0.9, and 1.0, where a factor equal to 1 means the original spectrum without warping and less than 1 means spectral frequencies are expanded correspondingly.



(a) (b)

Fig. 2. (a) A schematic plot of a bilinear warping function performance; (b) samples of warped spectra with different warping factors.

A single warping factor for all the utterances of a speaker is effective in VTLN. In [12], different frequency warping factors are employed for different acoustic classes due to the fact that all acoustic classes do not reveal the similar spectral variation caused by physiological differences. We also investigate the spectral variations of different phones. We use long vowel parts of speech from two speakers to represent their spectral variation difference. Five main vowels, each segment (about 40ms) with the same phone contexts, are selected. We estimate a warping factor of a bilinear function for each vowel pair by a grid search in the sense of minimizing spectral distance. Table 1 shows the optimal warping factors *a* for five vowels from a male to a female and from a female to a female, indicating that the spectral variations of these five vowels are distinctive. In a female-to-female conversion as shown in Table 1 (b), the spectral frequency axis needs to be compressed for vowels /i:/, /ə:/, and /o:/, while extended for vowels /a:/ and /u:/ according to the warping factors, so a single bilinear warping function for all spectra seems inappropriate for a female-to-female conversion.

Table 1. Five long vowels and their corresponding warping factors.

(a) Male to female

Vowels	/a:/	/i:/	/u:/	/ə:/	/o:/
Factors	0.90	0.88	0.85	0.89	0.93

(b) Female to female

Vowels	/a:/	/i:/	/u:/	/ə:/	/o:/
Factors	0.98	1.01	0.99	1.03	1.02

2.2 Our Approach to Frequency Warping

Assigning different warping factors to the spectra of different phones should be more effective than having a single warping factor for all spectra, according to the discussion in section 2.1. We propose a frequency warping approach based on a time-varying bilinear function to reduce the weighted spectral distance between the source speaker and the target speaker. The parallel spectra from the source and the target speakers are used to train a GMM, where a bilinear warping factor is estimated for each Gaussian component by minimizing the weighted spectral distance. The weight assigned to spectral distance is correlated with spectral formant, which is sensitive to the perception of speaker characteristics [25]. The warping factor for each frame of the source speaker is dynamically generated from the trained GMM. Fig. 3 illustrates the procedure of warping factor estimation in our approach with the following steps:

1. Prepare the parallel utterance pairs from the source and the target speakers.
2. Apply dynamic time warping (DTW) to time-align the source and the target speakers' features, where the cost function (or distance) for DTW is the minimizing weighted likelihood ratio (D_{WLR}), a weighted log spectral distance.
3. Use parallel feature pairs to train a GMM by maximizing weighted likelihood training.
4. Find a warping factor for each GMM component by a full grid search by minimizing the weighted distance between the source and the target speakers' spectra associated with the mean vector of that component.
5. Warp each frame of the source speaker by the bilinear warping function with the warping factor dynamically estimated by our approach.
6. Repeat steps 2 to 5 until the overall weighted distance between the warped source spectra and the target spectra for convergence is met.

If the parallel utterance pairs are not available in the above step 1, we suggest (a) using a TTS system built by source speaker data to synthesize speech (features) with the same texts as those of target speaker's; or (b) finding the nearest frame from source features for each target frame to generate the parallel feature pairs [26].

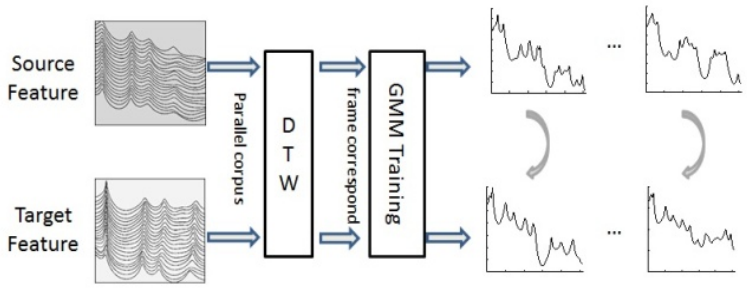


Fig. 3. The procedure of warping factors estimation by a GMM-based approach.

The cost function for DTW in step 2 is to minimize the weighted likelihood ratio [27], D_{WLR}, as

$$D_{WLR}\left(\log P_s(\omega), \log P_g(\omega)\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\log P_s(\omega) - \log P_g(\omega))(P_s(\omega) - P_g(\omega))d\omega, \quad (2)$$

where $P_s(\omega)$ and $P_g(\omega)$ are the spectra of the source speaker and the target speaker, respectively. The D_{WLR} between two spectra weights more on spectral peaks than on spectral valleys. It is consistent with human perception for speech formants (acoustic resonances of the human vocal tract). The time-align source and the target speakers' features by this cost function should emphasize formant alignment more than other cost functions. Eq. (2) can be seen as a weighted log spectral distance, which puts more weights on spectral frequencies associated with formants. It also can be seen as a Kullback-Leibler distance, a measure of (dis)similarity between two probability distributions in probability and information theory, between the two spectra. WLR has been applied to noisy speech recognition [17]. However, the computational complexity of D_{WLR} between two spectra is high, so we use a weighted line spectrum pair (LSP) distance to approximate it, as follows:

