PAPER Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application

A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical Excitation Generation

Kou TANAKA[†a)], ***Nonmember***, **Tomoki TODA**[†b)], ***Member***, **Graham NEUBIG**[†c)], ***Nonmember***, **Sakriani SAKTI**[†d)], ***and*** **Satoshi NAKAMURA**[†e)], ***Members***

SUMMARY This paper presents an electrolaryngeal (EL) speech enhancement method capable of significantly improving naturalness of EL speech while causing no degradation in its intelligibility. An electrolarynx is an external 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 to EL speech as noise. To address these issues, there are mainly two conventional approached to EL speech enhancement through either noise reduction or statistical voice conversion (VC). The former approach 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 approach 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 approach using a noise reduction method for enhancing spectral parameters and statistical voice conversion method for predicting excitation parameters. Moreover, we further modify the prediction process of the excitation parameters to improve its prediction accuracy and reduce adverse effects caused by unvoiced/voiced prediction errors. The experimental results demonstrate the proposed method yields significant improvements in naturalness compared with EL speech while keeping intelligibility high enough.

key words: *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 is 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. 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 Fig. 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 very 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.

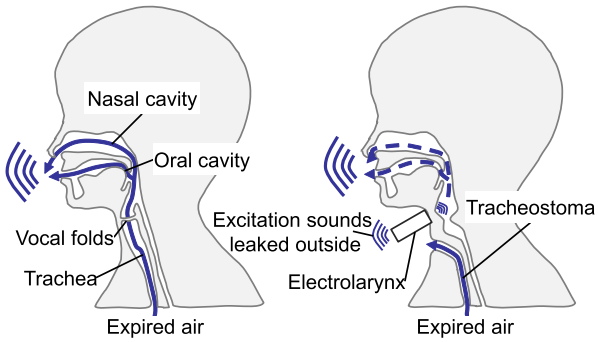

Fig. 1 Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

To address these issues of EL speech, two approaches have conventionally been adopted. 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 natural-

Manuscript received September 28, 2013.
Manuscript revised January 18, 2014.
[†]The authors are with Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630–0192 Japan.
a) E-mail: ko-t@is.naist.jp
b) E-mail: tomoki@is.naist.jp
c) E-mail: neubig@is.naist.jp
d) E-mail: ssakti@is.naist.jp
e) E-mail: s-nakamura@is.naist.jp
DOI: 10.1587/transinf.E97.D.1429

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers

ness 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]. Even if naturalness is significantly improved compared to EL speech, the enhanced speech suffering from degradation from another perspective will have less chance of being accepted by actual users. In particular, intelligible speech production is the most essential factor for laryngectomees to communicate with others. Therefore, it is required to develop an EL speech enhancement technique causing no degradation in intelligibility compared to the original EL speech.

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 improve prediction accuracy of the excitation parameters and reduce adverse effects caused by unvoiced/voiced prediction errors, we modify the prediction process of the excitation parameters. 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. Conventional EL Speech Enhancement

2.1 Enhancement Based on Spectral Subtraction (SS)

SS is a method for restoration of the amplitude spectrum of a speech signal that has been observed together with the leaked noise of an electrolarynx. This is done by subtracting 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. The enhanced speech signal is generated using the processed amplitude spectrum and the original phase extracted from the noisy speech signal.

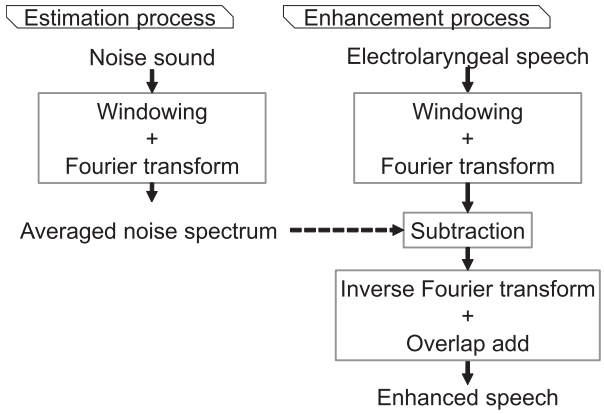

Fig. 2 EL speech enhancement based on SS.

