AN EVALUATION OF ALARYNGEAL SPEECH ENHANCEMENT METHODS BASED ON VOICE CONVERSION TECHNIQUES

Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Nara, 630-0192 Japan

E-mail: {hironori-d, kei-naka, tomoki, sawatari, shikano}@is.naist.jp

ABSTRACT

In this study, we evaluate our proposed methods for enhancing alaryngeal speech based on statistical voice conversion techniques. Voice conversion based on a Gaussian mixture model has been applied to the conversion of alaryngeal speech into normal speech (AL-to-Speech). Moreover, one-to-many eigenvoice conversion (EVC) has also been applied to AL-to-Speech to enable the recovery of the original voice quality of laryngectomees even if only one arbitrary utterance of the original voice is available. VC/EVC-based AL-to-Speech systems have been developed for several types of alaryngeal speech, such as esophageal speech (ES), electrolaryngeal speech (EL), and body-conducted silent electrolaryngeal speech (silent EL). These proposed systems are compared with each other from various perspectives. The experimental results demonstrate that our proposed systems yield significant enhancement effects on each type of alaryngeal speech.

Index Terms**—** alaryngeal speech, speech enhancement, voice conversion, eigenvoice conversion, performance evaluations

1. INTRODUCTION

Laryngectomees who have undergone total laryngectomy due to an accident or laryngeal cancer cannot produce speech sounds in a conventional manner because their vocal folds have been removed. Therefore, they require an alternative speaking method to produce speech sounds using sound sources generated in a special manner without vibrating their vocal folds. The produced speech is called alaryngeal speech.

There are various methods of producing alaryngeal speech. In this study, we focus on the three types of alaryngeal speech shown in Fig. 1: esophageal speech (ES), electrolaryngeal speech (EL), and body-conducted silent electrolaryngeal speech (silent EL). ES and EL are the most popular types of alaryngeal speech in Japan. In ES, alternative excitation sounds are generated by releasing gases from or through the esophagus. Thus, ES can be produced without any equipment. However, it is difficult to learn the skills to produce ES. On the other hand, EL is produced using an electrolarynx, a medical device for mechanically generating the sound source signals that are conducted into the oral cavity from the skin on the lower jaw. It is much easier to learn how to speak using the electrolarynx than to learn how to produce ES. However, because the electrolarynx needs to generate sufficient loud-sound source signals to make the produced speech sufficiently audible, sound source signals are easily emitted outside, disturbing speech communication. To resolve this issue, a speaking method for silent EL has been proposed [1]. A new sound source unit is used to generate less audible sound source signals. Since the produced speech also becomes less audible, it is detected with a non audible murmur (NAM) microphone, which is a body-conductive microphone capable of detecting extremely soft speech from the neck below the ear. These three types of alaryngeal speech allow laryngectomees to produce speech sounds again. However, their sound quality and intelligibility are severely degraded compared with those of normal speech uttered by non-laryngectomees. Moreover, alaryngeal speech sounds are of similar quality regardless of the speaker. This problem is called degradation of speaker individuality.

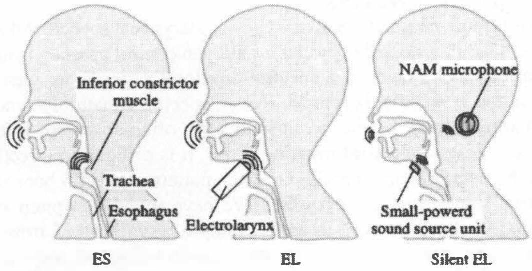

Fig. 1. Methods of producing three types of alaryngeal speech (ES, EL, and silent EL).

