INTERSPEECH 2009 BRIGHTON

Efficient Modeling of Temporal Structure of Speech For Applications in Voice Transformation

Binh Phu Nguyen and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology

{npbinh, akagi}@jaist.ac.jp

Abstract

Aims of voice transformation are to change styles of given utterances. Most voice transformation methods process speech signals in a time-frequency domain. In the time domain, when processing spectral information, conventional methods do not consider relations between neighboring frames. If unexpected modifications happen, there are discontinuities between frames, which lead to the degradation of the transformed speech quality. This paper proposes a new modeling of temporal structure of speech to ensure the smoothness of the transformed speech for improving the quality of transformed speech in the voice transformation. In our work, we propose an improvement of the temporal decomposition (TD) technique, which decomposes a speech signal into event targets and event functions, to model the temporal structure of speech. The TD is used to control the spectral dynamics and to ensure the smoothness of transformed speech. We investigate the TD in two applications, concatenative speech synthesis and spectral voice conversion. Experimental results confirm the effectiveness of TD in terms of improving the quality of the transformed speech.

Index Terms: spectral modification, voice transformation, temporal decomposition

1. Introduction

Voice transformation is a process of changing certain perceptual properties of speech while leaving other properties unchanged. Voice transformation has many applications in our lives. For example, we employ voice transformation techniques to create various wave sounds from a pre-recorded database in a Text-to-Speech system. In foreign language learning, it will be much easier to listen when slowing down the speed of sounds. To enhance the hearing abilities of deaf people, we can adjust the frequency of sounds so that it is located in their hearing portion.

The goals of voice transformation systems are to generate wave sounds from a pre-recorded speech database, or to alter styles of speech utterances without losing the utterance content, etc. The styles which can be changed include the speaker's gender, the speaker's identity, or the speaker's emotion and so on.

Spectral modification lies at the heart of the voice transformation. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing. Most methods of spectral modification process speech signals in the time-frequency domain. The basic idea of spectral processing is to convert a time-domain digital signal into its representation in a time-frequency domain. In the time axis, most of them process speech signals frame-by-frame. They do not ensure the smoothness of synthesized speech after modification, which leads to the degradation of modified speech quality. One study [1] points out that spectral dynamics is more important than spectral distortion in human perception. Therefore, it is necessary to have a new method for ensuring the smoothness of the transformed speech.

In the area of voice transformation, many methods have been proposed to solve discontinuities of speech signals after modification. For example, in the concatenative synthesis area, Plumpe et al. [2] introduce a HMM-based smoothing technique. A large training database is required to estimate the HMM parameters, and this point is a limitation. Wouters and Macon [3] propose a method which controls spectral dynamics. In this approach, synthesis is performed by combining information from two tiers of speech units, denoted concatenation units and fusion units. The concatenation units specify initial estimates of the spectral trajectories for an utterance, while the fusion units characterize the spectral dynamics at the join points between concatenation units. These two unit tiers are fused during synthesis to obtain natural spectral transitions throughout the synthesized speech. Preparation of a fusion unit for each concatenation point is required. Kain et al. [4] also propose a new method of controlling spectral dynamics which has same idea with the work of Wouters and Macon [3]. They smooth the trajectory of formant frequencies. In [4], it is not necessary to prepare the fusion units. Apart from that, this method considers the smoothness of energy. Since this method use formant frequencies as parameters to interpolate between two segments, some steps in this method need to be manually performed. Therefore, a new method for concatenative speech synthesis which is automatically performed is needed. In the spectral voice conversion area, to maintain a continuous transformation in consecutive frames, Chen et al. [5] smooth the converted features along the time axis by employing a median filter and a low pass filter. Applying these filters can lead to a loss of temporal resolution, and it is a relatively crude implementation. Duxans et al. [6] include dynamic information in their GMM-based voice conversion system to take into the relations between frames. According to Duxans et al. [6], this method does not improve the performance of a GMM-based voice conversion system. In [7], Toda at al. include dynamic features and the global variance to solve the discontinuities of spectral conversion in the time domain. This method improves the quality of the converted speech.

