IMPLEMENTATION OF F0 TRANSFORMATION FOR STATISTICAL SINGING VOICE CONVERSION BASED ON DIRECT WAVEFORM MODIFICATION

Kazuhiro Kobayashi[1], *Tomoki Toda*[2], *Satoshi Nakamura*[1]

Graduate School of Information Science, Nara Institute of Science and Technology, Japan[1]
Information Technology Center, Nagoya University, Japan[2]

ABSTRACT

This paper presents a technique for transforming F_0 in a framework of statistical singing voice conversion with direct waveform modification based on spectrum differential (DIFFSVC). The DIFFSVC method converts voice timbre of singing voices of a source singer into that of a target singer without using vocoder-based waveform generation. Although this method achieves high sound quality of the converted singing voices, its use is limited to only intra-gender conversion without the need of F_0 transformation. To make it possible to also use the DIFFSVC method for cross-gender conversion, we propose a method to transform F_0 of an input singing voice for the DIFFSVC. The proposed method is also based on direct waveform modification using overlap-add process and filtering process. Results of subjective evaluations demonstrate that the proposed DIFFSVC method with F_0 transformation significantly improves sound quality of the converted singing voices while preserving the conversion accuracy of singer identity in the cross-gender conversion compared to the conventional SVC with vocoder.

Index Terms— statistical singing voice conversion, cross-gender conversion, direct waveform modification, spectral differential, F_0 transformation

1. INTRODUCTION

Singers can expressively control F_0, rhythm, and voice timbre of singing voices while also conveying linguistic information of the lyrics. However, they usually have a difficulty in changing their voice timbre or their F_0 range (e.g., changing their own voice timbre and F_0 range into those of another specific singer) owing to physical constraints in speech production. If singers could freely control their voice timbre and F_0 range beyond their physical constraints, it would open up entirely new ways for singers to express more varieties of expression.

Singing voice conversion (SVC) is a technique for directly converting a source singer's singing voice into another target singer's singing voice [1, 2]. One of the typical methods is based on statistical voice conversion (VC) techniques [3, 4]. A conversion model is trained in advance using a parallel data set of song pairs sung by the source and target singers. The trained conversion model makes it possible to convert the acoustic features of the source singer's singing voice into those of the target singer's singing voice in any song while keeping the linguistic information of the lyrics unchanged. Recently eigenvoice conversion (EVC) techniques [5, 6] have been successfully applied to SVC [7] to develop a more flexible SVC framework capable of achieving conversion between arbitrary source and target singers, even if their parallel data set is not available. Moreover, in order to manually control voice timbre by manipulating intuitively understandable parameters, such as a perceived age, a voice timbre control technique based on regression approaches to VC [8] has also been successfully applied to SVC [9]. Furthermore, a real-time VC technique [10] can also be applied to these SVC frameworks. Therefore, these techniques enable singers to sing songs in real time with their desired voice timbre and F_0 range, not limited by physical constraints.

Although these SVC frameworks have great potential to bring a new singing expression to singers, there remain several problems to be solved. One of the biggest problems is that sound quality of the converted singing voice is significantly degraded compared to that of the natural singing voice. The SVC frameworks usually use vocoder [11] to produce the converted singing voice from the converted acoustic features. Consequently, sound quality of the converted singing voice suffers from various errors, such as F_0 extraction errors, modeling errors in spectral parameterization, and over-smoothing effects on the converted acoustic features. These issues are indeed hard to be addressed even by using high-quality vocoder systems [12, 13, 14, 15].

As a new SVC framework not suffering the errors caused by using vocoder-based waveform generation, we have proposed a SVC method based on direct waveform modification with spectral differential (DIFFSVC) [16]. This method directly filters a waveform of an input natural singing voice with time-variant spectral differential between the source and target singers, which is statistically estimated with the statistical conversion model [16]. Namely, the vocoder-based waveform generation is no longer needed in this method by effectively using excitation signals of the input natural singing voices. Moreover, the over-smoothing effects are well alleviated by also considering global variance (GV) [4] in the DIFFSVC [17]. We have reported that the DIFFSVC yields significant improvements in quality of the converted singing voices while preserving conversion accuracy in singer identity compared to the conventional SVC. On the other hand, the use of the DIFFSVC method is restricted to only intra-gender SVC without the necessity of F_0 transformation because the excitation signals of the input natural singing voices are directly used. It is expected that the DIFFSVC method will be widely available for not only intra-gender SVC but also cross-gender SVC by implementing F_0 transformation without using the vocoder-based waveform generation.