As one of the techniques for effectively changing voice quality while keeping linguistic contents unchanged, statistical voice conversion (VC) has been studied for around two decades [2, 3, 4]. In particular, speaker conversion based on a Gaussian mixture model (GMM) [3] has been widely studied and its performance has been significantly improved [4]. A GMM of the joint probability density of acoustic features between a source speaker's voice and a target speaker's voice is trained in advance using parallel data consisting of dozens of utterance pairs of the source and target speakers. The trained GMM is capable of converting the acoustic features of the source speech to those of the target speech in a probabilistic manner. Moreover, to make the training process more flexible, one-to-many eigenvoice conversion (EVC) [5] has been proposed as a method of converting a single speaker's voice into an arbitrary speaker's voice. This method enables us to control the speaker individuality of the converted speech by manipulating a small number of parameters or by automatically adjusting them to an arbitrary target speaker using only a few target speech samples as adaptation data.

To effectively enhance alaryngeal speech, we have proposed an alaryngeal speech enhancement method that converts alaryngeal speech into normal speech using VC techniques. The proposed method is called alaryngeal speech-to-speech (AL-to-Speech) [6, 7]. AL-to-Speech yields significant improvements in speech quality since the converted speech is basically generated according to the statistical properties of the acoustic features of normal speech. Moreover, to enable flexible control of the converted voice quality, one-to-many EVC has also been applied to the conversion of AL-to-Speech [6]. EVC-based AL-to-Speech allows laryngectomees to recover their original natural voice quality even if only one arbitrary utterance of their natural speech is available. Using these techniques, we have preliminarily developed AL-to-Speech systems for ES, EL, and silent EL [8].

In this paper, we describe our proposed AL-to-Speech systems

This work was supported in part by MIC SCOPE. The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan for permission to use the STRAIGHT analysis-synthesis method.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

based on VC/EVC for individual types of alaryngeal speech and evaluate their effectiveness. A comparison between VC and EVC in AL-to-Speech and a comparison among individual AL-to-Speech systems are conducted from various perspectives. The enhancement effects of the individual systems are demonstrated from the experimental results.

2. ALARYNGEAL SPEECH

2.1. Esophageal speech (ES)

ES sounds more natural than other types of alaryngeal speech. Additionally, a speaker skilled in producing ES can control prosody using residual organs. However, a spectral envelope varies more unstably than that in normal speech. Moreover, specific unnatural sounds caused at producing the excitation sounds are often observed. Even if we can perceive pitch information in ES, it is difficult to directly extract F_0 patterns corresponding to pitch patterns from ES because excitation signals are less periodic. We have found that pitch information is also perceived in an ES sample resynthesized from a mel-cepstrum sequence including power coefficients and noise excitation. Therefore, it is expected that pitch information of ES is included in a spectral envelope.

2.2. Electrolaryngeal speech (EL)

EL sounds mechanical owing to artificial excitation signals. Although a spectral envelope stably varies according to each phoneme, it is distorted by the sound source signals leaked from the electrolarynx. The electrolarynx used in this paper generates sound source signals with almost constant F_0 values and a high periodicity. Excitation parameters such as F_0 and aperiodic components are easily extracted from EL but are less informative since they capture only the acoustic characteristics of the artificial excitation signals.

2.3. Body-conducted silent electrolaryngeal speech (silent EL)

Silent EL sounds much more unnatural than EL owing to its lower-powered sound source signals and body conduction. It basically has similar acoustic characteristics to EL except that 1) the signal-to-noise ratio of silent EL is much lower than that of EL and 2) a severe attenuation of high-frequency components over 3 or 4 kHz is induced by the lack of radiation characteristics from the lips and by the effect of the low-pass characteristics of the soft tissue.

3. AL-TO-SPEECH

In AL-to-Speech, the spectrum, aperiodic components, and F_0 of the target normal speech are independently estimated using GMMs or eigenvoice GMMs (EV-GMMs) from the spectrum of alaryngeal speech. To eliminate unstable fluctuations observed in a spectrum sequence of alaryngeal speech and to compensate for a spectral structure collapsed by the production mechanisms of individual alaryngeal speech, a segment feature vector extracted from multiple frames around a current analyzed frame is used as the input feature. We use three joint static and dynamic feature vectors of the spectrum, aperiodic components, and F_0 extracted from target normal speech as the output features.

