AN EVALUATION OF EXCITATION FEATURE PREDICTION IN A HYBRID APPROACH TO ELECTROLARYNGEAL SPEECH ENHANCEMENT

Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

ABSTRACT

We implement removing micro-prosody with low-pass filtering and avoiding Unvoiced/Voiced (U/V) prediction as part of a hybrid approach to improve statistical excitation prediction in electrolaryngeal (EL) speech enhancement. An electrolarynx is a device that artificially generates excitation sounds to enable laryngectomees to produce EL speech. Although proficient laryngectomees can produce quite intelligible EL speech, it sounds very unnatural due to the mechanical excitation produced by the device. Moreover, the excitation sounds produced by the device often leak outside, adding noise to EL speech. To address these issues, in our previous work, we proposed a hybrid method using a noise reduction method for enhancing spectral parameters and voice conversion method for predicting excitation parameters. In this paper, we evaluate the effect of removing micro-prosody with low-pass filtering and avoiding U/V prediction in the hybrid enhancement process.

Index Terms— speaking aid, electrolaryngeal speech, hybrid approach, statistical excitation prediction, unvoiced/voiced information

1. INTRODUCTION

Speech is one of the most common media of human communication. Unfortunately, there are many people with disabilities that prevent them from producing speech freely, leading to communication barriers. One example of people who cannot produce speech freely are laryngectomees, who have undergone an operation to remove the larynx including the vocal folds for reasons such as an accident or laryngeal cancer. Larengectomees cannot produce speech in the usual manner because they no longer have their vocal folds. Electrolaryngeal (EL) speech is produced by one of the major alternative speaking methods for laryngectomees as shown in Figure 1. EL speech is produced using an electrolarynx, which is an electromechanical vibrator that is typically held against the neck to mechanically generate artificial excitation signals. The generated excitation signals are conducted into the speaker's oral cavity, and EL speech is produced by articulating the conducted excitation signals. Compared with other types of alaryngeal speech, EL speech is relatively intelligible. However, the excitation sounds are usually emitted outside as noise causing degradation of sound quality, and naturalness is very low owing to its mechanical sound quality caused by the mechanically generated excitation signals.

To address these issues of EL speech, two approaches have conventionally been adopted. One is based on noise reduction [1] [2] and the other is based on statistical voice conversion (VC) [3] [4]. The former approach aims to reduce the effect of the excitation sounds leaked from the electrolarynx by using noise reduction techniques, such as spectral subtraction (SS) [5]. This noise reduction process causes no degradation in intelligibility but yields only small improvements in naturalness as the mechanical excitation sounds remain essentially unchanged. On the other hand, the latter method is capable of significantly improving naturalness by converting acoustic parameters of EL speech into those of natural voices using statistical VC techniques [6] [7]. The use of statistics extracted from a parallel data set consisting of EL speech and natural voices makes it possible to achieve more complex conversion processes than those of other signal processing approaches, such as formant manipulation [8]. However, VC-based approaches usually cause degradation in intelligibility owing to errors in conversion [4].