In this paper, we implement SS for EL speech enhancement, as shown in Fig. 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 Fig. 1.

2.2 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 Fig. 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].

2.2.1 Training Process

The spectral components of EL speech are unstably and spectral structures of some phonemes are often collapsed due to the production mechanism of EL speech. To address these issues, we use spectral segment features ex-

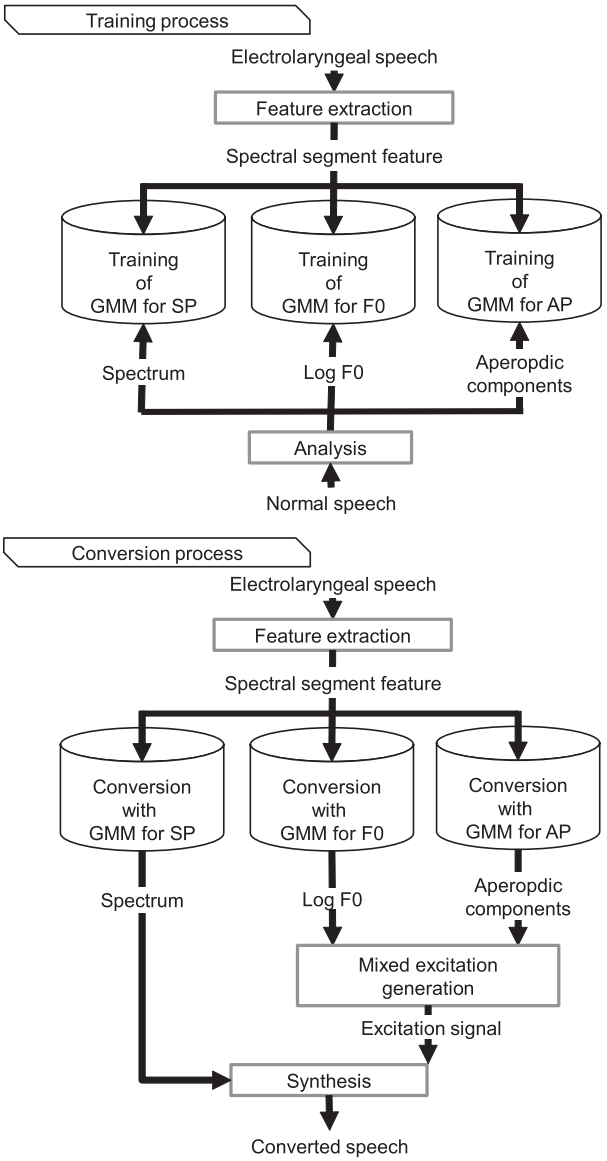

Fig. 3 EL speech enhancement based on VC.

tracted from multiple frames as follows:

$$\boldsymbol{X}_t = \boldsymbol{C}\boldsymbol{X'}_t + \boldsymbol{d} \tag{3}$$

where $\boldsymbol{X'}_t = [\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top]^\top$ is a joint vector generated by concatenating a spectral parameter vector $\boldsymbol{x}_t$ at the current frame and those at $\pm i$ preceding and succeeding frames. Because this joint vector includes redundant information, dimension reduction with principal component analysis (PCA) is performed for the joint vector $\boldsymbol{X'}_t$ in order to extract spectral segment features $\boldsymbol{X}_t$ at frame t, where $\boldsymbol{C}$ and $\boldsymbol{d}$ are a transformation matrix and a bias vector extracted by PCA, respectively. On the other hand, we assume 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 features 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:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \tag{4}$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{5}$$

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.