$$d_{\omega} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_t(i)(\omega_{s,t}(i)\text{-}\omega_{g,t}(i))^2}, \quad (3)$$

$$w_t(i) = \frac{1}{\omega_{g,t}(i) - \omega_{g,t}(i-1)} + \frac{1}{\omega_{g,t}(i+1) - \omega_{g,t}(i)}, \quad (4)$$

where $\omega_{s,t}(i)$ and $\omega_{g,t}(i)$ are the ith order of the tth frame LSP features of the source speaker and the target speaker, respectively. N is the order (dimensions) of the LSP vector used for distance calculation. LSP is a way of uniquely representing the linear predictive coding (LPC) coefficients. It has good properties, *e.g.*, good interpolation and robustness to quantization [31], for modeling. Another good property of LSP is that a cluster of two or more LSPs indicates a spectral peak (formant) [19]. Therefore, we use the inverse of the distance between adjacent LSPs as weights in LSP distance computation, as in Eq. (4). We minimize the spectral distance between the source and the target speakers by warping the spectral frequency of the source speaker, so the target speaker's spectrum is the goal and the adjacent LSPs of the target speaker are used to estimate weights in Eq. (4). Similarly, the inverse harmonic mean weighting function, same as in Eq. (4), is used for vector quantization in speech coding [18] and directly applied to spectral parameters in HMM-based speech synthesis [19]. The perceptually sensitive spectral information (formants) is located mainly in the frequency range below 4kHz and LPC analysis is not stable at higher frequency. We only use those LSPs, $\omega_{s,t}(i)$ and $\omega_{g,t}(i)$, below 4kHz for Eq. (3).

To be consistent with the cost function of DTW, we employ maximizing weighted likelihood to GMM training in step 3. The objective function

$$Q = \sum_t \{\log \sum_m c_m \prod_i [N(z_t(i), \mu_m(i), \Sigma m(i))]^{w_t(i)}\} \quad (5)$$

is maximized, where $z_t = [\omega_{s,t}^{\mathrm{T}}, \omega_{g,t}^{\mathrm{T}}]^{\mathrm{T}}$ is a parallel feature pair of the tth frame composed of a source feature vector $\omega_{s,t}$ and a target feature vector $\omega_{g,t}$; μ_m and Σ_m are the mean vector and covariance matrix of the mth Gaussian component, respectively. Here the

diagonal covariance matrix is used. The term $w_t(i)$ is defined in Eq. (4). We apply the expectation-maximization (EM) algorithm to obtain GMM parameters as follows:

$$\mu_m(i) = \frac{\Sigma_t L_{m,t} z_t(i)}{\Sigma_t L_{m,t}}, \tag{6}$$

$$\Sigma m(i) = \frac{\sum_t L_{m,t}(z_t(i) - \mu_m(i))(z_t(i) - \mu_m(i))^{\mathrm{T}}}{\Sigma t Lm,t}, \tag{7}$$

$$c_m = \frac{\sum_t Lm,t}{\Sigma_t \Sigma_m L_{m,t}}, \tag{8}$$

$$Lm,t = \frac{c_m \prod_i \left[N(z_t(i),\mu_m(i),\Sigma m(i))\right]^{w_t(i)}}{\sum_m c_m \prod_i \left[N(z_t(i),\mu_m(i),\Sigma m(i))\right]^{w_t(i)}}, \tag{9}$$

where i, t, and m are the LSP dimension index, frame index, and Gaussian component index, respectively.

After GMM training, each frame, *e.g.*, the tth spectrum associated with $\omega_{s,t}$ of the source speaker is warped by the bilinear warping function with the factor $\hat{\alpha}_t$ obtained by

$$\hat{\alpha}_t = \sum_m \left(\frac{c_m N(\omega_{s,t}; \mu_{s,m}, \Sigma_{s,m})}{\Sigma_m c_m N(\omega_{s,t}; \mu_{s,m}, \Sigma_{s,m})} \right) a_m, \tag{10}$$

Where c_m, $\mu_{s,m}$ and $\Sigma_{s,m}$ are the weight, mean vector and covariance matrix of the mth Gaussian component of the source speaker, respectively. The bilinear warping function factor for the mth Gaussian component is expressed as a_m (m = 1, …, M), obtained by

$$\hat{\alpha}_m = \arg\min_{\alpha_m} D_{WLR}(\log \Psi_{a_m}(P_{s,m}(\omega)), \log P_{g,m}(\omega)). \tag{11}$$

Where $P_{s,m}(\omega)$ and $P_{g,m}(\omega)$ are the source and the target speakers' spectra associated with the mean vector of mth Gaussian component, respectively, and Ψa_m is a bilinear warping function with warping factor, a_m, defined in Eq. (1). $F0_{s,t}$ for the tth frame of source speaker is also adjusted according to

$$\widehat{F0}_{s,t} = \frac{(F0_{s,t} - u_s)}{\sigma_s} \cdot \sigma_g + u_g, \tag{12}$$

where u_S, u_g, σ_s and σ_g are the mean and the standard deviation of the source and the target speakers' F0s. The warped spectra and the adjusted F0s of the source speaker are hence used for retraining HMMs, as illustrated in Fig. 1.