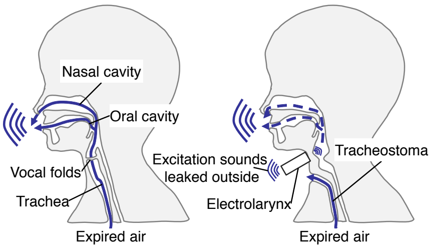

Fig. 1. Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

To significantly improve naturalness compared with EL speech while preserving intelligibility, as a first step we proposed a hybrid approach using SS-based noise reduction for enhancing spectral parameters and VC for predicting excitation parameters [9]. Although laryngectomees cannot produce excitation sounds in the usual manner, they can still articulate by changing the shape of their vocal tract. Thus, there is not a large difference between spectral parameters of natural speech and EL speech. On the other hand, the excitation parameters of EL speech are highly unnatural owing to mechanically generated excitation sounds, compared with those of natural speech. In order to develop an EL speech enhancement method that allows for the large improvements of naturalness realizable by VC while ameliorating its adverse effects, we propose a hybrid approach based on SS and VC. Moreover, to address the issue that modeling of discontinuous F_0 patterns is difficult [10], we also proposed the use of continuous F_0 pattern without any unvoiced frames to generate the excitation signals [11]. However, compared with excitation parameters of the natural voice, excitation parameters of the enhanced voice are still degraded owing to U/V prediction errors in the VC-based enhancement process. In addition, we noticed that attempting to model micro-prosody, rapid movements that cannot be modeled accurately with a GMM, causes an adverse effect on model training.

In this paper we make two changes to the parameter generation process to overcome these obstacles. First, we propose removing micro-prosody to improve prediction accuracy. Second, to improve prediction accuracy and reduce adverse effects caused by U/V prediction errors, we improve and evaluate the continuous F0 method we preliminarily proposed in [9]. We conduct an experimental evaluation, including dictation tests, preference tests, and objective test of excitation parameter prediction.

Part of this work was supported by JSPS KAKENHI Grant Number 22680016.

2. HYBRID APPROACH TO EL SPEECH ENHANCEMENT

The hybrid approach [9] adopts SS-based noise reduction for enhancing spectral parameters and VC method for predicting excitation parameters, as shown in Figure 2.

SS [12] is a method for restoration of the amplitude spectrum of a speech signal that has been observed with additive noise. In this method, assuming that the additive noise signal is stationary, the enhanced clean speech component $|\hat{S}(\omega,t)|$ is extracted through subtraction of the averaged amplitude spectrum of the noise $|\hat{L}(\omega)|$ from the amplitude spectrum of the noisy speech signal $|Y(\omega,t)|$ as follows:

$$|\hat{S}(\omega,t)|^{\gamma} = \begin{cases} |Y(\omega,t)|^{\gamma} - \alpha|\hat{L}(\omega)|^{\gamma} & \left(\frac{|\hat{L}(\omega)|^{\gamma}}{|Y(\omega,t)|^{\gamma}} < \frac{1}{\alpha}\right) \\ 0 & (otherwise) \end{cases} \tag{1}$$

where ω is frequency, t is time frame, α $(\alpha > 1)$ is an over-subtraction parameter, and γ is an exponential domain parameter.

VC [3] attempts to convert EL speech of laryngectomees into normal speech of non-disabled speakers. It consists of training and conversion processes. In training process, the conversion models are constructed by using parallel data of EL speech and normal speech. Let us assume the spectral segment features of EL speech $\boldsymbol{X}_t$, extracted from $\pm C$ frames around current frame, and a static feature vector $\boldsymbol{y}_t$ of each type of the normal speech parameters at frame t. As an output speech feature vector, we use $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of the static and dynamic features, where $\top$ denotes transposition of the vector. We independently train GMMs to model the joint probability densities [13] of the spectral segment feature of EL speech and U/V information, log F_0 values and aperiodic components [14] of the output feature vectors of individual target parameters of normal speech using the corresponding joint feature vector set as follows:

$$\begin{aligned} &P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) \\ &\quad = \textstyle\sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \end{aligned} \tag{2}$$

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. The parameter set of the GMM is $\boldsymbol{\lambda}$, which consists of mixture-component weights α_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. Note that the U/V information is expressed by log F_0 values at voiced frames and by ZERO values at unvoiced frames. Log F_0 values are values of continuous F_0 patterns generated by using spline interpolation to produce F_0 values at unvoiced frames.

In the conversion process, individual speech parameters of the target normal speech are independently estimated from the spectral segment features extracted from the EL speech using each of the trained GMMs as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{3}$$

where $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$, $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$, and $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ are time sequence vectors of the input spectral segment features, the output features, and the converted static features of each target speech parameter over an utterance, respectively. The matrix $\boldsymbol{W}$ is a transform to extend the static feature vector sequence into the joint static and dynamic feature vector sequence [15]. Note that the log F_0 values of the enhanced voice are estimated based on U/V information predicted by a GMM used for predicting U/V information. After estimating time sequences of the converted F_0 and aperiodic components, a mixed excitation signal is generated using the converted F_0 and aperiodic components [16]. Finally, the enhanced speech signal is synthesized by filtering the generated excitation signal based VC with the parameterized spectral parameters enhanced by SS.