In this section, we describe AL-to-Speech based on one-to-many EVC. This method entails a training, adaptation, and conversion process. The details of AL-to-Speech based on VC are shown in ref. [6].

3.1. Training process

As conversion models for estimating spectrum and aperiodic components of normal speech from the spectral segment of alaryngeal speech, EV-GMMs are trained using multiple parallel data sets consisting of utterance pairs of a laryngectomee and many prestored target speakers. Let us assume a source spectral segment feature vector, X_t, and a target joint static and dynamic feature vector, Y_t, at frame t. The EV-GMM models the joint probability density of the source and target feature vectors as;

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

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

where $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ is the target-speaker-dependent weight vector for controlling target voice quality. $\top$ denotes the transposition of the vector. $\boldsymbol{\lambda}^{(EV)}$ is a canonical EV-GMM parameter set consisting of the weight α_m, the mean vector $\boldsymbol{\mu}_m^{(X)}$, the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$, the bias vector $\boldsymbol{b}_m$, and the eigenvectors $\boldsymbol{A}_m = [\boldsymbol{a}_m(1), \cdots, \boldsymbol{a}_m(J)]$ for the m^{th} mixture component, where the number of eigenvectors is J.

The EV-GMM is adapted to a new target speaker by adjusting the weight vector so that the marginal likelihood for given target speech features is maximized [5]. This adaptation process is effective if speaker-dependent characteristics are well captured by short-term features, such as spectrum and aperiodic components. On the other hand, it is essentially difficult to control speaker-dependent characteristics captured by long-term features, such as F_0 patterns. Therefore, instead of the EV-GMM, a well-trained speaker-dependent GMM is used to estimate F_0 patterns from the spectral segment sequence of alaryngeal speech. In AL-to-Speech for ES, to develop the GMM for estimating F_0 patterns corresponding to the perceived pitch information of ES, we use F_0 values extracted from normal speech uttered by a non-laryngectomee as an imitating prosody of ES in the training as the output features [6]. To develop a GMM for F_0 estimation in EL and silent EL, speaker-dependent GMMs are separately trained for all prestored target speakers. Then, the GMM achieving the highest F_0 estimation accuracy is manually selected.

3.2. Adaptation and conversion processes

Assuming that a few speech samples uttered by laryngectomees before undergoing total laryngectomy are available as adaptation data, the EV-GMM is flexibly adapted to the target voice quality by automatically determining the weight vector in a text-independent manner [5]. The weight vectors of the EV-GMMs for the spectral and aperiodic estimations are independently estimated using the spectral features and the aperiodic components extracted from the given target speech samples. The converted spectral feature vectors and aperiodic components are independently estimated using the adapted EV-GMMs. On the other hand, in the F_0 estimation, the global speaker-dependent characteristics of F_0 patterns are simply controlled. A log-scaled F_0 sequence is first estimated with the selected speaker-dependent GMM, and then further converted so that its mean μ_x and standard deviation σ_x are equal to those of the adaptation speech data, μ_y and σ_y, as follows:

$$\log y_t = \frac{\sigma_y}{\sigma_x}(\log x_t - \mu_x) + \mu_y, \tag{3}$$

where x_t and y_t denote the F_0 value estimated with the GMM and the converted F_0 value at frame t, respectively. The maximum likelihood estimation method considering not only the explicit relationship between static and dynamic features, but also global variance [4], is adopted in the estimation of the converted features.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We recorded 50 phoneme-balanced sentences of ES uttered by one Japanese male laryngectomee, those of EL and silent EL uttered by another Japanese male laryngectomee, and those of normal speech uttered by each of 40 Japanese non-laryngectomees. The speech data of 30 non-laryngectomees were used for training and those of the other 10 non-laryngectomees were used as the target data for evaluation. From the 50 recorded sentences of each speaker, 40 were used as the training or adaptation data and the remaining 10 were used as the test data. The sampling frequency was set to 16 kHz.