In this paper, we propose an F_0 transformation method for the DIFFSVC method to make it possible to also use the DIFFSVC for the cross-gender SVC. In the cross-gender SVC, the F_0 transformation is often needed because a male singer and a female singer usually sing a song on different keys (e.g., one octave difference). To transform F_0 while keeping an advantage of the DIFFSVC, i.e., not using the vocoder-based waveform generation, the proposed F_0 transformation method uses direct waveform modification based on overlap-add process and filtering process. Moreover, to avoid a mismatch of spectral envelop before and after F_0 transformation, which causes significant performance degradation of the DIFFSVC, the modification process is applied to residual signals extracted from the input natural singing voices. We conduct subjective evaluations, demonstrating that the proposed DIFFSVC with F_0 transformation significantly improves sound quality of the converted singing voices while preserving the conversion accuracy of singer identity in the cross-gender conversion compared to the conventional SVC with vocoder.

2. STATISTICAL SINGING VOICE CONVERSION WITH DIRECT WAVEFORM MODIFICATION (DIFFSVC)

DIFFSVC consists of a training process and a conversion process. In the training process, a joint probability density function of spectral features of the source singer and the target singer is modeled with a GMM in a traditional manner [18]. Then, a differential GMM for modeling the joint probability density function of the source spectral feature and the spectral differential feature between the source and target singers is analytically derived from the trained traditional GMM.

As the spectral features of the source and target singers, we employ $2D$-dimensional joint static and dynamic feature vectors $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ of the source and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ of the target consisting of D-dimensional static feature vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and their dynamic feature vectors $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ at frame t, respectively, where $\top$ denotes the transposition of the vector. Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the static and dynamic differential feature vectors, where $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$. A joint probability density between the source and differential spectral features modeled by the differential GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{D}_t | \boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(D)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XD)} \\ \boldsymbol{\Sigma}_m^{(DX)} \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right) \quad (1)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component.

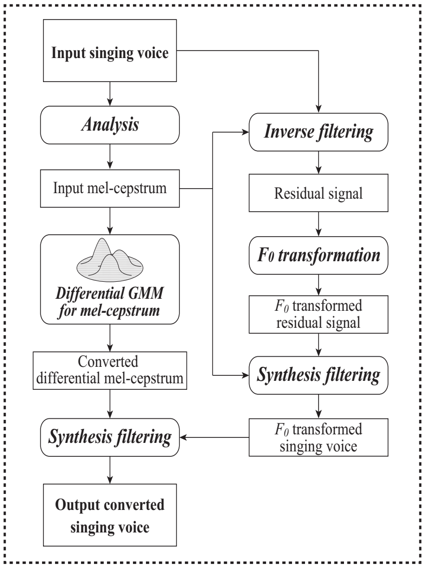
Input singing voice
Analysis
Inverse filtering
Input mel-cepstrum
Residual signal
Differential GMM for mel-cepstrum
F0 transformation
F0 transformed residual signal
Converted differential mel-cepstrum
Synthesis filtering
Synthesis filtering
F0 transformed singing voice
Output converted singing voice

Fig. 1. Conversion process considering F_0 transformation in DIFFSVC.

In the conversion process, the converted spectral feature differential sequence is estimated from the source feature sequence based on the differential GMM in the same manner as the maximum likelihood estimation of speech parameter trajectory with the traditional GMM [4]. A time sequence vector of the source static and dynamic features and that of the static and dynamic feature differential are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{D} = [\boldsymbol{D}_1^\top, \cdots, \boldsymbol{D}_T^\top]^\top$, respectively, where T is the number of frames over a time sequence. A time sequence vector of the converted spectral feature differential $\hat{\boldsymbol{d}} = [\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d} \quad (2)$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature sequence vector into the joint static and dynamic feature sequence vector [19]. In order to alleviate the over-smoothing effect often observed in the converted spectral sequence, we can also consider the GV of the target spectral features [17] in Eq. (2). Finally, voice timbre of the source singer is converted into that of the target singer by directly filtering a speech waveform of the input natural singing voice with the converted spectral feature differential.