2.2.2 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:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})^\omega \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{6}$$

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 function 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, which is set to the ratio of the number of dimensions between vector $\boldsymbol{v}(\boldsymbol{y})$ and $\boldsymbol{Y}$, i.e., $1/(2T)$ in this paper. 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.

3. Proposed EL Speech Enhancement Based on a Hybrid Approach

3.1 Enhancement based on a Hybrid Approach

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 improve 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 Fig. 4. As laryngectomees have the capability to properly articulate the excitation signals, spectral parameters of EL speech do not have 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 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.

3.2 Improvement of Statistical Excitation Prediction

We propose several methods to improve statistical excitation prediction in the hybrid enhancement process.

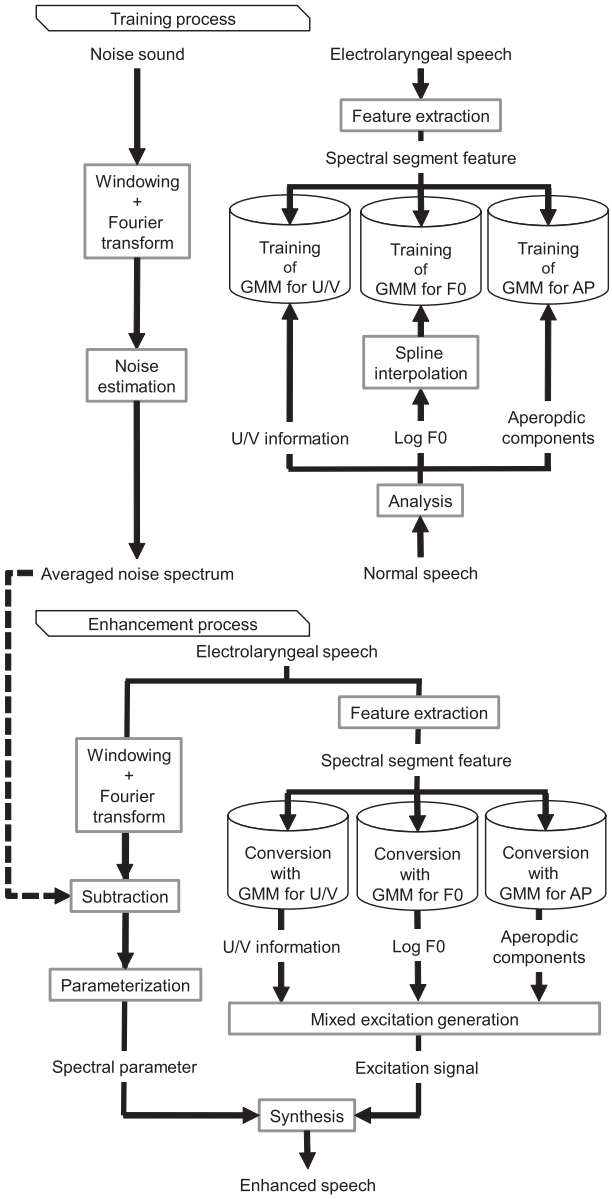

Fig. 4 EL speech enhancement based on the proposed hybrid approach.

3.2.1 Continuous F_0 Patterns (CF0)

The F_0 patterns extracted from natural voices are discontinuous because the F_0 values are not observed at the unvoiced and silent frames. In the conventional VC-based enhancement method, a constant value clearly different from F_0 values (e.g., a value much less than the minimum F_0 value) is used to represent F_0 values at those frames, and both unvoiced/voiced prediction and F_0 value estimation are performed with a single GMM in the same manner as described in [10]. However, it is not straightforward to accurately model such a discontinuous F_0 pattern.