2.3 Speaker Adaptation of HMM-based Speech Synthesis

The HMMs retrained by the warped spectra and the adjusted F0s should be closer to

the target speaker than the models trained by the unwarpped features would be. The frequency warping only reallocates the source speaker's spectral formants on the frequency axis, while the formant bandwidth, the spectral tilt and the spectral intensity are almost unchanged. We further adapt the HMMs by MLLR and CMLLR [20, 21]. In MLLR adaptation, the mean vectors of Gaussian mixture components of HMMs are transformed as

$$\hat{\mu} = A\mu + b, \tag{13}$$

where A is the transform matrix, b is the bias vector, μ is the mean vector of warped model. And $\hat{\mu}$ is the adapted mean vector. In CMLLR adaptation, both mean vectors and covariance matrices of Gaussian mixture components of HMMs are transformed as

$$\begin{aligned} \hat{\mu} &= A\mu + b, \\ \hat{\Sigma} &= \Sigma A \Sigma^{T}, \end{aligned} \tag{14}$$

where μ and $\sum$ are the mean vector and the covariance matrix, respectively. The transform matrix A is estimated by maximizing likelihood of adaptation data O from the target speaker,

$$\hat{A} = \arg\max_{A} P(O \mid \lambda, A), \tag{15}$$

where λ is the parameter set of warped HMMs. Transform matrix A is estimated by the Baum-Welch algorithm.

3. EXPERIMENTS AND RESULTS

The experiments in this paper are performed on the CMU ARCTIC database [22], which is designed for speech synthesis.

3.1 Experimental Setup

Two speech corpora with the voices of two U.S. English speakers, a male (bdl) and a female (clb), are used as the source speakers' data. Another speech corpus with the voice of another U.S. English female speaker (slt) is used as the target speaker's data. The data of each source speaker consist of 900 training, 100 developing, and 100 testing utterances. The target speaker's data are made up of 100 adaptation and 100 testing utterances selected from the target speaker's corpus. The transcriptions of the source speakers' developing and testing data are same as the target speaker's adaptation and testing data. Speech signals in both corpora are sampled at 16 kHz, windowed by a 25-ms window with a 5-ms shift. *F*0s and spectral envelopes are estimated by STRAIGHT [23]. The LPCs of 40th order are transformed into static LSPs and their dynamic counterparts. Five-state, left-to-right HMM phone models, where each state is modeled by a single Gaussian, diagonal covariance output distribution, are adopted to train the source speaker's HMMs. Richer phonetic and prosodic contexts are used to capture the co-articulation effects in HMM modeling. We use state tying via a clustered decision tree to

cluster long context models into generalized ones to predict unseen context in test robustly. The HMM training was done by HTS toolkit [28].

As described in section 2, aligned utterance pairs from the source speaker's developing data and the target speaker's adaptation data are used to train a GMM for getting time-varying warping factors. The source speaker's training data are warped by a bilinear function. The HMMs trained by the warped features of the source speaker are further adapted by the target speaker's adaptation data. In the MLLR and CMLLR adaptation, a regression class tree is constructed to classify the Gaussians in the model set into many classes, and a set of transformations is estimated according to the amount of adaptation data. The Gaussians, which are close in acoustic space, are grouped into a class and share a transform matrix. For our adaptation experiment, we need to set the number of the classes, the splitting threshold for each node of the tree, the block size, and the bandwidth of the transform matrix. To get better performance with the limited adaptation data, a global transform is generally constructed to the model set firstly, *i.e.*, all Gaussians of the model set share one transform matrix, then more transform matrices are estimated to adapt the model set based on the global transformation.

We run the experiments with different configurations, which are listed as follows:

- CFW: Conventional frequency warping method, where all frames share one warping factor obtained by a full grid search.
- DFW: Our frame-dependent frequency warping method, as described in section 2.2, to warp the source speaker's spectra.
- FUW+MLLR: HMMs are trained by the unwarpped features, *i.e.*, without spectral frequency warping and $F0$ adjustment, of the source speaker and further adapted by MLLR with the adaptation data of the target speaker.
- DFW+MLLR (CMLLR): HMMs are trained by the warped features of the source speaker and further adapted by MLLR (and CMLLR) with the adaptation data of the target speaker.