3. PROPOSED F_0 TRANSFORMATION METHOD FOR DIFFSVC

To make it possible to apply the DIFFSVC method to the cross-gender SVC as well, we propose an F_0 transformation method based on direct waveform modification without using the vocoder-based waveform generation. Figure 1 shows a conversion process of the DIFFSVC with the proposed F_0 transformation. To avoid a change of spectral envelope before and after the F_0 transformation, the proposed method modifies not the input singing voice waveform directly but its residual signal. Three basic modification processes are implemented: 1) F_0 transformation based on time-scaling with waveform similarity-based overlap-add (WSOLA) [20] and resampling, 2) generation of high frequency components, and 3) comb filtering [21] to suppress aperiodic components. The second and third processes are needed when transforming F_0 to lower values, e.g., in female-to-male SVC. In this paper, we assume that F_0 is transformed with a constant rate, e.g., one octave is always increased in male-to-female SVC. Figure 2 shows an example of a spectrogram of an input natural singing voice and these of signals generated in the proposed F_0 transformation process.

3.1. F_0 transformation based on WSOLA and resampling

In this paper, we use time-scaling with WSOLA and resampling for transforming F_0. For example, if we double F_0 to increase one octave in male-to-female SVC, we halve duration of the input signal with WSOLA while keeping F_0 and voice timbre, and then we perform up-sampling to restore the original duration. If we decrease F_0 to half in female-to-male SVC, we double the duration of the input signal with WSOLA, and then we perform down-sampling. This process is equivalent to expand or shrink a frequency axis. Therefore, it also transforms voice timbre (i.e., spectral envelope) as well as F_0. This is quite inconvenient in the DIFFSVC method because the differential GMM also needs to be modified to deal with the change of spectral envelope.

To keep spectral envelope unchanged, we apply the F_0 transformation process based on WSOLA and resampling to a residual signal of the input natural singing voice. First, an inverse filter based on mel-cepstrum extracted from the input natural singing voice is applied to its waveform signal for extracting its residual signal, which spectral envelope is basically flat but still consists of harmonic and aperiodic components (shown in Fig. 2 (b)). Then, the F_0 transformation based on WSOLA and re-sampling is performed to generate the F_0 transformed residual signal (shown in Fig. 2 (c)). Finally, it is filtered with the mel-cepstrum to generate the F_0 transformed input singing voice while preserving its spectral envelope.

3.2. Generation of high frequency components

When F_0 is decreased in female-to-male SVC, high frequency components of the residual signal are lost (shown in Fig. 2 (c)). It is obvious that the use of this residual signal causes adverse effects. As high-frequency components of a speech signal tend to be less periodic and be well modeled with noise components, we generate them using a noise excitation signal. First, the high-pass filtered noise excitation signal is added to the F_0 transformed residual signal (shown in Fig. 2 (d)). Then, the resulting residual signal with the generated high frequency components is filtered with the mel-cepstrum to generate the F_0 transformed input singing voice (shown in Fig. 2 (e)).