To simplify characteristics of the parameter sequence to be modeled, we apply a continuous F_0 pattern to the statistical excitation prediction. In the training process, con-

tinuous F_0 patterns of normal speech are generated by using spline interpolation to produce F_0 values at unvoiced frames, as shown in the middle of Fig. 5. The resulting continuous F_0 patterns are used as the target parameter in the GMM training. The conventional discontinuous F_0 patterns are also modeled with another GMM to predict U/V information. In the conversion process, a continuous F_0 pattern and U/V information are separately predicted using the corresponding GMMs. A discontinuous F_0 pattern to be used in synthesis is finally generated by combining them. The effectiveness of using the continuous F_0 pattern has also been reported in the field of statistical parametric speech synthesis [14].

3.2.2 Remove 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, an 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 those patterns. To achieve this, we propose the use of a method to smooth the continuous F_0 patterns with low-pass filtering [15] as shown in the middle of Fig. 5. The smoothed continuous F_0 patterns are then modeled with the GMM.

3.2.3 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.

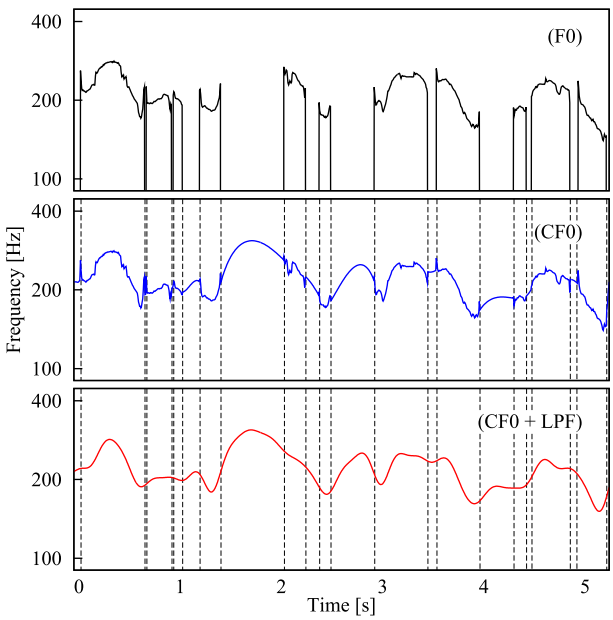

Fig. 5 Each type of F_0 patterns. Top figure is target F_0 counters, middle is continuous F_0 patterns using the spline interpolation, and bottom is smoothing continuous F_0 patterns using the low-pass filter.

4. Experimental Evaluations

4.1 Experimental Conditions

We objectively and subjectively compared the performance of the proposed system with the conventional systems. 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 [16] were used as the spectral parameters. The shift length was set to 5 ms. STRAIGHT analysis was also used for spectral extraction of EL speech but F_0 values were set to 100 Hz without F_0 extraction, which was equivalent to F_0 values of excitation signals generated by the electrolarynx used in the experiments. For the segment feature extraction, the current ± 4 frames were used to extract a 50-dimensional feature vector. PCA was conducted to determine the transformation matrix and the bias vector in Eq. (3) using all EL speech samples in the training data. 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. In the SS-based spectral enhancement method, the number of FFT points was set to 512 and individual parameters were set to $\alpha = 2.0$, $\beta = 0, 0$, and $\gamma = 1.0$, which were manually determined by listening to the enhanced speech so that its voice quality was improved as much as possible. The cut-off frequency of the low-pass filter was set to 10 Hz. Note that in the VC-based enhancement method, mel-cepstral and aperiodic distortion is shown in Table 1.

Table 1 Conversion accuracy in enhancement methods with VC.

Mel-cepstral distortion without power information	5.09 dB
Aperiodic distortion	3.19 dB

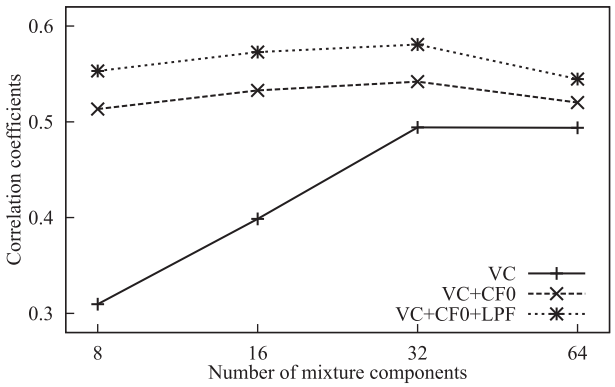

Fig. 6 Conversion accuracy for the F_0 patterns of each systems.

