Direct F_0 Control of an Electrolarynx based on Statistical Excitation Feature Prediction and its Evaluation through Simulation

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

10.21437/Interspeech.2014-7

Abstract

An electrolarynx is a device that artificially generates excitation sounds to enable laryngectomees to produce electrolaryngeal (EL) speech. Although proficient laryngectomees can produce quite intelligible EL speech, it sounds very unnatural due to the mechanical excitation produced by the device. To address this issue, we have proposed several EL speech enhancement methods using statistical voice conversion and showed that statistical prediction of excitation parameters, such as F_0 patterns, was essential to significantly improve naturalness of EL speech. In these methods, the original EL speech is recorded with a microphone and the enhanced EL speech is presented from a loudspeaker in real time. This framework is effective for telecommunication but it is not suitable to face-to-face conversation because both the original EL speech and the enhanced EL speech are presented to listeners. In this paper, we propose direct F_0 control of the electrolarynx based on statistical excitation prediction to develop an EL speech enhancement technique also effective for face-to-face conversation. F_0 patterns of excitation signals produced by the electrolarynx are predicted in real time from the EL speech produced by the laryngectomee's articulation of the excitation signals with previously predicted F_0 values. A simulation experiment is conducted to evaluate the effectiveness of the proposed method. The experimental results demonstrate that the proposed method yields significant improvements in naturalness of EL speech while keeping its intelligibility high enough.

Index Terms: laryngectomee, electrolarynx, electrolaryngeal speech, statistical excitation prediction, simulation evaluation

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. Laryngectomees 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. Naturalness is also very low owing to the mechanical sound quality and artificial fundamental frequency (F_0) patterns 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.

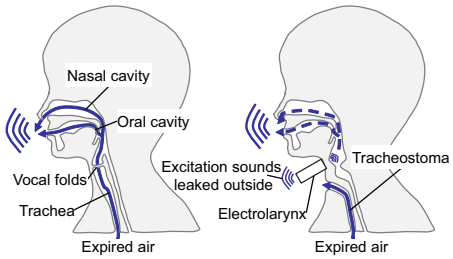

Figure 1: Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

To address these issues of EL speech, several EL speech enhancement methods have been proposed. These methods include enhancement methods based on a simple signal processing framework, *e.g.*, noise reduction to alleviate the issues of the leaked excitation [1] and rule-based formant manipulation in analysis-synthesis [2]. Recently, statistical approaches to EL speech enhancement have been proposed [3] to convert alaryngeal speech into target normal speech while keeping linguistic information unchanged. We recently proposed a hybrid approach [4] using noise reduction [1] [5] for enhancing spectral parameters and statistical voice conversion [6] [7] for predicting excitation parameters. Our experimental results demonstrated that the hybrid approach achieved significant improvements of naturalness while causing no degradation in intelligibility compared to the original EL speech. We have also found that the use of F_0 patterns statistically predicted from EL speech is very effective for improving naturalness of EL speech.

Traditional EL speech enhancement systems need to record the original EL speech with a microphone and present the enhanced EL speech with a loudspeaker. This enhancement process can be achieved in real time, and therefore it is very effective for facilitating human-to-human conversation. However, there is an essential drawback: both the original EL speech and the enhanced EL speech can be heard in the vicinity of the speaker. If these systems are used for telecommunication, no problem is caused because it is possible to present only the enhanced speech to the listener. On the other hand, this technology is not suitable for face-to-face conversation because the original EL speech is always heard by the listener.

In this paper, we propose an EL speech enhancement system effective for any situation, including face-to-face conver-

sation. F_0 patterns of the excitation signals produced by the electrolarynx are directly controlled using statistical excitation prediction. Namely, an F_0 value at a current frame is predicted in real time from the EL speech produced by the laryngectomee articulating the excitation signals with previously predicted F_0 values. Consequently, the proposed system has the potential to allow laryngectomees to directly produce enhanced EL speech with more natural F_0 patterns than the original EL speech, and present only the enhanced EL speech to the listener. As the first step toward implementation of the proposed system, a simulation experiment is conducted in this paper to demonstrate that the proposed system is capable of achieving significant improvements in naturalness of EL speech while preserving its high intelligibility.