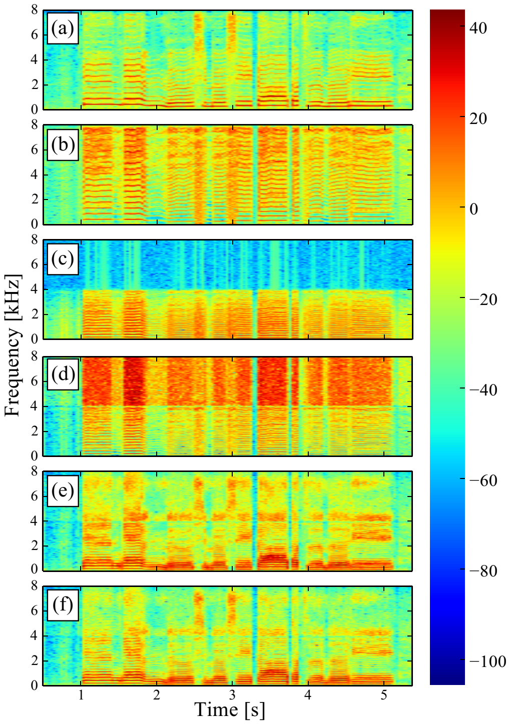

Fig. 2. An example of spectrograms of several signals: (a) input natural singing waveform, (b) residual signal, (c) F_0 transformed residual signal, (d) F_0 transformed residual signal after noise component generation in high frequency bands, (e) F_0 transformed singing voice, and (f) F_0 transformed singing voice after comb filtering.

3.3. Suppression of aperiodic components

Another adverse effect caused by decreasing F_0 in female-to-male SVC is that aperiodic components in high frequency bands are shifted to in low frequency bands (e.g., aperiodic components in 4–8 kHz frequency bands are shifted to in 2–4 kHz frequency bands if F_0 is decreased to half). Because aperiodic components in high frequency bands are significantly higher than those in low frequency bands, this shift tends to make the F_0 transformed input singing voice sound too husky. To address this issue, we use the feed-back comb filter to suppress aperiodic components. Its transfer function at frame t is given by

$$H_t(z) = \frac{1-a}{1-az^{-(f_s/f_t)}} \tag{3}$$

where a is a parameter to control a frequency response of the comb filter. f_s and f_t denote a sampling frequency of the input waveform and the converted F_0 value at frame t, respectively. Figure 3 shows an example the frequency response of the comb filter. In this paper, we apply the comb filter to the F_0 transformed input singing voice. An example of the F_0 transformed input singing voice after filtering is shown in Fig. 2 (f). Note that the comb filter is not applied to unvoiced frames.

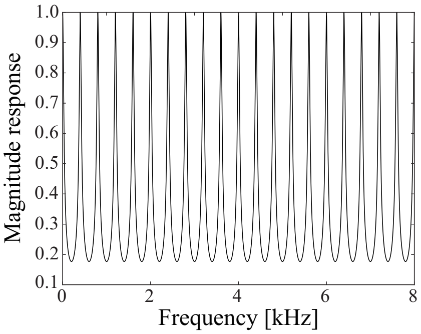

Fig. 3. Magnitude response of the comb filter (setting $a = 0.7$, $f_s = 16000$, $f_t = 400$ in Eq. (3)).

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We evaluated sound quality and singer identity of the converted singing voices to compare the conventional SVC method with vocoder-based waveform generation [4] and the proposed DIFFSVC method in cross-gender SVC. We used singing voices of 21 Japanese traditional songs, which were divided into 152 phrases, where the duration of each phrase was approximately 8 seconds. 3 males and 3 females sang these phrases. The sampling frequency was set to 16 kHz.

STRAIGHT [12] was used to extract spectral envelopes, which was parameterized to the 1-24th mel-cepstral coefficients as the spectral feature. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [22] was used as the synthesis filter.

We used 80 randomly selected phrases for the GMM training and the remaining 72 phrases were used for evaluation. The speaker-dependent GMMs were separately trained for all cross-gender singer pairs. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64. We transformed F_0 by 2.0 times for male-to-female conversion and 0.5 times for female-to-male conversion. The parameter a for comb filter was set to 0.6.

Two preference tests were conducted. In the first test, sound quality of the converted singing voices was evaluated. The converted singing voice samples of the conventional SVC and proposed DIFFSVC methods for the same phrase were presented to listeners in random order. The listeners selected which sample had better sound quality. In the second test, conversion accuracy in singer identity was evaluated. A natural singing voice sample of the target singer was presented to the listeners first as a reference. Then, the converted singing voice samples of the conventional SVC and proposed DIFFSVC methods for the same phrase were presented in random order. The listeners selected which sample was more similar to the reference natural singing voice in terms of singer identity. The number of listeners was 8 and each listener evaluated 72 sample pairs. They were allowed to replay each sample pair as many times as necessary.