The 0th through 24th mel-cepstral coefficients were used as spectral parameters. Mel-cepstrum analysis [9] was employed for alaryngeal speech and STRAIGHT analysis [10] was employed for normal speech. The frame shift was 5 ms. To extract the spectral segment feature of ES, current and ±8 frames were used for spectral and aperiodic estimation and current and ±16 frames were used for F_0 estimation. For EL and silent EL, current and ±8 frames were used for each parameter estimation. These frame lengths were preliminarily optimized. As the source excitation features of normal speech, we used log-scaled F_0 values and aperiodic components for designing mixed excitation.

The EV-GMMs for spectral and aperiodic component estimation were trained for each type of alaryngeal speech. The numbers of eigenvectors and mixture components were set to 29 and 64 in every EV-GMM, respectively. The EV-GMMs were adapted to the target speakers using 1, 2, 4, 8, 16, or 32 utterances of their normal speech data. For AL-to-Speech based on VC, the GMMs for spectral and aperiodic estimation were trained using a parallel dataset for each type of alaryngeal speech and normal speech of each target speaker. The number of training utterance pairs was set to 1, 2, 4, 8, 16 or 32. The number of mixture components was optimized manually depending on the training data size. Individual speaker-dependent GMMs for F_0 estimation were trained for all the 40 non-laryngectomees. The GMM yielding the most natural F_0 pattern was then selected by listening to the converted speech. The same F_0 estimation process was performed for the EVC-based AL-to-Speech and VC-based AL-to-Speech.

4.2. Objective evaluation

We evaluated the effectiveness of AL-to-Speech based on EVC/VC for each type of alaryngeal speech with root mean square error (RMSE) on aperiodic components. The result of mel-cepstral distortion is shown in ref. [8]. Figure 2 shows RMSE on aperiodic components as a function of the number of adaptation utterances used in EVC or of utterance pairs used in VC. EVC shows a significantly smaller RMSE than VC in each type of alaryngeal speech enhancement when the amount of the target normal speech data is small. Even if only one arbitrary utterance of the target normal speech is available in EVC, its conversion performance is almost equivalent to or better than that of VC using 16 parallel utterance pairs. It is also observed that ES yields the best conversion accuracy and silent EL yields the worst among the three types of alaryngeal speech. Note that similar results have been observed in mel-cepstral distortion [8].

We also evaluated the F_0 estimation accuracy in AL-to-Speech for each type of alaryngeal speech using F_0 correlation coefficient and Unvoiced/Voiced (U/V) error between converted speech and target normal speech. To demonstrate the F_0 estimation accuracy for various speakers in AL-to-Speech, the results calculated using individual speaker-dependent GMMs for the 40 non-laryngectomees are shown in Table 1. For ES, the results for another non-laryngectomee who uttered normal speech so that its pitch sounded similar to that of ES are also shown as "ES pitch." ES yields the best estimation accuracy among the three types of alaryngeal speech. Additionally, the estimation accuracy is significantly improved using the GMM developed with the normal speech, the F_0 patterns of which correspond well to the pitch patterns of ES.

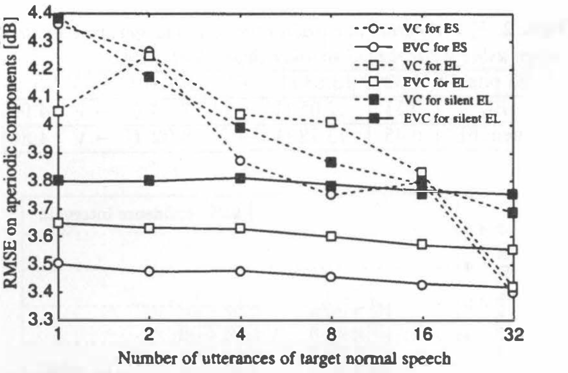

Fig. 2. RMSE on aperiodic components as a function of the number of utterances of target normal speech (i.e., utterance pairs in VC or adaptation utterances in EVC).