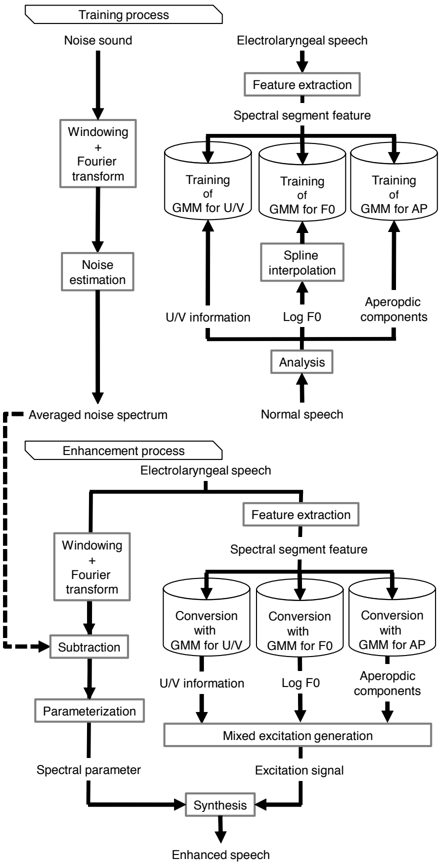

Fig. 2. EL speech enhancement based on a hybrid approach.

3. IMPROVEMENT OF STATISTICAL EXCITATION PREDICTION

In addition to the overall hybrid framework, we propose two improvements to statistical excitation prediction in the enhancement process.

3.1. Removing Micro-Prosody with Low-Pass Filter (LPF)

Rapid movements, called micro-prosody, are often observed in F_0 patterns extracted from natural voices. However, it is difficult to accurately model and reproduce these movements with a GMM. Moreover, the impact of micro-prosody on naturalness of synthetic speech is much smaller than that of F_0 patterns corresponding to phrase and accentual components. Therefore, it is helpful to make the GMM focus on modeling only these patterns. To achieve this, we propose the use of a method to smooth the continuous F_0 patterns with low-pass filtering [17] as shown in the bottom of Figure 3. The smoothed continuous F_0 patterns are then modeled with the GMM.

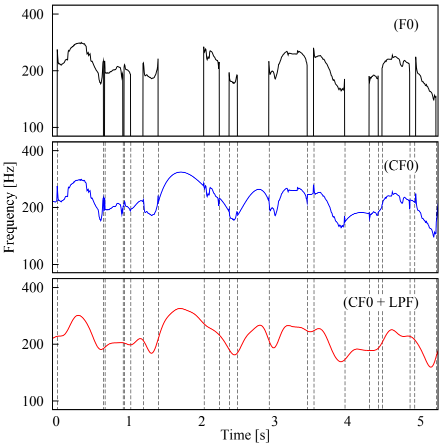

Fig. 3. Each type of F_0 pattern. The top figure is a target F_0 pattern, the middle is a continuous F_0 pattern using spline interpolation, and the bottom is a continuous F_0 pattern smoothed using the low-pass filter (cut-off frequency = 10 Hz).

3.2. Avoiding U/V Prediction Errors

In the excitation parameter prediction, U/V information is also predicted as mentioned above. Errors during this prediction process are also unavoidable and they may cause adverse effects in intelligibility.

