A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Spectral Subtraction and Statistical Voice Conversion

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

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

{ko-t, tomoki, ssakti, neubig, s-nakamura}@is.naist.jp

Abstract

We present a hybrid approach to improving naturalness of electrolaryngeal (EL) speech while minimizing degradation in intelligibility. 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, previous work has proposed methods for EL speech enhancement through either noise reduction or voice conversion. The former usually 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 significantly improves naturalness of EL speech using spectral and excitation parameters of natural voices converted from acoustic parameters of EL speech, but it usually causes degradation in intelligibility owing to errors in conversion. We propose a hybrid method using the noise reduction method for enhancing spectral parameters and voice conversion method for predicting excitation parameters. The experimental results demonstrate the proposed method yields significant improvements in naturalness compared with EL speech while keeping intelligibility high enough.

Index Terms: speaking-aid, electrolaryngeal speech, spectral subtraction, voice conversion, hybrid approach

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. Therefore, they require another method to produce speech without the vocal fold vibration.

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. There are several advantages of EL speech compared with other types of alaryngeal speech, such as esophageal speech: e.g., 1) it is easy to learn how to produce EL speech, 2) less physical power is needed to produce EL speech, and 3) EL speech is relatively intelligible. However, there are also some issues of EL speech: e.g., 1) the excitation sounds are usually emitted outside as noise causing degradation of sound quality, and 2) naturalness is low owing to its mechanical sound quality caused by the mechanically generated excitation signals. In particular, the latter issue is an essential drawback of EL speech caused by the difficulty of artificially generating natural F_0 patterns corresponding to linguistic content.

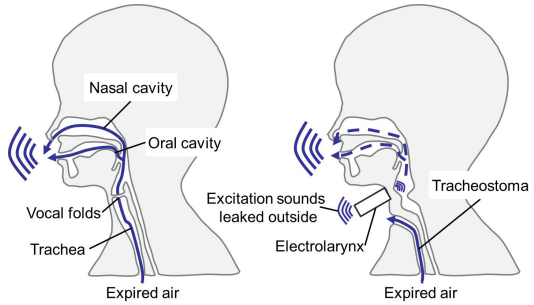

Figure 1: Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

To address these issues of EL speech, two main approaches have been proposed. One is based on noise reduction [1] and the other is based on statistical voice conversion (VC) [2] [3]. 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) [4]. 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 [5] [6]. 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 that of other signal processing approaches, such as formant manipulation [7]. For example, it is possible to convert from a spectral parameter sequence of EL speech into F_0 patterns of natural voices. However, VC-based approaches usually cause degradation in intelligibility owing to errors in conversion [3].

In this paper, to develop an EL speech enhancement method for significantly improving naturalness while preserving intelligibility in EL speech, we propose a hybrid method using the SS-based noise reduction method for enhancing spectral parameters and the VC method for predicting excitation parameters. Furthermore, to avoid degradation in intelligibility caused by unvoiced/voiced prediction errors, we also propose an estimation method of continuous F_0 patterns. We conduct an experimental evaluation, which demonstrates that the proposed method

yields significant improvements in naturalness compared with EL speech while causing no degradation in intelligibility.

2. Electrolaryngeal Speech Enhancement Based on Spectral Subtraction (SS)

SS is a method for restoration of the amplitude spectrum of a speech signal that has been observed with additive noise. This is done through subtraction of an estimate of the amplitude spectrum of the noise from the amplitude spectrum of the noisy speech signal. The noisy speech signal model in the frequency domain is expressed as follows:

$$Y(\omega,t) = S(\omega,t) + L(\omega,t) \tag{1}$$

where $Y(\omega,t)$, $S(\omega,t)$, and $L(\omega,t)$ are respectively components of the noisy speech signal, the clean speech signal, and the additive noise signal at frequency ω and time frame t. Assuming that the additive noise signal is stationary, the generalized SS scheme [8] is described as follows:

$$|\hat{S}(\omega,t)|^{\gamma} = \begin{cases} |Y(\omega,t)|^{\gamma} - \alpha|\hat{L}(\omega)|^{\gamma} & (\frac{|\hat{L}(\omega)|^{\gamma}}{|Y(\omega,t)|^{\gamma}} < \frac{1}{\alpha+\beta}) \\ \beta|\hat{L}(\omega)|^{\gamma} & (otherwise) \end{cases} \tag{2}$$

where α ($\alpha > 1$) is an over-subtraction parameter, β ($0 \leq \beta \leq 1$) is a spectral flooring parameter, γ is an exponential domain parameter, and $\hat{L}(\omega)$ is an estimate of the averaged amplitude spectrum of the additive noise signal.