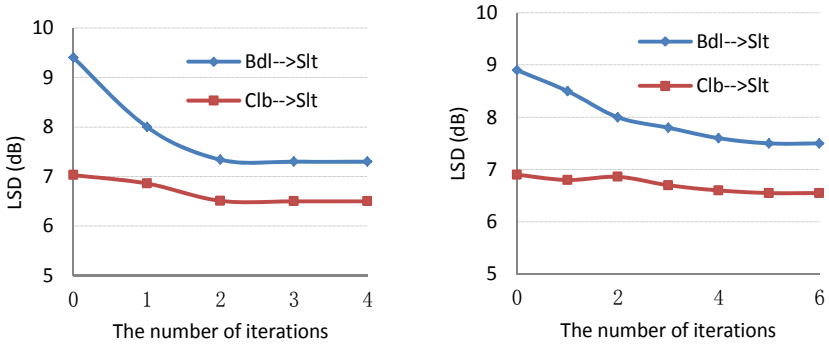
3.2 Evaluation Results and Analysis

3.2.1 Evaluation for warping function

Fifty aligned utterance pairs from the source speaker's developing data and the target speaker's adaptation data are used to estimate bilinear warping function by the conventional method (CFW) and our method (DFW). The resultant warping functions are performed on the source speakers' testing data. The log spectral distance (LSD) frame-by-frame between the spectra of the source and the target speakers' aligned testing data is calculated to show the performance of CFW and DFW. Although we use weighted spectral distance, *i.e.*, weighted likelihood ratio (D_{WLR}), as the objective function in our proposed approach, here we use LSD instead of D_{WLR} to show performance since LSD in dB scale is intuitive and generally used by measuring speech quality. Table 2 lists the average LSD by CFW and DFW. It shows that the original LSD between the source and the target speakers, *i.e.*, without frequency warping (or warping factor $a = 1$), is 9.4 dB for Bdl to Slt and 7.03 for Clb to Slt. The frequency warping can reduce the spectral distance between the source and the target speakers and result in these two speakers being

close or similar. Our frequency warping approach outperforms the conventional approach, *i.e.*, the average LSD between the target speaker's spectra and the source speaker's warped spectra using our approach is 0.89 dB and 0.31 dB lower than the conventional one.

Fig. 4 (a) shows the average LSD of the warped source and the target spectra by our approach as introduced in section 2.2 with different number of iterations. The LSD for the 0th iteration stands for the distance between the source and the target spectra according to initial frame alignment by DTW, the LSD for the 1st iteration means the distance between the warped source and the target spectra according to the frame alignment by DTW twice, and so on for the rest of the iterations. The average LSD between the warped source and the target spectra is shown to converge at two iterations; two or more times DTW can make the frame alignment more accurate; the frequency warping approach is more effective for inter-gender (bdl→Slt) conversion than for intra-gender conversion (Clb→Slt).



(a) Using parallel sentence pairs. (b) Using non-parallel sentence pairs.

Fig. 4. The average LSD of (warped) source and target spectra by our proposed frequency warping approach in different number of iterations by (a) using parallel sentence pairs and (b) using non-parallel sentence pairs

Table 2. Average LSD between the target speaker's spectra and the source speaker's (warped) spectra using the conventional method and our method.

LSD (*dB*)	*Without Warping*	*CFW*	*DFW*
Bdl→Slt	9.4	8.30	7. 41
Clb→Slt	7.03	6.95	6.64

However, it is not easy to get parallel utterance pairs from the source and the target speakers in real applications. We suggest two methods to generate parallel feature pairs in section 2.2. We try one of them here and show the average LSD with the number of iterations in Fig. 4 (b). In the 0-th iteration, for each frame of the target feature set, we search a nearest frame from the source feature pool (excluding those extracted from parallel utterances) by minimizing the weighted LSP distance to generate parallel feature pairs. To capture the temporal dynamic information during frame selection, the current frame and its preceding and succeeding frames are used to calculate the distance between

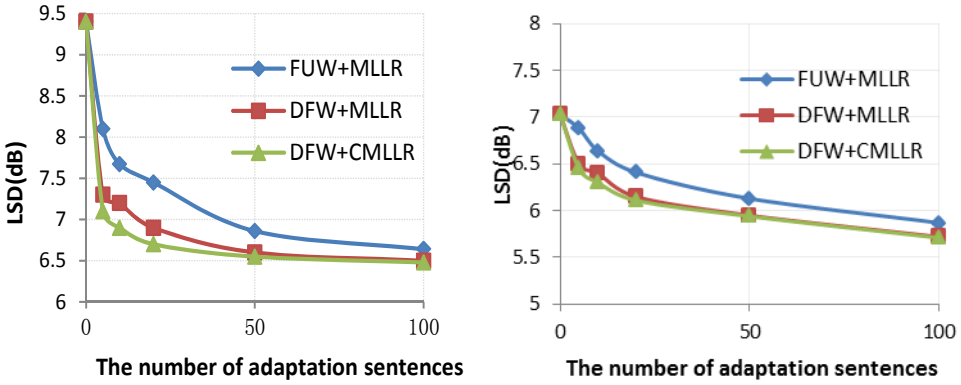
the source and the target frames. The weights for preceding-, current-, and succeeding-frame distances are set to 1, 3, and 1, as suggested in [24]. We show the average LSD between generated parallel feature pairs in Fig. 4 (b). The implementation of our approach to the generated parallel feature pairs is exactly the same as to the parallel feature pairs. In the 1-st iteration, we warp all source features by the warping factor obtained from 0-th iteration, regenerate parallel feature pairs by searching the nearest frame from the warped source feature pool, and recalculate LSD shown in Fig. 4 (b). The following iterations are so on. Fig. 4 (b) shows that using nonparallel data achieves a similar performance to that of using parallel data, but converges slower. We use only voiced frames of both parallel and nonparallel data for GMM training in our frequency warping approach, since formants only exist in voiced frames and LPC analysis is not so stable for unvoiced frames.