One of the effective ways to solve the discontinuity problem of the voice transformation applications is to develop a method for modeling the temporal evolution of speech. In the literature, a hidden Markov model (HMM) is well-known to model the temporal trajectories of speech parameters. However, two major drawbacks of the HMM are discussed: the assumption of conditional independence of successive states is grossly unrealistic, and the HMM has to rely on a large amount of training data to (partially) capture the temporal evolution. Another technique, the temporal decomposition (TD) technique [8], is also used to model the temporal evolution of speech, and it can overcome the two drawbacks of the HMM.

In the remaining paper, we first present our improvements of the temporal decomposition (TD) technique [8, 9] to model the temporal structure of speech. Based on our modeling of temporal decomposition of speech, we then introduce our new methods in two applications of the voice transformation, concatenative speech synthesis and spectral voice conversion, to improve the quality of the transformed speech.

2. Temporal decomposition

2.1. Introduction

Modeling the temporal trajectories of speech parameters gives the advantages to speech processing applications. This section presents the TD technique [8, 9] as an efficient model of temporal structure of speech.

Atal proposes a method based on the temporal decomposition of speech into a sequence of overlapping event functions and corresponding event targets [8], as given in Eq. (1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \le n \le N \tag{1}$$

where $\mathbf{a}_k$ is the spectral parameter vector corresponding to the k^{th} event target. The temporal evolution of this target is described by the k^{th} event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of the n^{th} spectral parameter vector $\mathbf{y}(n)$ produced by the TD model. *N* and *K* are the number of frames in the speech segment and the number of event functions, respectively ($N \gg K$). The TD does not need to assume the independence of event targets, and the TD bases on only the speech segment when estimating the event targets and event functions (spectral evolution). These are advantages when comparing with the HMM.

The original method of TD is known to have two major drawbacks, high computational costs and high parameter sensitivity to the number and locations of events. A number of modifications have been explored to overcome these drawbacks. In this study, we employ the MRTD algorithm [9]. The reasons for using the MRTD algorithm in this work are two-fold: (i) the MRTD algorithm enforces a new property on event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [9]; (ii) event targets can convey the speaker's identity [10]. In the MRTD algorithm, LSF parameters are chosen for the input of TD, because LSFs have good linear interpolation attributes. In addition, the temporal pattern of the excitation parameters can also be described by using the same event functions evaluated for the spectral parameters $\phi_k(n)$ in Eq. (1) and excitation targets $\mathbf{b}_k$.

Since the same event functions evaluated for the spectral parameters are also used to model the temporal pattern of the excitation parameters, we only need to modify these targets, $\mathbf{a}_k$, $\mathbf{b}_k$, and the corresponding event functions $\phi_k(n)$ for modifying the speech signals, instead of modifying the speech signals frame by frame. The smoothness of modified speech will be ensured by the shape of the event functions $\phi_k(n)$. This leads to easy modification of the speech signals in time-frequency domain, as well as ensuring the smoothness of the speech signals between frames, and thereby enhances the quality of modified speech.

2.2. Modeling of the event function using polynomial fitting

MRTD algorithm uses a spectral stability criterion to determine the initial event locations [9]. This algorithm is useful for applications in speech coding [9] and speaker identification [10]. However, this algorithm does not ensure one-to-one correspondence between events and phonemic units, which makes it difficult for applications in voice transformation (e.g. alignment between two utterances), speech perception (e.g. sharing the event functions, event targets).

In [11], we present a new method for the determination of event locations based on phonemes, and a new method for modeling the event function by using the nonlinear least square method as follows.

$$R = -\left(\frac{t}{d}\right)^{S} + e \tag{2}$$

where t is time variance, which indicates duration between a spectral parameter vector and the first event of the modeling event function, d is the duration of two consecutive events, e is the maximum value of the event function ϕ_k, and e is equal to 1. The polynomial fitting was done in $0 \le \phi_k \le 1$. The value of S indicates slope of event function. Shape of the event function can be changed according to the values of d and S. As a result, it is possible and flexible to control the event function by changing the value of d and S. More details of our methods can be found in [11].

3. Applications to voice transformation

In this paper, to show the effectiveness of our modeling for ensuring the smoothness of the transformed speech, we investigate the TD in two applications in voice transformation: concatenative speech synthesis and spectral voice conversion. Moreover, in voice transformation applications, we modify not only vocal tract information but also excitation information. Since the excitation and vocal tract information are not independent, modifying them separately often degrades the quality of converted speech. Therefore, a high quality analysis-synthesis framework, STRAIGHT [12] is utilized in this paper.