2. Statistical Excitation Prediction

The proposed method uses a statistical voice conversion techniques [8] [9] to predict F_0 patterns of normal speech produced by a non-disabled person from spectral parameters of EL speech produced by a laryngectomee. It consists of training and prediction processes as shown in Figure 2. A conversion model to predict F_0 patterns is trained in advance using a parallel data set consisting of utterance pairs of EL speech by a laryngectomee and normal speech by a target non-disabled speaker. The prediction process is based on maximum likelihood estimation of speech parameter trajectories considering global variance (GV) [9].

2.1. Training Process

In the training process, first source and target features are extracted from the parallel data. As the source features, spectral segment features of EL speech are extracted from mel-cepstra at multiple frames around the current frame [10]. As the target features, F_0 values are extracted from natural speech. Continuous F_0 patterns [11] are generated from the originally extracted F_0 values by spline interpolation to produce F_0 values at unvoiced frames, and low-pass filtering is used to remove micro-prosody. The effectiveness of using continuous F_0 patterns in statistical excitation prediction was reported in [12].

We assume the spectral segment features of EL speech $\boldsymbol{X}_t$ and a log-scaled F_0 value y_t of normal speech at frame t. As an output feature vector, we use $\boldsymbol{Y}_t = [y_t, \Delta y_t]$ consisting of the static and dynamic features. We train a Gaussian mixture model (GMM) to model the joint probability density [13] of the source and target features using the corresponding joint feature vector set generated by performing automatic frame alignment for the parallel data set, which is given by:

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

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

where $\top$ denotes transposition. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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. The mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ consists of a source mean vector $\boldsymbol{\mu}_m^{(X)}$ and a target mean vector $\boldsymbol{\mu}_m^{(Y)}$. The covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ consists of source and target 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 F_0 patterns of the target normal speech [9].

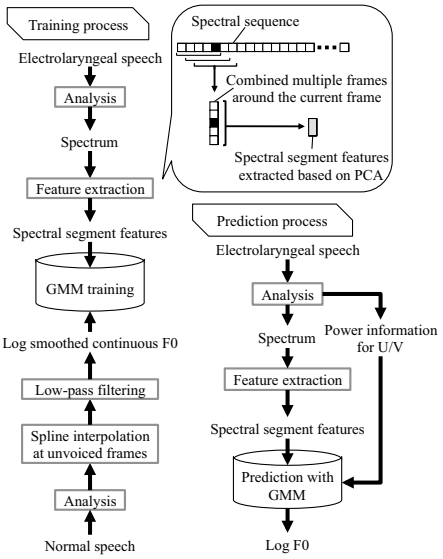

Figure 2: The training and prediction process.

2.2. Prediction Process

The continuous F_0 patterns of the target normal speech are predicted from the spectral segment features of EL speech using the trained GMM as follows:

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

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 $\hat{\boldsymbol{y}} = [\hat{y}_1, \cdots, \hat{y}_t, \cdots, \hat{y}_T]^\top$ are time sequence vectors of the spectral segment features, the joint static and dynamic target F_0 features, and the predicted F_0 features 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 [14]. 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. Finally, silence frames are automatically detected using waveform power and unvoiced excitation signals are generated only at those frames. Note that a real time prediction process can be achieved by using a computationally efficient real-time voice conversion method [15] based on the low-delay conversion algorithm to approximately solve Eq. (3) [16].

3. Statistical Method for Directly Controlling F_0 Patterns of Electrolarynx

3.1. Proposed System

The process of our proposed system is shown in Figure 3. This system allows laryngectomees to directly produce enhanced EL speech, and consists of two main processes: an articulation process and a prediction process. In the articulation process, the excitation signals generated from the electrolarynx are articu-

lated by laryngectomees in the same manner as the traditional speaking method using the electrolarynx. In the prediction process, F_0 patterns of the excitation signals are predicted from EL speech based on the real time statistical excitation prediction.