In this paper, we implement SS for EL speech enhancement, as shown in Figure 2. The averaged amplitude spectrum of the additive noise signal is estimated in advance using the excitation signals generated from the electrolarynx. In order to record only the excitation signals leaked from the electrolarynx as accurately as possible, the excitation signals are recorded with a close-talking microphone while keeping speaker's mouth closed. The excitation signals are generated with the electrolarynx held in the usual manner, as shown in Figure 1.

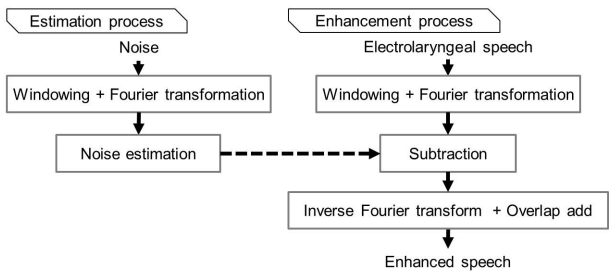

Figure 2: EL speech enhancement based on SS.

3. Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion (VC)

EL speech enhancement based on VC [2] attempts to convert EL speech of laryngectomees into normal speech of non-disabled speakers. It consists of training and conversion processes, as shown in Figure 3. To achieve the conversion from EL speech into normal speech, three conversion models are used to separately estimate spectrum, F_0, and aperiodic components, which capture the noise strength of an excitation signal on each frequency band [9]. These models are trained in advance using a parallel data set consisting of utterance pairs of a laryngectomee and a target non-disabled speaker. Conversion employs maximum likelihood estimation of speech parameter trajectories considering global variance (GV) [6]. This framework is the same as in conversion from body-conducted unvoiced speech into normal speech [10].

3.1. Training Process

Let us assume the spectral segment features of EL speech $\boldsymbol{X}_t$ 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 three GMMs to model the joint probability densities [11] of the spectral segment feature of EL speech and each of the output feature vectors of individual target parameters of normal speech using the corresponding joint feature vector set as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^{\top}, \boldsymbol{Y}_t^{\top}]^{\top}; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \tag{3}$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{4}$$

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. A 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. The mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ consists of an input mean vector $\boldsymbol{\mu}_m^{(X)}$ and an output mean vector $\boldsymbol{\mu}_m^{(Y)}$. The covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ consists of input and output covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ and cross-covariance matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$. We also train a Gaussian distribution modeling the probability density of the GV for the spectrum parameter of the target normal speech.

3.2. 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:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})^{\omega} \quad \text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{5}$$

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 [12]. The GV probability density is given by $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$, where $\boldsymbol{v}(\boldsymbol{y})$ is the GV of the target static feature vector sequence $\boldsymbol{y}$ and $\boldsymbol{\lambda}^{(v)}$ is a parameter set of the Gaussian distribution for the GV. The GV likelihood weight is given by ω. The GV likelihood is usually considered only in the spectral estimation, i.e., ω is set to zero in the F_0 estimation and the aperiodic estimation. After estimating time sequences of the converted spectrum, F_0, and aperiodic components, a mixed excitation signal is generated using the converted F_0 and aperiodic components [13]. Finally, the converted speech signal is synthesized by filtering the generated excitation signal with the converted spectral parameters.

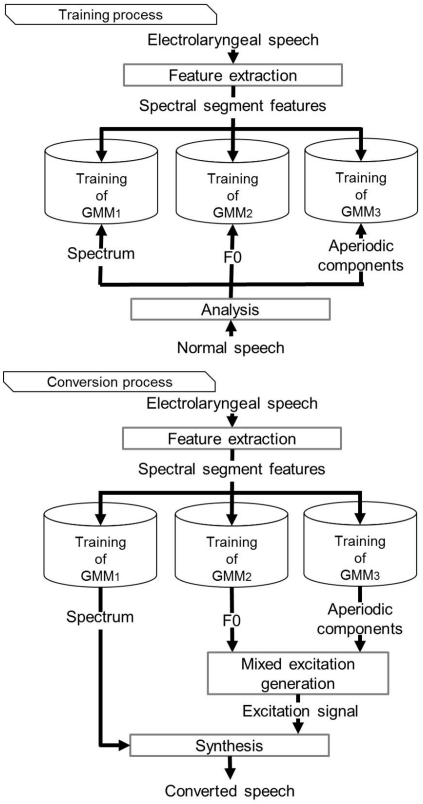

Figure 3: EL speech enhancement based on VC.