3.1. Proposed spectral smoothing for concatenative speech synthesis

Since controlling spectral dynamics can improve the quality of concatenation speech, we propose a new method for concatenative speech synthesis based on temporal decomposition [8, 9]. Our algorithm is described as follows.

First, LSF parameters are extracted from STRAIGHT spectral envelope [12]. MRTD is employed in the next step to decompose the LSF parameters of each speech segment into event targets and event functions. The same event function evaluated for LSF parameters are used to decompose the fundamental frequency and gain to get fundamental frequency targets and gain targets. In the ideal case, the last target of the first speech segment and the first target of second speech segment are identical. However, in concatenative speech synthesis, two event targets are often different. We need to modify these targets to smooth the transition between two speech segments. Since each event target is a valid LSF parameter, we should modify event targets so that they become a valid LSF parameter. In our algorithm, the modified event target is calculated by applying following equation.

$$LSF_i^{modified} = \beta LSF_i^{last\ \ ET} + (1-\beta) LSF_i^{first\ \ ET} \tag{3}$$

where $i = 1 \ldots P$, P is the order of LSF. The $LSF^{last\ \ ET}$ and $LSF^{first\ \ ET}$ are the LSF parameters of the last event target of the first speech segment and the first event target of the second speech segment, respectively. β is the weight factor, and satisfies $0 \le \beta \le 1$. We can adjust the value of β to control the degree of modification of each concatenation part in accordance with their importance. In this paper, we choose $\beta = \frac{1}{2}$. The optimal value of β for each concatenation point will be investigated in our future work. After combining the last event target of the first speech segment and the first event target of the second speech segment, we also modify the fundamental frequency targets and gain targets to smooth all of the most important parameters in the concatenation point. The modified event targets, modified fundamental frequency targets and modified gain targets are then re-synthesized as modified LSFs, modified fundamental frequency information and modified gain information by TD synthesis, respectively. In the next step, the modified LSF parameters and modified gain information are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the synthesized speech. Note that when we modify these targets, the spectral and source information of adjacent frames around on the concatenation point are also modified, and the smoothness is ensured by the shape of the event functions.

3.2. Proposed spectral voice conversion using temporal decomposition and Gaussian mixture model

Until now, GMM-based spectral voice conversion methods are regarded as some of the most successful techniques. However, the quality of the converted speech is still far from natural. There are three main problems: insufficient smoothness of the converted spectra between frames, the insufficient precision of GMM parameters and over-smooth effect happens in each converted frame.

The first problem is discussed in the Introduction. The second problem is described as follows. In the training phase of the

GMM-based methods, both unstable frames, which often come from transition parts between phonemes, and stable frames are used to model the distribution of acoustic features. This leads to addition of noise to the GMM parameters. To overcome this drawback, some solutions have been proposed. Kumar and Verma [13] explicitly partition acoustic space of a speaker into phones by using the phonetic alignments. After that, GMM parameters are used for finer modeling of each phone. This approach can prevent the interference of frames between phones. However, it still uses unstable frames in each phone. Liu et al. [14] segment frames according to each phoneme, and eliminate unstable frames in each phoneme by proposing a method for identifying stable frames based on limitation of maximal variation range for the first three formant frequencies. After getting the stable frames, Liu et al. also use GMM parameters to model the distribution of acoustic features. Nguyen and Akagi [15] use event targets as spectral vectors to estimate GMM parameters, instead of using spectral parameters of aligned frames. However, all methods in [13, 14, 15] do not take into account the relations between frames when estimating the GMM parameters. The GMM parameters therefore are more precisely estimated when considering the relations between frames. Defining solutions for the third problem, over-smooth effect happens in each converted frame, is beyond the scope of this section

This section addresses two of the three main issues mentioned above, the insufficient smoothness of the converted spectra between frames and the insufficient precision of GMM parameters. Our proposed method focuses on spectral voice conversion, and is based on the GMM method [16, 17]. The processing flow of our spectral voice conversion system is described as follows.