As EL speech is totally voiced speech, no degradation is caused even if the converted speech is generated by regarding all speech frames as voiced frames. To further reduce the possibility of degradation in intelligibility caused by U/V prediction errors, we also propose the use of continuous F_0 patterns without any unvoiced frames for speech segments to generate the excitation signals. In the conversion process, continuous F_0 patterns are predicted over all frames. Then, only silence frames are automatically detected using waveform power and unvoiced excitation signals are generated only at those frames. Unvoiced phoneme sounds cannot be generated in this method, as in the original EL speech, but the converted speech does not suffer from wrongly predicted unvoiced frames. Note that the difference between this paper and [9] is a more comprehensive evaluation of the performance.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental Conditions

We conducted two objective tests for the excitation parameter prediction, and as well as a subjective evaluation consisting of a dictation test on intelligibility and a preference test on naturalness. In our experiments, the source speaker was one laryngectomee and the target speaker was one non-disabled speaker. Both speakers recorded 50 phoneme-balanced sentences. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance-pairs were used for evaluation. Sampling frequency was set to 16 kHz. In the VC-based enhancement methods, the 0th through 24th mel-cepstral coefficients extracted by STRAIGHT analysis [18] were used as the spectral parameters. The shift length was set to 5 ms. For the segment feature extraction, current ± 4 frames were used. In the VC-based enhancement method, the numbers of mixture components were set to 16 for the aperiodic estimation, and the aperiodic distortion was 3.19 dB.

In the objective tests, we evaluated the effect of removing micro-prosody with LPF for the training data. As measures to evaluate the prediction accuracy of the excitation features, we used the correlation coefficient and U/V error rate on F_0 components between the converted speech parameters and the natural target speech parameters. Note that we set the cut-off frequency of LPF to 5 Hz, 10 Hz, 20 Hz, or 40 Hz, and also set the number of GMM mixture components for F_0 estimation and U/V prediction to 8, 16, 32, or 64.

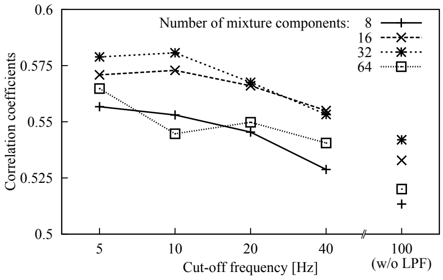

Fig. 4. Relationship between cut-off frequency of LPF and F_0 correlation coefficients.

In the dictation test, in order to demonstrate the effect of avoiding U/V prediction errors on intelligibility, we evaluated the following five types of speech samples:

EL original EL speech

SS speech enhanced by the SS-based enhancement method

Hybrid (V) speech enhanced by the proposed hybrid enhancement method without U/V prediction

Hybrid (U/V) speech enhanced by the proposed hybrid enhancement method with U/V prediction

Hybrid (target U/V) speech enhanced by the proposed hybrid enhancement method with ideal U/V information

where the proposed hybrid enhancement method was the method based on SS+VC+CF0+LPF. As the ideal U/V information, we used target U/V information obtained by performing DTW between the enhanced speech parameters using the VC-based enhancement method and the natural target speech parameters. Note that the VC-based enhancement method generally causes a significant degradation (around 3% recognition rate reduction) in intelligibility compared with EL speech as reported in [19]. In the preference test, in order to demonstrate the effect of avoiding U/V prediction errors on naturalness, we evaluated the following three types of speech samples:

Hybrid (V)

Hybrid (U/V)

Hybrid (target U/V)

All tests were performed by 5 listeners. Each listener evaluated 10 samples per system. Note that, in these tests, we set the cut-off frequency of LPF to 10 Hz, and the number of GMM mixture components for F_0 estimation and U/V prediction to 32.

4.2. Experimental Results

Figure 4 shows the result of the evaluation for changing cut-off frequency of LPF to 5 Hz, 10 Hz, 20 Hz, or 40 Hz. We achieved an improvement in the accuracy of F_0 estimation, thanks to removing micro-prosody. From the results, we can see that the optimal number of GMM mixture components and cut-off frequency with LPF is 32 and 10 Hz, respectively. We can also see that rapid movements, such as micro-prosody, caused degradation of the prediction accuracy at more than 10 Hz.