To evaluate the effectiveness of the weighted likelihood ratio (or approximated by weighted LSP distance) used in our proposed frequency warping approach (DFW_I), we build a time-varying bilinear frequency warping function (DFW_II) by using maximum likelihood for GMM training and conventional LSD for DTW and iteration convergence. Since the optimization criteria used in DFW_I and DFW_II are different, the objective measure, LSD, is not appropriate to evaluate their performance. A subjective measure, the ABX test, is employed to estimate the distance from the warped source speaker to the target speaker. In this ABX test, the letter "A" represents 50 synthesized source sentences from the warped spectra by DFW_I, "B" denotes the same sentences as A but synthesized from the warped spectra by DFW_II, and "X" stands for 50 original utterances of the target speaker. Ten subjects were asked to select a more similar one between A and B for sentence X. The test results show that 76% of sentences from X sound similar to A and the rest of the sentences sound similar to B. The ABX test results indicate that the weighted likelihood ratio used in our approach can significantly improve the quality of warped spectra on speaker similarity.

3.2.2 Evaluation for further MLLR-based adaptation

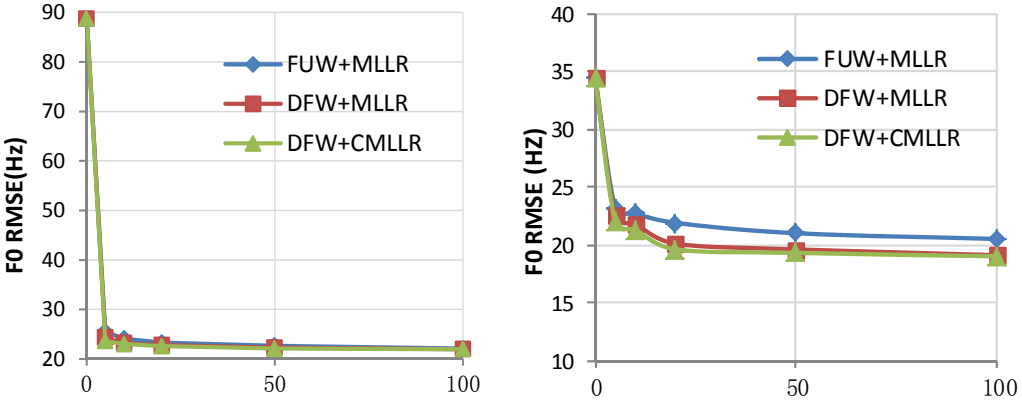
Figs. 5 and 6 show the LSD and *F*0 root mean square error (RMSE) of the testing sentences of the original target speaker and generated by adapted HMMs by MLLR (CMLLR) with 5, 10, 20, 50, and 100 adaptation sentences of target speakers. To make the comparison valid, we also use 5, 10, 20, 50, and 100 aligned sentence pairs from the source speaker's developing data and the target speaker's adaptation data to train a GMM for getting time-varying warping factors in our approach. The resultant numbers of Gaussian components are 2, 3, 5, 8, and 12, respectively. For each experimental configuration, we compare LSD and *F*0 RMSE at different MLLR-based adaptation settings, *e.g.*, the different bandwidth of transform matrix and the number of transform matrices, and report the optimal values in Figs. 5 and 6. The average LSD and *F*0 RMSE of testing sentences between the original target speaker and generated by adapted HMMs both decrease with the increasing number of adaptation sentences. When the number of adaptation sentences is less than 50, the performance of MLLR-based adaptation for source model trained with the warped features is significantly better than the source model trained with the unwrapped features. With more and more adaptation sentences, the benefits brought from spectral frequency warping and *F*0 adjustment become less and less,

but still exist. CMLLR adaptation outperforms MLLR adaptation when the adaptation data size is small, *i.e.*, the number of adaptation sentences is less than 20, while CMLLR and MLLR adaptations perform almost the same with the more adaptation data.



(a) Male (bdl)→female (slt). (b) Female (clb)→female (slt).

Fig. 5. The average LSD of the testing sentences of the original target speaker (slt) and those generated by adapted source speakers' (bdl and clb) HMMs by MLLR and CMLLR with 5, 10, 20, 50 and 100 sentences of the target speaker.



(a) Male (bdl)→female (slt). (b) Female (clb)→female (slt).