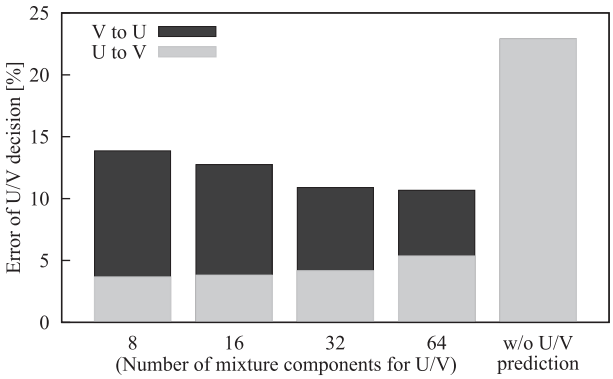

Fig. 7 U/V error rate of each systems.

4.2 Objective Evaluations

We evaluated the effectiveness of the proposed preprocessing for the training data, including the continuous F_0 estimation method and the low-pass filter. Note that the effectiveness of the continuous F_0 estimation method has already been reported in [14]. 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. As for the evaluation of the F_0 correlation coefficient, we set the number of GMM mixture components to 8, 16, 32, or 64. We evaluate three systems using the normal F_0 patterns extracted from target natural voices, the continuous F_0 patterns interpolated with the normal F_0 patterns using spline interpolation, and the smoothing continuous F_0 patterns extracted from continuous F_0 patterns through the low-pass filter. On the other hand, for the evaluation of U/V error rate, we also set the number of GMM mixture components for VC to 8, 16, 32, or 64.

Figure 6 shows the result of the evaluation for the F_0 correlation coefficient. In the case of using only the VC method, the F_0 correlation coefficient depends on the number of GMM mixture components, and is maximized with 64 mixture components. However, as for the proposed method, VC + CF0 and VC + CF0 + LPF, those are not depended well. Especially, in the small number of GMM mixture components, a significant degradation is not observed compared with using only VC methods. Moreover, it can be observed that the F_0 correlation coefficient is improved by the continuous F_0 estimation and also improved by using the low-pass filter. Note that in the case of the use of the continuous F_0 estimation method and the low-pass filter, the F_0 correlation coefficient is maximized with 32 mixture components.

Figure 7 shows the result of the evaluation for U/V error rate. We have found that large errors in the F_0 estimation tend to be observed at short voiced segments that are sometimes generated in only the VC-based enhancement method. This improvement is similar to that yielded by the continuous F_0 modeling in HMM-based speech synthesis [14]. As the number 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. The V-to-U errors still exist without 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 significantly increases owing to the continuous F_0 estimation method. Note that as we mentioned in Sect. 3.2.3, this increase causes no adverse effect compared with EL speech because EL speech is totally voiced speech.

4.3 Subjective Evaluations

We conducted two opinion tests of listenability and naturalness and a dictation test on intelligibility. In this paper, the term "listenability" is used to indicate a score that was measured by asking the listener to subjectively evaluate how easy it was to understand the utterance. The term "intelligibility" is used to indicate a score that was calculated by asking the listener to write down the content of the utterance, and measuring the accuracy of transcription. The term "naturalness" is used to indicate a score that was measured by asking the listener to subjectively evaluate whether the evaluated speech is similar to natural human speech or not. In the opinion tests, each listener evaluated each speech quality of the enhanced voices using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). We evaluated the following five types of speech samples:

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

On the other hand, in the dictation test, in order to demonstrate the effectiveness of avoiding U/V prediction errors, we evaluated the following five types of speech sam-