4. Hybrid Approach to Electrolaryngeal Speech Enhancement

The SS-based EL speech enhancement method essentially estimates EL speech produced by the lips while reducing the impact of leaked excitation sounds. Even if the leaked excitation sounds are completely removed, improvements in naturalness yielded by this method will be small because the produced EL speech intrinsically suffers from the lack of naturalness caused by highly artificial F_0 patterns and the mechanical excitation sound quality. On the other hand, this method does not cause any significant degradation in intelligibility of EL speech. In other words, this method may cause small improvements, but very rarely degradations in speech quality.

The VC-based EL speech enhancement method has the potential to significantly improving naturalness of EL speech by converting EL speech into normal speech. As the converted speech signal is generated from statistics of normal speech parameters, it does not suffer from the artificial F_0 patterns and mechanical sound quality. However, the conversion process in this method is quite complex, and therefore, errors in conversion are inevitable. These errors tend to cause degradation in intelligibility of converted speech as adverse effects.

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. The proposed EL speech enhancement method is shown in Figure 4. As laryngectomees have the capability to properly articulate the excitation signals, spectral parameters of EL speech do not need to be changed greatly to generate intelligible speech. Therefore, we use the spectral parameters refined with SS without applying VC. On the other hand, it is essentially difficult to generate natural excitation signals exhibiting natural F_0 patterns in EL speech production. Therefore, we use VC to estimate the excitation parameters: i.e., F_0 and aperiodic components. The proposed hybrid method can be expected to yield much larger improvements in naturalness compared with the SS-based enhancement method thanks to the use of more natural excitation signals generated from statistics of normal speech. It also can be expected to alleviate the degradation in intelligibility observed in the conventional VC-based enhancement method by avoiding errors in spectral conversion.

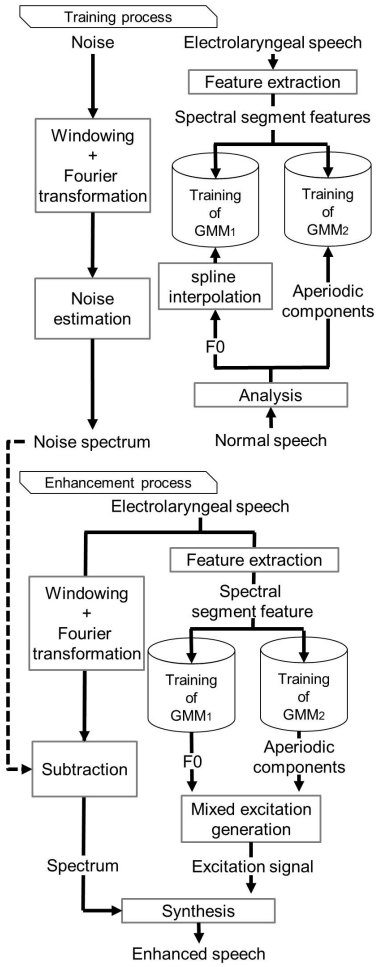

Figure 4: EL speech enhancement based on the proposed hybrid approach.

In the excitation parameter estimation based on VC, unvoiced/voiced (U/V) information is also predicted in the manner described in [10]. However, as EL speech is totally voiced speech, it is possible that significant improvements in naturalness can be yielded even if U/V information is not added to the converted speech. To further reduce the possibility of degradation in intelligibility caused by the U/V prediction errors, we also propose the use of continuous F_0 patterns without any unvoiced frames to generate the excitation signals. In the training process, continuous F_0 patterns of normal speech are generated by using spline interpolation to add F_0 values to unvoiced frames, and the GMM is trained on this modified data. In the conversion process, continuous F_0 patterns are estimated over all frames. As it is straightforward to automatically detect silence frames in EL speech simply using waveform power, unvoiced excitation signals are generated only at those frames. At the other voice active

frames, voiced excitation signals are always generated. Although unvoiced phoneme sounds cannot be generated in this method, the converted speech does not suffer from wrongly predicting voiced frames as unvoiced frames. Because unvoiced phoneme sounds also can not be generated in the original EL speech, this method causes no adverse effect.

5. Experimental Evaluation

5.1. Experimental Conditions

In our experiments, the source speaker was a laryngectomee and the target speaker was a 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 [14] were used as the spectral parameters. The shift length was set to 5 ms. For the segment feature extraction, current $\pm$ 4 frames were used. The numbers of mixture components were set to 32 for the spectral and aperiodic estimation, 64 for the F_0 estimation, and 32 for continuous F_0 estimation.