In this process, there are two main problems. One is misalignment between the articulated sounds and F_0 patterns. Real-time statistical excitation prediction causes a constant processing delay of 50 to 70 msec as reported in [15]. Namely, predicted F_0 values from instances 70 msec before are used to generate the excitation signals at the next frame. Consequently, the EL speech produced by the proposed system always suffers from misalignment between the articulated sounds and F_0 patterns caused by this delay. It is necessary to investigate whether or not this misalignment causes perceivable degradation in the EL speech.

The other problem is acoustic mismatches between the training and prediction processes. The EL speech produced by using the proposed system is affected by the predicted F_0 values. Therefore, spectral parameters extracted from it are also affected by them. This has the potential to cause acoustic mismatches between the training and prediction processes. In statistical excitation prediction, spectral analysis based on fast Fourier transform (FFT) is usually used to significantly reduce computational cost. Because the FFT-based spectral analysis easily captures periodicity of the excitation signals, the source features (i.e., the mel-cepstral segment features) are strongly affected by the predicted F_0 values. To address this issue, we investigate two approaches, a model-based approach and a feature extraction approach. The former approach uses the conversion model widely accepting EL speech with various F_0 values. For the original EL speech samples in the parallel data set, analysis-resynthesized EL speech samples are generated by modifying the F_0 values. FFT spectral features are extracted from these generated samples and the resulting source features are used in the GMM training. In this paper, global linear transformation is used for modifying the F_0. On the other hand, the latter approach uses a spectral analysis method robust to periodicity of the excitation signals. STRAIGHT analysis [17] is used in this paper. To significantly reduce computational cost of STRAIGHT analysis, the predicted F_0 value is directly used in spectral analysis to avoid the F_0 extraction process.

3.2. A Simulation Experient

As the first step for implementation of the proposed system, we investigate the performance of the proposed system by a simulation experiment in this paper. The simulated implementation of the proposed system is also shown in Figure 3. In the prediction process, not the low-delay conversion algorithm but the batch conversion algorithm is employed. The conversion accuracy of the two algorithms is almost equivalent.

At first, 1) we extract spectral envelope parameters and aperiodic components [18] from the original EL speech in advance by using STRAIGHT analysis. These features capture acoustic properties depending on articulation and the excitation signals leaked out from the electrolarynx, except for periodicity of the excitation signals. These are used to approximate the EL speech production process. Then, 2) spectral segment features are extracted from EL speech, and F_0 patterns of normal speech are predicted from them based on the statistical excitation prediction. 3) The predicted F_0 patterns are delayed to consider the delay time caused by real time prediction process. 4) Using the delayed F_0 patterns and the extracted aperiodic components, excitation signals are generated using the mixed

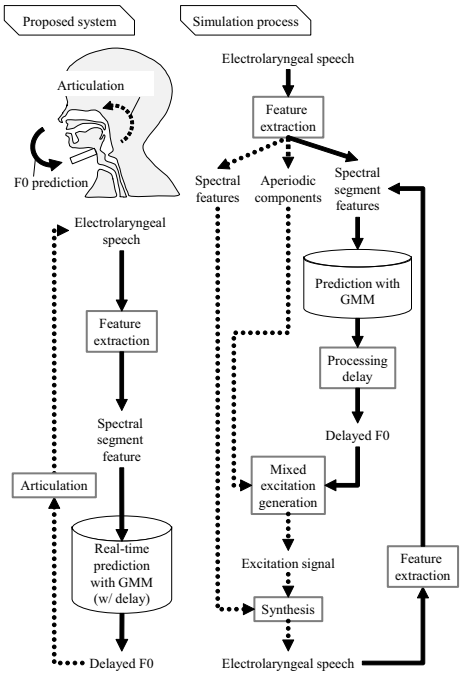

Figure 3: The proposed system and its simulation implementation.

excitation model [19]. 5) Finally, the enhanced EL speech is approximately synthesized by filtering the generated excitation signals with the extracted spectral envelope parameters. Note that this is a result of using the spectral segment features extracted from the original EL speech, and therefore it is not affected by the predicted F_0 patterns. To consider the impact of the predicted F_0 patterns on the spectral segment features, 6) the spectral segment features are extracted again from the synthesized EL speech and F_0 pattern prediction is also performed again using the extracted spectral segment features. Step 3) to step 6) are iteratively repeated until the predicted F_0 patterns converge. If they converge, the proposed system may be expected to work stably because the EL speech produced with the predicted F_0 patterns is consistent with that used in the spectral segment feature extraction. We experimentally investigate whether or not the predicted F_0 patterns converge.