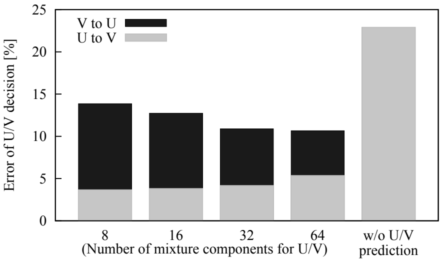

Fig. 5. U/V error rate of each system.

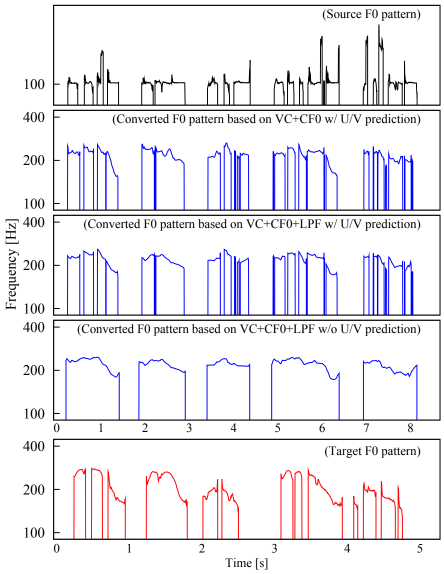

Fig. 6. Each type of F_0 pattern.

Figure 5 shows the result of the evaluation for U/V error rate. As the number of mixture components grows larger, V-to-U error rate decreases while U-to-V increases. With 64 mixture components, the U/V error rate is minimized. On the other hand, without U/V prediction, the U/V error rate is constant. In particular, the V-to-U error rate is practically zero. In actuality, the V-to-U errors still exist with the continuous F_0 estimation method owing to errors in the automatic silence frame detection with waveform power, but they are almost negligible. However, U-to-V errors significantly increase in the continuous F_0 estimation method. Note that as we mentioned in Section 3.2, this increase causes no adverse effect compared with EL speech because EL speech is totally voiced speech.

Figure 7 shows a result of the dictation test on intelligibility. We found that the hybrid methods do not cause any degradation in intelligibility compared with EL speech. Furthermore, in the hybrid method avoiding U/V prediction by using the continuous F_0 estimation method, the intelligibility is preserved, similarly to the hybrid method using ideal U/V information. Hence, it can be said that U/V prediction is not always required. On the other hand, the hybrid methods tend to degrade intelligibility slightly compared to SS, owing to several issues, such as the effect of synthesis using a vocoder and using 24-dimensional mel-cepstral coefficients as spectral features. We show samples of F0 patterns in Figure 6

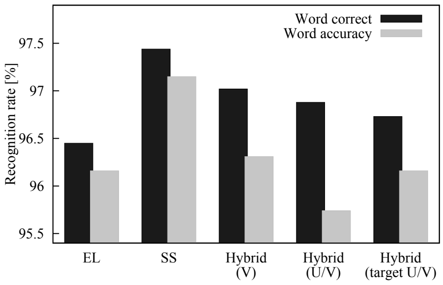

Fig. 7. Result of dictation test on intelligibility.

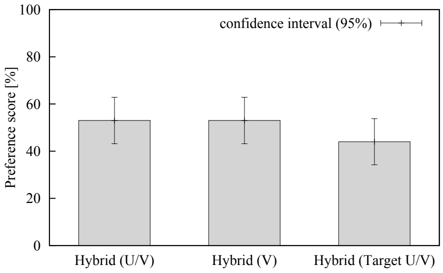

Fig. 8. Result of preference test on naturalness.