In the training phase, LSF parameters (extracted from STRAIGHT spectral envelope [12]) are decomposed into event targets and event functions by using the MRTD [9]. Each phoneme is represented by five event targets. In these five event targets, two edge event targets coincide with edge event targets of adjoining phonemes and the beginning of a phoneme is more important than the ending of a phoneme. We formulate a vector of phoneme-based features of event targets $\mathbf{EV} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \mathbf{a}_3^T, \mathbf{a}_4^T]$, where $\mathbf{a}_k (1 \leq k \leq 4)$ is the k^{th} event target in each speech segment (a phoneme). $\mathbf{EV} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \mathbf{a}_3^T, \mathbf{a}_4^T]$ represents sequences of four consecutive event targets in a phoneme, and can therefore explicitly characterize the relationship between these vectors. Moreover, each event target $\mathbf{a}_k$ in the MRTD algorithm [9] is a valid LSF coefficient. An important property of LSFs $\{LSF_i\}$ is that they are ordered $(0, \pi)$, as follows.

$$0 < LSF_1 < LSF_2 < \ldots < LSF_P < \pi \tag{4}$$

where P is the order of LSF. To prevent a bad initialization in estimation of GMM parameters, we normalize the vectors of phoneme-based features of event targets extracted from each phoneme in utterances of source and target speakers, **x** and **y**, as follows.

$$\mathbf{x} = [\mathbf{a}_{s1}^T, \mathbf{a}_{s2}^T + \pi, \mathbf{a}_{s3}^T + 2\pi, \mathbf{a}_{s4}^T + 3\pi]^T \tag{5}$$

$$\mathbf{y} = [\mathbf{a}_{t1}^T, \mathbf{a}_{t2}^T + \pi, \mathbf{a}_{t3}^T + 2\pi, \mathbf{a}_{t4}^T + 3\pi]^T \tag{6}$$

where $\mathbf{a}_{sk}$, $\mathbf{a}_{tk}$ are the k^{th} event targets in each phoneme of the source and target speakers, respectively. As a result, the vectors **x** and **y** are ordered $(0, 4\pi)$. We align the phoneme-based features, **x** and **y**, and formulate a set of joint vectors of event targets between source and target speakers $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q]$ where $\mathbf{z}_i = [\mathbf{x}_i^T, \mathbf{y}_i^T]^T$, and $\mathbf{x}_i, \mathbf{y}_i$ are event target sets of i^{th} phoneme of source speaker and the corresponding event target of the target speaker, respectively. Our transformation procedure is the same with that in the conventional GMM-based method [17], except that the vectors for the transformation procedure are the sets on normalized phoneme-based features, **x** and **y**, in Eqs. (5) and (6). When getting the converted phoneme-based features, we convert these vectors back to event targets. The converted event targets are re-synthesized as converted LSF by MRTD synthesis. Then, the converted LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the converted speech. Note that our method does not deal with prosodic, energy conversion. To implement a complete voice conversion system, our work should be integrated with some methods for prosodic, energy conversion, such as in [18].

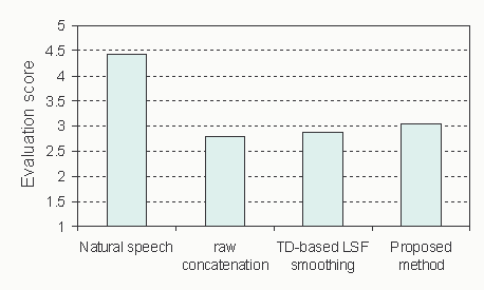

Figure 1: *Results of subjective tests of concatenative speech synthesis.*

4. Experiments and results

This section evaluates the effectiveness of our proposed methods in voice transformation. We evaluate our spectral smoothing method in Subsection 4.1 and spectral voice conversion method in Subsection 4.2.

4.1. Concatenative speech synthesis

Stimuli consisted of the five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) in a consonant-vowel-consonant (CVC) context. We selected a dataset consisting of five words containing the five Japanese vowels from the ATR Japanese speech database [19]. We exchanged the vowels in these words, and smoothed the borders by using different methods. Some synthesized words were meaningless. The main analysis conditions for these experiments are as follows. Sampling frequency is 16 kHz, the order of LSF is 32.