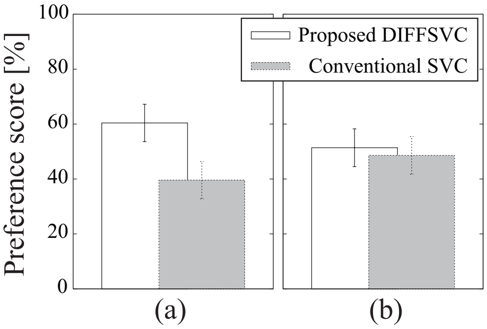

Fig. 4. Results of preference tests on (a) sound quality of converted singing voice and (b) conversion accuracy of singer individuality.

4.2. Experimental results

Figure 4 (a) indicates the result of the preference test for the sound quality. The proposed DIFFSVC method generates the converted speech with better sound quality than the conventional SVC method. This is because the proposed DIFFSVC method can effectively avoid several errors caused by using the vocoder-based waveform generation. Figure 4 (b) indicates the result of the preference test for the singer identity. The conversion accuracy of the singer identity of the proposed DIFFSVC method is not significantly different from that of the conventional SVC method. These results demonstrate that the proposed DIFFSVC method is capable of converting voice timbre and transforming F_0 with higher sound quality while causing no degradation in the conversion accuracy of singer identity compared to the conventional SVC method. As shown in [17] these results are almost the same as in intra-gender DIFFSVC. Therefore, they reveal that the DIFFSVC can also be applied to cross-gender SVC by using the proposed F_0 transformation while preserving its advantages.

5. CONCLUSION

In order to apply statistical singing voice conversion based on direct waveform modification (DIFFSVC) to cross-gender singer conversion, we have proposed a F_0 transformation method for the DIFFSVC method, which is also based on direct waveform modification with overlap-add and filtering processes. The experimental results have demonstrated that the proposed DIFFSVC method makes it possible to convert voice timbre and F_0 with higher sound quality while not causing any adverse effects on the conversion accuracy of singer identity compared to the conventional SVC method based on vocoder-based waveform generation even in the cross-gender singer conversion. We plan to also implement a conversion process of aperiodic components into the DIFFSVC framework.

6. ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Numbers: 26280060, and by the JST OngaCREST project.

7. REFERENCES

[1] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," *Proc. INTERSPEECH*, pp. 2162–2165, Sept. 2010.

[2] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech (Japanese edition)*, vol. 110, no. 297, pp. 71–76, Nov. 2010.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[5] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, pp. 1249–1252, Apr. 2007.

[6] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *Proc. INTERSPEECH*, pp. 1623–1626, Sept. 2009.

[7] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," *Proc. APSIPA ASC*, Nov. 2012.

[8] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive voice-quality control based on one-to-many eigenvoice conversion," *Proc. INTERSPEECH*, pp. 2158–2161, Sept. 2010.

[9] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G.Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *IEICE Trans. on Inf. and Syst.*, vol. 97, no. 6, pp. 1419–1428, 2014.

[10] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, Sept. 2012.

[11] H. Dudley, "Remaking speech," *JASA*, vol. 11, no. 2, pp. 169–177, 1939.

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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, no. 3-4, pp. 187–207, Apr. 1999.

[13] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," *Proc. SMAC*, pp. 287–292, Aug. 2013.

[14] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. SAP*, vol. 9, no. 1, pp. 21–29, 2001.

[15] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J-STSP*, vol. 8, no. 2, pp. 184–194, 2014.

[16] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," *Proc. INTERSPEECH*, pp. 2514–2418, Sept. 2014.

[17] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," *Proc. INTERSPEECH*, pp. 2754–2758, Sept. 2015.

[18] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, vol. 1, pp. 285–288, 1998.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp. 1315–1318, June 2000.

[20] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *Proc. ICASSP*, pp. 554–557 vol.2, Apr. 1993.

[21] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *JASA*, vol. 60, no. 4, pp. 911–918, 1976.

[22] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," *Proc. ICSLP*, pp. 1043–1045, 1994.