Figure 8 shows a result of the preference test on naturalness. We found that the avoiding U/V prediction in the hybrid methods does not cause any degradation in naturalness compared with the use of ideal U/V information in the hybrid methods. Therefore, we can avoid the U/V prediction process without adverse effects. On the other hand, we also found that the use of predicted U/V information by VC-based method does not cause any degradation in naturalness compared with the use of ideal U/V information in the hybrid methods. Due to the inconsistency between the excitation signal created by the VC-based method and the spectral parameters created by the SS-based method, when using ideal U/V information we observe no clear improvement in naturalness, which is in contrast to the results for the other method. Because EL speech is totally voiced, the spectral parameters enhanced by the SS-based method are also entirely voiced parameters. On the other hand, as for the generated excitation signal based on VC-based method, because excitation parameters are predicted using statistics of normal speech, excitation parameters are voiced or unvoiced parameters. Hence, the use of ideal U/V information in the hybrid methods does not perform well due to the inconsistency between spectral parameters at voiced frames and excitation parameters at unvoiced frames.

5. CONCLUSION

In this paper, we proposed a method of removing micro-prosody with LPF as part of the hybrid approach to further improve the excitation feature estimation. Moreover, we evaluate the effect of avoiding U/V prediction errors that cause degradation in intelligibility. As a result of an experimental evaluation, it has been demonstrated that removing micro-prosody is capable of improving the excitation feature estimation. Furthermore, in the hybrid method that avoids U/V prediction, the intelligibility and naturalness is maintained to the level of the hybrid method using ideal U/V information. Hence, it can be said that U/V prediction is not always required in EL voice enhancement.

6. REFERENCES

[1] H. Liu, Q. Zhao, M.X. Wan, and S.P. Wang, “Enhancement of electrolarynx speech based on auditory masking,” *IEEE Trans. Biomedical Engineering,* vol. 53, no. 5, pp. 865–874, May 2006.

[2] S.K. Basha and P.C. Pandey, “Real-Time Enhancement of Electrolaryngeal Speech by Spectral Subtraction,” *Proc. NCC,* 1569507449, pp. 516–520, Feb, 2012.

[3] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” *SPECOM,* vol. 54, no. 1, pp. 134–146, Jan 2012.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques,” *Proc. ICASSP,* pp. 5136–5139, May 2011.

[5] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. 27, no. 2, pp. 113–120, Apr 1979.

[6] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Processing,* vol. 6, no. 2, pp. 131–142, Mar 1998.

[7] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech, and Language,* Vol. 15, No. 8, pp. 2222–2235, Nov 2007.

[8] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi, “Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec,” *IEEE Trans. Biomedical Engineering,* vol. 57, no. 10, pp. 2448–2458, Oct 2010.

[9] K. Tanaka, T. Toda, G .Neubig, S. Sakti, and S. Nakamura, “A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Spectral Subtraction and Statistical Voice Conversion,” *Proc. INTERSPEECH,* pp.3067–3071, Aug. 2013.

[10] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Trans. Audio, Speech, and Language,* Vol. 20, No. 9, pp. 2505–2517, Nov 2012.

[11] K. Yu and S. Young, “Continuous F_0 modelling for HMM based statistical parametric speech synthesis,” *IEEE Trans. Audio, Speech, and Language,* Vol. 19, No. 5, pp. 1071–1079, Jul 2011.

[12] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, “A parametric formulation of the generalized spectral subtraction method,” *IEEE Trans. Speech and Audio Processing,* Vol. 6, No. 4, pp. 328–337, Jul 1998.

[13] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proc. ICASSP,* pp. 285–288, May 1998.

[14] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system STRAIGHT,” *Proc. 2nd MAVEBA,* Sep 2001.

[15] 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.

[16] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” *Proc. Interspeech,* pp. 2266–2269, Sep 2006.

[17] A. Sakurai and K. Hirose, “Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours,” *Proc. ICSLP,* Vol. 2, pp. 817–820, Oct 1996.

[18] H. Kawahara, I. Masuda-Katsuse, and A. 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,” *SPECOM,* Vol. 27, No. 3-4, pp. 187–207, Apr 1999.

[19] H. Doi., “Augmented speech production beyond physical constraints using statistical voice conversion -Alaryngeal speech enhancement and singing voice quality control-,” *NAIST Doctoral Dissertation,* NAIST-IS-DD1061014, Mar 2013.