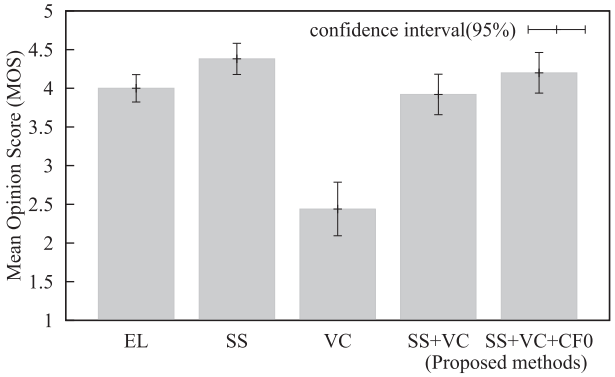

Fig. 8 Result of opinion test on listenability.

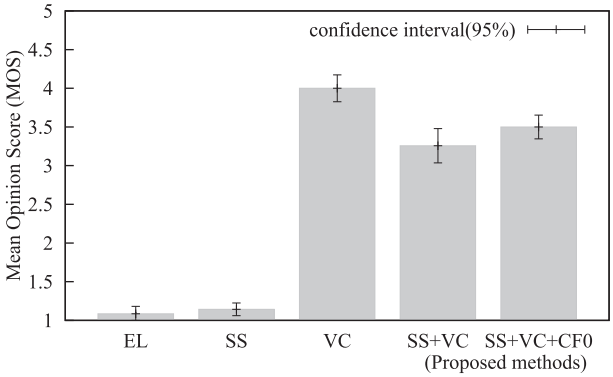

Fig. 9 Result of opinion test on naturalness.

ples:

EL original EL speech
SS speech enhanced by the SS-based enhancement method
Hybrid(V) speech enhanced by the proposed hybrid enhancement method with continuous F_0 estimation
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 is a method based on SS+VC+CF0+LPF. As the reference U/V information, we use 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. Intelligibility was evaluated using word correct rate and word accuracy, which were calculated as follows:

$$\texttt{word correct rate [\%]} = \frac{C}{S+D+C} \tag{7}$$

$$\texttt{word accuracy [\%]} = \frac{C-I}{S+D+C} \tag{8}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and C is the number of correct words. Note that the VC-based enhancement method generally causes a significant degradation in intelligibility (around 3% word recognition rate reduction) compared to EL speech as reported in [3]. All tests were performed by 5 listeners. Each listener evaluated 50 samples, 10 samples per system [†].

First, in Fig. 8, we show the results of the subjective opinion test on listenability. 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 a degradation compared with EL but it still causes a very small degradation compared with SS. This adverse effect on listenability is not observed in the proposed hybrid methods (SS+VC and SS+VC+CF0) thanks to no spectral conversion error.

[†]Some samples are available at the URL: http://isw3.naist.jp/~ko-t/eval/publication/IEICE/HAESE/index.html

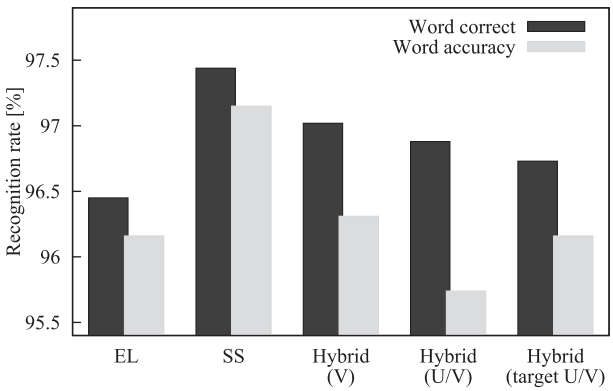

Fig. 10 Result of opinion test on naturalness.

Figure 9 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 continuous F_0 estimation is effective for improving naturalness as well.

Figure 10 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 that avoided 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 by using vocoder and using 24-dimensional mel-cepstral coefficients as spectral features.