To evaluate the performance of our proposed method, we performed subjective experiments regarding speech quality. We compared our proposed method with two other methods. In the first method, we only concatenated speech segments together (the raw concatenation method); in the second method, we only smoothed spectral parameters by using TD, but we did not smooth F0 and energy (TD-based LSF smoothing method). We presented the synthesized sounds to eight Japanese graduate students with normal hearing ability, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Results of the subjective tests are shown in Fig. 1. These results indicate that the quality of words modified by using our proposed method is the best in all three methods. Fig. 2 shows the parts of the LSF contours before and after modification at the concatenation points by replacing the vowel *"u"* in the word *"takumi"* by the vowel *"e"* in the word *"jiten"*.

4.2. Spectral voice conversion

The corpus used for the experiments is a dataset consisting of 460 sentences spoken once each by two speakers (one male & one female) in the MOCHA-TIMIT English speech database [20]. In our experiments, two different voice conversion tasks were investigated: male-to-female (M2F) and female-to-male (F2M) conversion. For each kind of conversion, we used 300 pair utterances for training and 30 other pair utterances for evaluation.

To evaluate the performance of our proposed method, we performed subjective experiments regarding speech quality and speaker individuality. Six graduate students known to have normal hearing ability were recruited for the listening experiments. We compared our proposed method (the phoneme-based TD+GMM method) with two other methods. The first

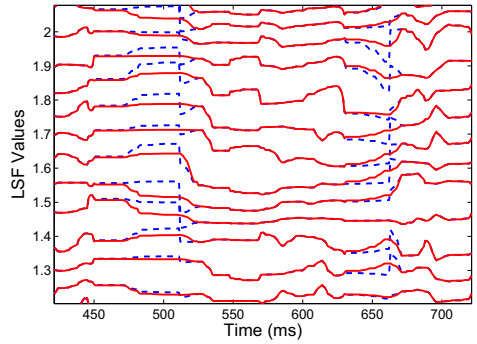

Figure 2: *Parts of the LSF contours before and after modification at the concatenation points by replacing the vowel* "u" *in the word* "takumi" *by the vowel* "e" *in the word* "jiten". *The dot line indicates the LSF contours of the two speech units before modification. The solid line indicates the LSF contours of the two speech units after modification by using our proposed method.*

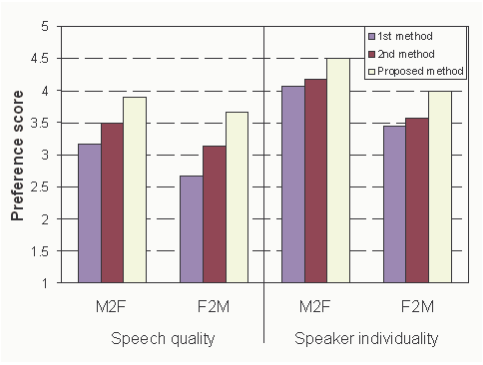

Figure 3: *Results of subjective tests of spectral voice conversion regarding speech quality and speaker individuality.* 1^{st} *method: conventional GMM method,* 2^{nd} *method: TD+GMM without considering phoneme relations, our proposed method: phoneme-based TD+GMM.*

method used for comparison is the conventional method (the GMM method) [17]. The second method used for comparison also employed event targets for training, and the transformation procedure was performed for each event target (the TD+GMM method). The difference between the second method and our proposed method is that the second method does not take into account the relations between event targets in training and transformation procedures. Since we only focus on spectral voice conversion, we automatically copy the prosody information and energy from the utterances of the target speaker to converted utterances. In addition, because the problem of the over-smooth effect in each converted frame is outside the scope of this section, without loss of generality, all three methods utilize the same transformation mapping function of the conventional method [17]. The main analysis conditions for these experiments are as follows. Sampling frequency is 16 kHz, the order of LSF is 32, and the number of Gaussian components is 128.

We randomly presented each of ten converted utterances from both kinds of conversion (male-to-female and female-to-male) to listeners, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). In the test of speaker individuality, an ABX test was conducted. A represents the source speaker, B represents the target speaker, and X represents the converted speech, which supplied from each of two test methods. The listeners were asked to select if X was closer to A or B, and adjusted the score from 1 (very similar to A) to 5 (very similar to B) according to his/her perception of speech individuality when comparing. Results of the subjective tests are shown in Fig. 3. These results indicate that the quality of utterances converted using our proposed method better than that using the conventional method (GMM method) [17] and the second method (TD+GMM method).