Table 1. *F_0 estimation accuracies for various target speakers using corresponding target-speaker-dependent GMMs*

	Correlation	U/V error [%]
ES	0.58	12.39 ($V \rightarrow U : 6.59, U \rightarrow V : 5.80$)
EL	0.40	13.20 ($V \rightarrow U : 4.92, U \rightarrow V : 8.28$)
Silent EL	0.42	14.02 ($V \rightarrow U : 6.89, U \rightarrow V : 7.13$)
ES pitch	0.68	8.36 ($V \rightarrow U : 4.30, U \rightarrow V : 4.05$)

The final results for the 10 target non-laryngectomees from the test data are shown in Table 2. The GMM for "ES pitch" was used in ES enhancement, and manually selected speaker-dependent GMMs were used in the ES/silent EL enhancement. Namely, the speaker used in the model training is different from the target speakers. It is observed that, for EL and silent EL, the estimation accuracy of the selected GMMs is higher than that of various speaker-dependent GMMs shown in Table 1, even though a speaker different from the target speakers is used in the training. To generate a natural F_0 pattern in AL-to-Speech, it is useful to select an optimum speaker for training rather than to directly use the same speaker as the actual target speaker since the F_0 estimation accuracy largely varies among different speakers. It is also observed that ES enhancement yields better F_0 correlation than the others.

4.3. Subjective evaluation

We conducted opinion tests of speech quality and intelligibility. In the opinion test of intelligibility, 8 listeners evaluated 9 types of speech including original alaryngeal speech and converted speech with AL-to-Speech based on VC/EVC in ES, EL, and silent EL. The VC-based AL-to-Speech used 32 utterance pairs for GMM training. On the other hand, only one utterance was used as adaptation data for the EVC-based AL-to-Speech. Each listener evaluated 135 speech samples. The experimental conditions for the opinion test of speech quality are shown in ref. [8]. We also conducted a preference test to evaluate speaker individuality. In the preference test, 6 listeners evaluated 6 types of speech consisting of converted speech by the VC/EVC-based AL-to-Speech in EL, ES, and silent EL. The training data used in VC and the adaptation data used in EVC were the same as those used in the opinion tests.

Figures 3 and 4 show the results of the opinion tests of speech quality and intelligibility, respectively. All the AL-to-Speech methods yield significant improvements in speech quality compared with that of the original alaryngeal speech. The speech quality of the enhanced silent EL is lower than that of the enhanced ES and enhanced EL but it is significantly higher than that of each type of original ala-

Table 2. *F_0 estimation accuracies for actual target speakers in evaluation using well-trained speaker-dependent GMMs*

ES pitch	0.62	13.88 ($V \rightarrow U$: 10.70, $U \rightarrow V$: 3.18)
EL	0.51	12.05 ($V \rightarrow U$: 7.13, $U \rightarrow V$: 4.92)
Silent EL	0.45	13.78 ($V \rightarrow U$: 8.92, $U \rightarrow V$: 4.86)

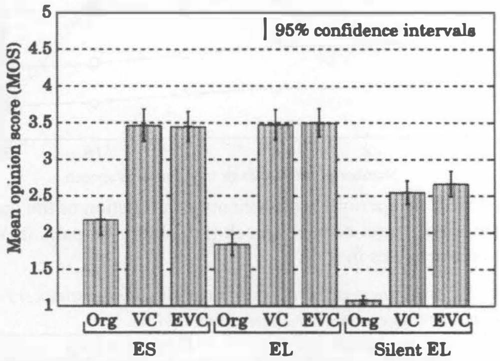

Fig. 3. Result of opinion test of speech quality. "Org", "VC", and "EVC" show original alaryngeal speech, converted speech by AL-to-Speech based on VC trained with 32 utterance pairs, and converted speech by AL-to-Speech based on one-to-many EVC adapted with one utterance of target speech, respectively.