These results suggest that the proposed hybrid approach to EL speech enhancement based on the continuous F_0 estimation and using the low-pass filter is effective in significantly improving naturalness of EL speech while avoiding degradation in listenability that is often observed in the conventional VC-based enhancement method.

5. Conclusion

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 improve the excitation features estimation, we implemented continuous F_0 estimation and low-pass filtering as part of the proposed approach. Moreover, we proposed a method for avoiding U/V prediction errors causing degradation in intelligibility. 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. Furthermore, U/V prediction is not always required for EL speech enhancement.

Acknowledgements

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

References

[1] H. Liu, Q. Zhao, M.X. Wan, and S.P. Wang, "Enhancement of electrolarynx speech based on auditory masking," IEEE Trans. Biomed. Eng., vol.53, no.5, pp.865–874, May 2006.
[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," SPECOM, vol.54, no.1, pp.134–146, Jan. 2012.
[3] D. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," IEEE Trans. Audio. Speech Language, vol.22, no.1, pp.172–183, Jan. 2014.
[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.27, no.2, pp.113–120, April 1979.
[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech Audio Process., vol.6, no.2, pp.131–142, March 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 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. Biomed. Eng., 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 Audio Process., vol.6, no.4, pp.328–337, July 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, Sept. 2001.
[10] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," IEEE Trans. Audio Speech 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, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. Interspeech, pp.2266–2269, Sept. 2006.
[14] K. Yu and S. Young, "Continuous F_0 modelling for HMM based statistical parametric speech synthesis," IEEE Trans. Audio Speech Language, vol.19, no.5, pp.1071–1079, July 2011.
[15] 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.
[16] 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, April 1999.

Kou Tanaka graduated from the Department of Mathematics and Informatics, Faculty of Human Development, Kobe University in Japan in 2012. He is currently in the master's course at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in Japan. He is a student member of ISCA, and ASJ.

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visting Researcher at the NICT, Kyoto, Japan, since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher at the ATR until March 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, speech analysis, speech production, and speech recognition. He received the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award from the TAF, the 2007 ISS Best Paper Award and the 2010 ISS Young Researcher's Award in Speech Field from the IEICE, the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award and the 26th Awaya Prize Young Researcher Award from the ASJ, the 2009 Young Author Best Paper Award from the IEEE SPS, the Best Paper Award (Short Paper in Regular Session Category) from APSIPA ASC 2012, the 2012 Kiyasu Special Industrial Achievement Award from the IPSJ, and the 2013 Best Paper Award (Speech Communication Journal) from EURASIP-ISCA. He was a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009. He is a member of IEEE, ISCA, IPSJ, and ASJ.

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign, U.S.A, in 2005, and his M.E. and Ph.D. in informatics from Kyoto University, Kyoto, Japan in 2010 and 2012 respectively. He is currently an assistant professor at the Nara Institute of Science an Technology, Nara, Japan. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

Sakriani Sakti received her B.E degree in Informatics (cumlaude) from Bandung Institute of Technology, Indonesia, in 1999. In2000, she received "DAAD-Siemens Program Asia 21st Century" Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR and U-STAR (2006–2011). She also served as a visiting professor of Computer Science Department, University of Indonesia (UI) in 2009–2011. Currently, she is an assistant professor of the Augmented Human Communication Lab, NAIST, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE amd IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was a director of ATR Spoken Language Communication Research Laboratories in 2000–2008, and a vice president of ATR in 2007–2008. He was a director general of Keihanna Research Laboratories, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently a professor and a director of Augmented Human Communication laboratory, Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of spoken dialog system, speech-to-speech translation. He is one of the leaders of speech-to-speech translation research projects including C-STAR, IWSLT and A-STAR. He headed the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received LREC Antonio Zampoli Award, the Commendation for Science and Technology by the Ministry of Science and Technology in Japan. He is an elected board member of ISCA, International Speech Communication Association, and an elected member of IEEE SPS, speech and language TC.