5. Conclusions

In this paper, we have presented the effectiveness of TD in voice transformation applications, concatenative speech synthesis and spectral voice conversion. The event targets are considered to be "ideal" spectral parameters, can convey speaker's identity. The event functions are regarded as modelings of the spectral evolutions. Using the TD in voice transformation, we only need to modify the event targets and event functions, which leads to efficient and flexible modifications of speech. Experimental results show the effectiveness of our methods when applied to voice transformation in terms of improving the quality of modified speech.

Modeling the temporal structure of speech gives benefits to most of areas in the speech technology, such as speech coding, speech recognition, speaker verification and identification, speech modification. In our work, spectral evolution can be controlled by changing values of duration between two consecutive events, d, and slope of each event function, S. In addition, in our modeling, S indicates a slope of an event function, and S can be seen as dynamic information between multiple frames. It is of interest to investigate the incorporation between event targets $\mathbf{a}_k$ and the slopes of event functions S in speech processing applications, such as speech and speaker recognition.

6. Acknowledgments

This study was supported by SCOPE (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

7. References

[1] Knagenhjelm, H.P. and Kleijn, W.B., "Spectral dynamics is more important than spectral distortion," Proc. ICASSP, 732–735, 1995.
[2] Plumpe, M., Acero, A., Hon, H.W., and Huang, X., "HMM-based smoothing for concatenative speech synthesis," Proc. ICSLP, 1998.
[3] Wouters, J. and Macon, M., "Control of spectral dynamics in concatenative speech synthesis," IEEE Trans. Speech and Audio Proc., 30–38, 2001.
[4] Kain, A., Miao, Q., and van Santen, J., "Spectral control in concatenative speech synthesis," Proc. ISCA Workshop on Speech Synthesis, 2007.
[5] Chen, Y., Chu, M., Chang, E., Liu, J., and Liu, R., "Voice conversion with smoothed GMM and MAP adaptation," Proc. Eurospeech, 2413–2416, 2003.
[6] Duxans, H., Bonafonte, A., Kain, A., and van Santen, J., "Including dynamic and phonetic information in voice conversion systems," Proc. Interspeech, 1193–1196, 2004.
[7] Toda, T., Black, A.W., and Tokuda, K., "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech and Language Proc., 15(8): 2222–2235, 2007.
[8] Atal, B.S., "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP, 81–84, 1983.
[9] Nguyen, P.C., Ochi, T., and Akagi, M., "Modified restricted temporal decomposition and its application to low bit rate speech coding," IEICE Transactions on Information and Systems, E86-D: 397–405, 2003.
[10] Nguyen, P.C., Akagi, M., and Ho, T.B., "Temporal decomposition: A promising approach to VQ-based speaker identification," Proc. ICASSP, 184–187, 2003.
[11] Nguyen, B.P., Shibata, T., and Akagi, M., "High-quality analysis/synthesis method based on temporal decomposition for speech modification," Proc. Interspeech, 662–665, 2008.
[12] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27: 187–207, 1999.
[13] Kumar, A. and Verma, A., "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," Proc. ICASSP, 720–723, 2003.
[14] Liu, K., Zhang, J., and Yan, Y., "High quality voice conversion through combining modified GMM and formant mapping for Mandarin," Proc. ICDT, 10, 2007.
[15] Nguyen, B.P. and Akagi, M., "Control of spectral dynamics using temporal decomposition in voice conversion and concatenative speech synthesis," Proc. NCSP, 279–282, 2008.
[16] Stylianou, Y., Cappe, O., and Moulines, E., "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Proc., 6(2): 131–142, 1998.
[17] Kain, A. and Macon, M.W., "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, 285–288, 1998.
[18] Erro, D. and Moreno, A., "Weighted frequency warping for voice conversion," Proc. Interspeech, 1965–1968, 2007.
[19] Abe, M., Sagisaka, Y., Umeda, T., and Kuwabara, H., "Speech database user's manual," ATR Technical Report, TR-I-0166, 1990.
[20] Wrench, A., "The MOCHA-TIMIT articulatory database," Queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.