We conducted both objective and subjective evaluations. In the objective evaluation, conversion accuracy in the VC-based enhancement method was evaluated using mel-cepstral distortion, the U/V error rate, F_0 correlation coefficient, and aperiodic distortion between the converted speech parameters and the natural target speech parameters. In the subjective evaluations, we conducted two opinion tests of intelligibility and naturalness using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). The following five types of speech samples were evaluated by 5 listeners:

EL original EL speech

SS speech enhanced by the SS-based enhancement method

VC speech enhanced by the VC-based enhancement method

SS+VC speech enhanced by the proposed hybrid enhancement method with U/V prediction

SS+VC+CF0 speech enhanced by the proposed hybrid enhancement method with continuous F_0 estimation

5.2. Experimental Results

First, we show results of the objective evaluation for conversion accuracy for the enhancement methods with VC in Table 1. It can be observed that the F_0 correlation coefficient is improved slightly by the continous F_0 estimation. We have found that large errors in the F_0 estimation tend to be observed at short voiced segments that are sometimes generated in VC or SS+VC. This improvement is similar to that yielded by the continuous F_0 modeling in HMM-based speech synthesis [15]. Moreover, it is reasonable that the U to V error rate increases and the V to U error rate decreases in SS+VC+CF0 compared with those in VC or SS+VC. The V to U errors still exist in SS+VC+CF0 owing to errors in the automatic silence frame detection with waveform power but they are almost negligble.

Next, in Figure 5 we show the results of the subjective opinion test on intelligibility. It can be seen that a slight improvement is yielded by SS. On the other hand, VC causes significant degradation as reported in [3]. SS+VC doesn't cause degradation compared with EL but it still causes very small degradation compared with SS. This adverse effect on intelligibility is not observed in the proposed hybrid methods (SS+VC and SS+VC+CF0) thanks to no spectral conversion error.

Table 1: Conversion accuracy in enhancement methods with VC.

	VC	SS+VC	SS+VC+CF0
Mel-cepstral distortion without power infomation	5.09 dB	———	
U to V error rate	4.04 %		17.9 %
V to U error rate	4.42 %		0.09 %
F_0 correlation coefficient	0.53		0.55
Aperiodic distortion	3.19 dB		

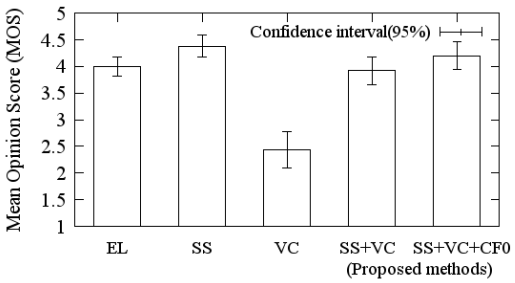

Figure 5: Result of opinion test on intelligibility.

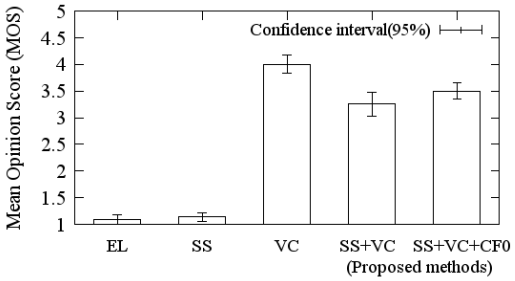

Figure 6: Result of opinion test on naturalness.

Figure 6 shows a result of the opinion test on naturalness. SS yields a very small improvement in naturalness. On the other hand, VC yields a significantly larger improvement. The proposed hybrid methods (SS+VC and SS+VC+CF0) also yield significantly larger improvements compared with SS as they are capable of generating more natural F_0 patterns. We can also observe that the continous F_0 estimation is effective for improving naturalness as well.

These results suggest that the proposed hybrid approach to EL speech enhancement based on the continous F_0 estimation is effective in significantly improving naturalness of EL speech while avoiding degradation in intelligibility that is often observed in the conventional VC-based enhancement method.

6. Conclusions

In this paper, we have proposed a hybrid approach to electrolaryngeal (EL) speech enhancement based on spectral subtraction for spectral parameter estimation and statistical voice conversion for excitation parameter prediction. To further avoid conversion errors causing degradation in intelligibility, the continuous F_0 estimation method has also been implemented for the proposed approach. As a result of an experimental evaluation, it has been demonstrated that the proposed approach is capable of significantly improving naturalness of EL speech while causing no adverse effect such as the degradation in intelligibility.

7. Acknowledgements

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

8. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.