4. Experimental Evaluation

4.1. Experimental Conditions

We conducted an objective test for evaluating prediction accuracy of F_0 patterns and two subjective evaluations on intelligibility and naturalness. The source speaker was one laryngectomee and the target speaker was one non-disabled speaker. Both speakers recorded 50 phoneme-balanced sentences. Sampling frequency was set to 16 kHz.

We employed FFT analysis or STRAIGHT analysis to extract the spectrum parameters of EL speech. Note that F_0 values of EL speech in STRAIGHT analysis were constantly set to 100 Hz instead of performing STRAIGHT F_0 analysis because F_0 of the excitation signals was almost equivalent to 100 Hz in the electrolarynx used by the laryngectomee. The frame shift length was set to 5 msec. The extracted spectral parameters were converted to the 0th through 24th mel-cepstral coefficients, which were used to extract the mel-cepstral segment feature as the source feature. The mel-cepstra at current $\pm$ 4 frames were used in this segment feature extraction. F_0 values

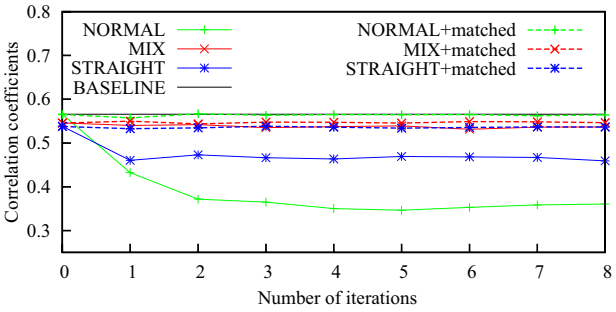

Figure 4: Prediction accuracy for F_0 correlation coefficient.

of normal speech were extracted with STRAIGHT F_0 analysis and continuous F_0 patterns were generated using a low-pass filter with 10 Hz cut-off frequency as the target feature. The mean F_0 value of normal speech was around 220 Hz.

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. The number of mixture components was set to 32. In the training data generation process described in Section 3.1, F_0 values were shifted to 150, 200, and 250 Hz, and totally 160 EL speech samples were used to train the GMM. The delay time in the simulation experiment was set to 70 msec.

The EL speech generated by the following four systems were mainly evaluated:

EL Original EL speech

BASELINE Enhanced speech that does not perform real-time F_0 predictioin, and that has no processing delay causing the F_0 misalignment. This is equivalent to the conventional hybrid EL speech enhancement method [4] without the noise reduction process.

MIX Enhanced speech with the processing delay using the GMM trained with the training data generation process.

STRAIGHT Enhanced speech with the processing delay using robust spectral analysis with STRAIGHT using the predicted F_0.

In the objective evaluation, the correlation coefficient between the predicted and natural F_0 patterns was calculated. To clarify the impact of the acoustic mismatches caused by the predicted F_0 on the F_0 estimation accuracy, we also evaluated a system "NORMAL" with the processing delay using the GMM without the training data generation process nor the robust spectral analysis. Moreover, the F_0 estimation accuracy not suffering from the acoustic mismatches was also evaluated in the systems, "MIX", "STRAIGHT", and "NORMAL" by shifting the predicted F_0 values so that their mean value was equal to that of the original EL speech used in the training (*i.e.*, 100 Hz), which were denoted as "MIX+matched", "STRAIGHT+matched", and "NORMAL+matched."

In the subjective evaluation, we conducted two opinion tests on intelligibility and naturalness. The opinion score was set to a 5-point scale (i.e., 1 (very poor) to 5 (excellent)). The number of listeners was 5 in each test. Each listener evaluated naturalness and intelligibility of "EL", "BASELINE", "MIX", and "STRAIGHT."