Fig. 6. The average *F*0 RMSE of the testing sentences of the original target speaker (slt) and those generated by adapted source speakers' (bdl and clb) HMMs by MLLR and CMLLR with 5, 10, 20, 50 and 100 adaptation sentences of the target speaker.

We synthesize the testing sentences to further evaluate our approach by a subjective listening test. To avoid the over-smoothing problem of parameter trajectories generated by adapted HMMs and the resultant buzzy speech, formant sharpening based on LSP frequency and global variance (GV) [32] as a constraint are used for LSP and *F*0 generation, respectively. The subjective measures are mean opinion scores (MOS) for speaker

similarity and naturalness. The speaker similarity MOS measures how close the two utterances are: one is from the original target speaker and the other is synthesized by adapted HMMs. The naturalness MOS indicates the voice quality of the synthesized speech. The MOS is expressed as a number ranging from 1 to 5, where 1 is the lowest and 5 is the highest.

A total of 900 sentences synthesized by 9 experimental configurations are used in the listening test. The configurations are (1) FUW, synthesized by the source speaker's HMMs trained with original (unwrapped) features; (2) CFW, synthesized by the source speaker's HMMs trained with warped features by the conventional frequency warping method (warping factor is estimated from 50 sentence pairs); (3) DFW, synthesized by the source speaker's HMMs trained with warped features by our method (a time-varying warping function is estimated from 50 sentence pairs); (4) FUW+MLLR, synthesized by HMMs firstly trained with original (unwrapped) features and further adapted by MLLR with 10 and 50 adaptation sentences; and (5) DFW+MLLR(CMLLR), synthesized by HMMs firstly trained with warped features (warping functions are estimated from 10 and 50 sentence pairs) and further adapted by MLLR(CMLLR) with 10 and 50 adaptation sentences.

A total of 36 listeners participated in the listening test. Ten listeners were well-educated native English speakers and the rest of the listeners were graduate students majoring in English. The male and female listeners were evenly distributed, and the listeners' age ranged from 20 to 50 years. Each listener marked the MOS of speaker similarity and naturalness for 50 synthesized sentences, which were randomly selected from the sentences synthesized by each configuration of inter-/intra-gender adaptation.

Fig. 7 shows the MOS results. Table 3 shows the corresponding results of significance tests (at 95% confidence intervals). It shows TRUE (T) where two systems are significantly different and FALSE (F) where they are not. Fig. 7 and Table 3 further demonstrate the effectiveness of our frequency warping approach. In particular, they indicate the following:

1. The synthesized speech by the HMMs trained with warped feature (CFW and DFW) sounds much closer to the target speaker than that of the HMMs trained with unwrapped feature (FUW). Our frequency warp method (DFW) is significantly better than the conventional frequency warping method (CFW) on speaker similarity;

Table 3. The results of significance tests (at 95% confidence intervals).

(a) CFW vs. DFW.

	Similarity	*Naturalness*
Sen Num	50	50
Bdl→Slt	*T*	*F*
Clb→Slt	*T*	*F*

(b) DFW vs. FUW_MLLR.

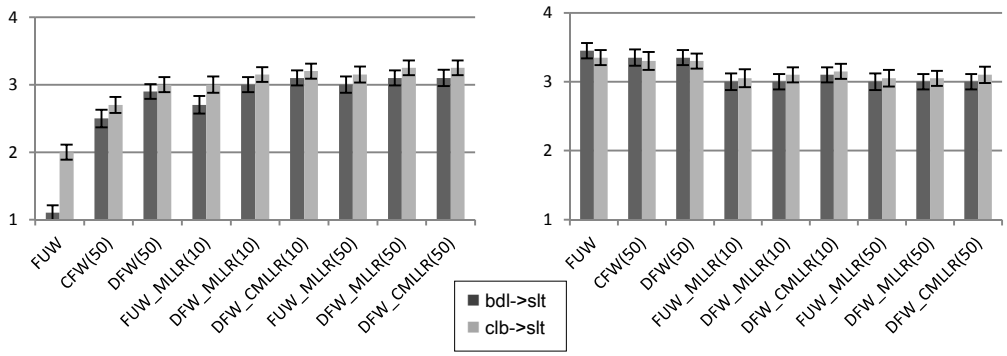
	Similarity	*Naturalness*
Sen Num	50	50
Bdl→Slt	*F*	*T*
Clb→Slt	*F*	*T*

(c) FUW_MLLR vs. DFW_MLLR.

	Similarity	*Naturalness*
Sen Num	10	50
Bdl→Slt	*T*	*F*
Clb→Slt	*F*	*F*

(d) DFW vs. DFW_MLLR.

	Similarity	*Naturalness*
Sen Num	50	50
Bdl→Slt	*T*	*T*
Clb→Slt	*T*	*T*



(a) MOS of speaker similarity. (b) MOS of naturalness.

Fig. 7. The MOS results (at 95% confidence intervals) of (a) speaker similarity and (b) naturalness for the testing sentences synthesized by different configurations.