ryngeal speech. The intelligibilities of ES and silent EL are also improved by AL-to-Speech. On the other hand, the intelligibility of EL slightly degrades from that of the original EL by AL-to-Speech, as observed in our previous work [7]. The speech quality and intelligibility enhanced by the EVC-based AL-to-Speech are almost equivalent to those enhanced by the VC-based AL-to-Speech. Note that the EVC-based method requires only one arbitrary utterance of the target normal speech whereas the VC-based method requires 32 utterance pairs of alaryngeal speech and the target normal speech.

Figure 5 shows the result of the preference test. We can observe the same tendency as that in Fig. 2. Enhanced ES yields the best speaker individuality and enhanced silent EL yields the worst among the three types of alaryngeal speech. We can observe that the VC-based methods slightly outperform the EVC-based methods in ES and EL. This tendency depends on the amount of available parallel data used for GMM training in VC-based methods, as shown in Fig. 2.

5. CONCLUSIONS

In this paper, we evaluated our proposed statistical enhancement methods based on voice conversion techniques (AL-to-Speech) for three types of alaryngeal speech: esophageal speech (ES), electrolaryngeal speech (EL), and body-conducted silent electrolaryngeal speech (silent EL). The experimental results suggested that 1) the proposed methods significantly improve the speech quality of each type of alaryngeal speech, 2) the proposed methods also improve the intelligibilities of ES and silent EL, 3) AL-to-Speech based on eigenvoice conversion (EVC) is capable of effectively adjusting the voice quality of enhanced speech to the target voice quality using only one arbitrary utterance of the target voice, and 4) AL-to-Speech for ES is the best in terms of speech quality, intelligibility, and speaker individuality.

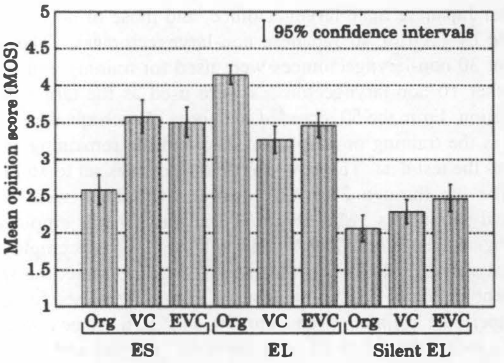

Fig. 4. Result of opinion test of intelligibility

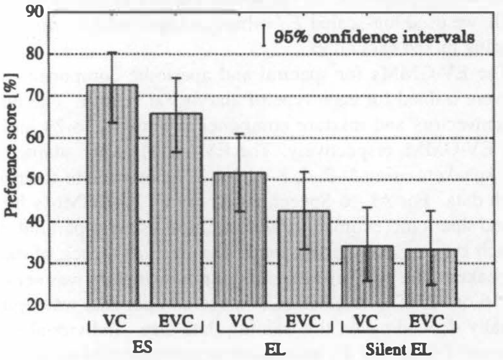

Fig. 5. Result of preference test of speaker individuality

6. REFERENCES

[1] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion," *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 7, pp. 1909–1917, July 2010.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, 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, "Eigenvoice conversion based on gaussian mixture model," *Interspeech*, pp. 2446–2449, Sept. 2006.

[6] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 9, pp. 2472–2482, Sept. 2010.

[7] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "The use of air-pressure sensor in electrolaryngeal speech enhancement based on statistical voice conversion," *Interspeech*, pp. 1628–1631, Sept. 2010.

[8] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid system based on one-to-many eigenvoice conversion for total laryngectomees," *APSIPA ASC*, pp. 498–501, Dec. 2010.

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

[10] 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," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.