4.2. Experimental Results

Figure 4 shows the result of the objective evaluation. We can see that correlation coefficients of all systems converge and the simulation process works reasonably well. If the acoustic mismatches are not caused by the predicted F_0, *i.e.*, in the systems

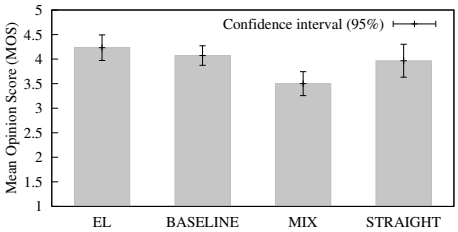

Figure 5: Result of opinion test on intelligibility.

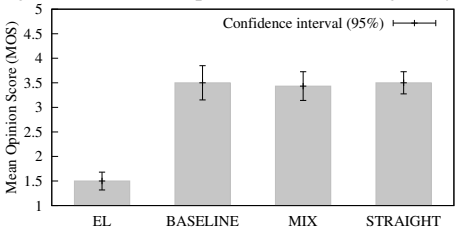

Figure 6: Result of opinion test on naturalness.

"+matched", the correlation coefficient is constant over the iterative process in the simulation. On the other hand, it can be observed from "NORMAL" that the correlation coefficient significantly degrades in the mismatched situations. This degradation is effectively alleviated by using the training data generation "MIX" or the robust spectral analysis "STRAIGHT."

Figure 5 shows the result of the opinion test on intelligibility. "BASELINE" causes no degradation in intelligibility compared to the original EL speech as reported in [4]. In the proposed systems, "STRAIGHT" can also preserve high intelligibility of the original EL speech but "MIX" causes slight degradation in intelligibility. We found that F_0 patterns generated in "MIX" sometimes varied unstably. Although we need to more carefully analyze these results, it is possible that the number of mixture components in "MIX" needs to be increased to accept more varieties of the mel-cepstral segment features.

Figure 6 shows the result of the opinion test on naturalness. The original EL speech is very unnatural but its naturalness can be significantly improved by "BASELINE" as reported in [4]. The proposed systems "MIX" and "STRAIGHT" can also significantly improve the naturalness. Because no statistically significant difference can be observed between "BASELINE" and the proposed systems "MIX" and "STRAIGHT", it is revealed that misalignment of F_0 patterns does not cause any degradation in naturalness.

5. Conclusions

In this paper, we proposed an electrolaryngeal (EL) speech enhancement system that directly controls F_0 values of the excitation signals generated by an electrolarynx based on statistical excitation prediction. We conducted simulation experiments to evaluate the effectiveness of the proposed system, investigating whether or not the enhanced EL speech is significantly affected by the processing delay of F_0 prediction and acoustic mismatches caused by the dynamically predicted F_0 values, which are always observed in the proposed system. The experimental results have shown that they cause no significant differences in either naturalness or intelligibility and the proposed system can significantly improve naturalness of EL speech while preserving its high intelligibility.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26280060.

7. References

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

[2] H. Sharifzadeh, I. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 10, pp. 2448–2458, October 2010.

[3] G. Aguilar-Torres, M. Nakano-Miyatake, and H. Perez-Meana, "Enhancement and restoration of alaryngeal speech signals," in *Electronics, Communications and Computers, 2006. CONIELECOMP 2006. 16th International Conference on*, February 2006, pp. 31–31.

[4] 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." in *Proc. INTERSPEECH*, August 2013, pp. 3067–3071.

[5] S. Basha and P. Pandey, "Real-time enhancement of electrolaryngeal speech by spectral subtraction," in *Communications (NCC), 2012 National Conference on*, February 2012, pp. 1–5.

[6] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," in *Proc. Speech Communication*, vol. 54, no. 1, January 2012, pp. 134 – 146.

[7] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 172–183, January 2014.

[8] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, March 1998.

[9] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, November 2007.

[10] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, November 2012.

[11] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, July 2011.

[12] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "An evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement," in *Proc. ICASSP*, May 2014, pp. 4521–4525.

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

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, June 2000, pp. 1315–1318.

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

[16] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, September 2008, pp. 1076–1079.

[17] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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," in *Proc. Speech Communication*, vol. 27, no. 3. Elsevier, April 1999, pp. 187–207.

[18] 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 synthesis system straight," in *Proc. MAVEBA*, September 2001, pp. 13–15.

[19] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, September 2006, pp. 2266–2269.