2. Our frequency warping method, DFW(50), can deliver natural speech with moderate speaker similarity, *i.e.*, the MOS of naturalness is 3.3 and the MOS of similarity is 2.95. By comparing these with MLLR adaptation with 50 sentences for the HMMs trained with unwarpped feature (FUW_MLLR(50)), DFW (50) is significantly better on naturalness and similar on speaker similarity;
3. MLLR adaption with 10 sentences for the HMMs trained with warped feature (DFW_MLLR(10)) significantly outperforms that of unwarpped feature (FUW_MLLR (10)) on speaker similarity for the conversion from male (bdl) to female (slt);
4. Compared with using frequency warping alone (DFW(50)), frequency warping plus MLLR-based adaptation (DFW_MLLR(50)) achieves higher performance on speaker similarity but lower performance on speech naturalness;
5. With the increae of adaptation data, the performance behaviors of MLLR-based adaptation are consistent with what we obtained in the objective measures. The MOS of speaker similarity is slightly improved while the MOS of naturalness is almost unchanged.

Our DFW approach can smoothly warp the spectrum and the pitch of the source speaker's speech unto the target speaker's space in both frequency and time domains. The resultant converted speech is rather natural, *i.e.*, perceptually no significant difference by comparing the synthesized speech with the source speaker's model. It is also observed that DFW can produce synthesized speech significantly natural compared with the synthesized speech produced by FUW_MLLR and DFW_MLLR. However, DFW normalizes the difference of vocal tract length inter speaker. It only reallocates the source speaker's spectral formants on the frequency axis towards those of the target speaker, while the other spectral characteristics, *e.g.*, formant bandwidth, spectral tilt, and spectral intensity, are almost unchanged. DFW also only converts the average $F0$ value and the dynamic range of $F0$ variation, while the conversion of prosodic control (tempo, intonation, loudness, *etc.*) is not considered. Therefore, it is difficult to make the converted speech from the source speaker very similar to the target speaker. If the end users have a high requirement for synthesized speech quality and are not so critical of the similarity to the special person or just want to change voice characteristics of the source

speaker to generate multiple voice fonts, DFW without MLLR would be a good choice.

DFW can be used as a good initialization for further MLLR-based adaptation. When the adaptation data are limited, say, to 10 sentences, MLLR-based speaker adaption has difficulty converting the source speaker's speech to the speech sounds of the target speaker with high quality. DFW_MLLR/CMLLR(10), in which the model used for adaptation is trained by speech parameters that are frequency warped toward the target speaker's space to equalize the vocal tract difference between speakers, can produce synthesized speech more similar to the target speaker's speech than the synthesized speech produced by FUW_MLLR/CMLLR(10), where the model used for adaption is trained by unwarpped speech parameters. This implies a critical impact of initialization onto the performance of the final adapted model, especially for a case where the target speaker's acoustic characteristics significantly differ from the source speaker's characteristics, *e.g.*, the conversion from male (bdl) to female (slt). DFW_MLLR/CMLLR can achieve an additive improvement over MLLR-based adaptation when the amount of adaptation data is small.

MLLR-based speaker adaptation approaches try to convert all voice characteristics' variability, *e.g.*, vocal tract length, mouth dimension, nasal cavity, accent, speaking rate, *etc.*, in a unified statistical way. However, the adaptation data are always not enough. Generally well-trained HMMs have a good phonetic and prosodic coverage. The HMM Gaussians, which are close in acoustic space, need to be grouped into a class and share a transform matrix in MLLR-based adaptation if the amount of adaptation data is not large enough. It may result in quality degradation of the synthesized speech, especially for the adaptation between two speakers who are largely different. Thus, it is observed in our experimental results that the naturalness of synthesized speech by the adapted model with MLLR-based adaption is worse than those of only taking DFW without MLLR.

3.3 Discussions

Frequency warping can be seen as a feature transform in the speaker adaptation of HMM-based speech synthesis. Our dynamic frequency warping method based on a bilinear function can transform features in a time-frequency smoothing way. The bilinear warping function performs a smooth transform over frequency, and frame-dependent warping factor estimated by the GMM probability of acoustic classes produces a smooth transform over time. Our frequency warping plus MLLR adaptation approach can be regarded as speaker transformation twice. One is a feature transform based on the criterion of minimizing weighted spectral distance, D_{WLR}, which weights the formant distance, between the source and the target speakers. The other is an HMM transform in the sense of maximum likelihood of the target speaker's data for the adapted model.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we use frequency warping for speaker adaptation in an HMM-based speech synthesis system. We first employ a GMM-based approach to generate a time-frequency smooth warping function then we warp the spectra of the source speaker towards the target speaker and retrain HMMs with LSPs extracted from the warped spectra.

Finally, we apply MLLR adaptation to retrained HMMs. The experimental results show that our approach outperforms the conventional MLLR-based speaker adaptation of speech synthesis system subjectively and objectively. Our approach is time-consuming, since it requires the retraining of the whole HMMs for every target speaker. In the next step, we will try to integrate feature transform based on the criterion of minimizing weighted spectral distance into model training [33, 34], like the implementation of CMLLR in the speaker adaptation of an HMM-based speech synthesis system.

REFERENCES

1. Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, Vol. 6, 2002, pp. 131-142.
2. T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, 2007, pp. 2222-2235.
3. A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 373-376.
4. K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 1315-1318.
5. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, 2009, pp. 66-83.
6. M. Blcmberg and K. Elenius, "Nonlinear frequency warping for speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1986, pp. 2631-2634.
7. P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Technical Reports CMU CS-97-148, School of Computer Science, Camergie Mellon University.
8. Z.-W. Shuang, R. Bakis, and Y. Qin, "Voice conversion based on mapping formants," in *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, 2006, pp. 219-223.
9. E. P. Neuburg, "Frequency warping by dynamic programming," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 573-575.
10. T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 841-844.
11. D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, 2007, pp. 1965-1968.

12. S. P. Rath and S. Umesh, "Acoustic class specific VTLN-warping using regression class trees," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, 2009, pp. 556-559.
13. L. Saheer, P. N. Garner, J. Dines, and H. Liang, "VTLN adaptation for statistical speech synthesis," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2010, pp. 4838-4841.
14. X. Zhuang, Y. Qian, F. K. Soong, Y.-J. Wu, and B. Zhang, "Formant-based frequency warping for improving speaker adaptation in HMM TTS," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 817-820.
15. W.-X. Gao and Q.-Y. Cao, "Frequency warping for speaker adaptation of text-to-speech synthesis," in *Proceedings of International Conference on Wireless*, *Mobile and Multimedia Networks*, 2010, pp. 307-310.
16. D. Paczolay, A. Kocsor, and L. Toth. "Real-time vocal tract length normalization in a phonolgical awareness teaching system," *Text Speech and Dialogue*, LNCS, Vol. 2807, Springer, Czech Republic, 2003, pp. 309-314.
17. C. Huang, Y-C Huang, F. K. Soong and J.-L. Zhou, "Weighted likelihood ratio (WLR) hidden Markov model for noisy speech recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2006, pp. 37-40.
18. R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 1991, pp. 641-644.
19. Y. Qian, F. K. Soong, Y. N. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, 2006, pp. 223-232.
20. M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, Vol. 12, 1998, pp. 75-98.
21. V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, 1995, pp. 357-366.
22. J. Kominek and A. Black, "The CMU ARCTIC databases for speech synthesis," Technical Report CMU-LTI-03-177, Language Technologies Institute, CMU, 2003. http://www.festvox.org/cmu_arctic.
23. H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 27, 1999, pp. 187-207.
24. T. Hirai, J. Yamagishi, and S. Tenpaku, "Utilization of an HMM-based feature generation module in 5ms segment concatenative speech synthesis," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2007, pp. 81-84.
25. R. Greisbach, O. Esser, and C. Weinstock, "Speaker identification by formant contours," in A .Braun, J.-P. Köster, eds., *Studies in Forensic Phonetics: BEIPHOL*, Vol. 64, Wissenschaftlicher Verlag, Trier, 1995, pp. 49-55.
26. D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan,

"Text-independent voice conversion based on unit selection," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2006, pp. 81-84.
27. M. Sugiyama, "LPC spectral matching measures for speech recognition," Ph.D. dissertation, School of Engineering, Tohoku University, 1984.
28. HTS, http://hts.sp.nitech.ac.jp/.
29. L. Saheer, J. Dines, and P. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," *IEEE Transactions on Audio*, *Speech and Language Processing*, Vol. 20, 2012, pp. 2134-2148.
30. L. Saheer, J. Yamagishi, P. Garner, and J. Dines, "Combing vocal tract length normalization with hierarchial linear transformations," in *Proceedings of International conference on Speech and Signal Processing*, 2012, pp. 4493-4496.
31. F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 1984, pp. 1.10.1-1.10.4.
32. T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proceedings of the 6th Annual Conference of the International Speech Communication Association*, 2005, pp. 2801-2804.
33. Y.-J. Wu and K. Tokuda, "Minimum generation error training by using original spectrum as reference for log spectral distortion measure," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2009, pp. 4013-4016.
34. Y.-J. Wu, L. Qin, and K. Tokuda, "An improved minimum generation error based model adaptation for HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of International Speech Communication Association*, 2009, pp. 1787-1790.

Weixun Gao (高伟勋) received a BS in the Department of Computer Science from Shanghai Normal University, China in 1995 and a MS in the College of Software from Fudan University, China in 2007. Currently he is pursuing the Ph.D. in School of Information Science and Technology, Donghua University, Shanghai, China.

Qiying Cao (曹奇英) is a Professor in the College of Computer Science and Technology, Donghua University, Shanghai, China. His research interests ubiquitous computing, intelligent information processing, network and information security, embedded technology and network home appliances and computer-aided technologies.