SICE Annual Conference 2013
September 14-17, 2013, Nagoya University, Nagoya, Japan

Statistical Voice Conversion Techniques for Alaryngeal Speech Enhancement

Tomoki Toda and Hironori Doi

Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
(Tel: +81-743-72-5261; E-mail: tomoki@is.naist.jp)

Abstract: This position paper gives a brief overview of our developed technologies for enhancing alaryngeal speech (AL speech) uttered by laryngectomees. There are several alternative speaking methods for laryngectomees to produce AL speech. However, any type of AL speech suffers from lack of naturalness and speaker individuality (identity). To address this issue, we have developed statistical voice conversion techniques for AL speech enhancement. Our developed techniques are capable of converting AL speech into normal speech in a probabilistic manner on the basis of statistics extracted from training data consisting of utterance pairs of AL speech and normal speech. Moreover, they are also capable of flexibly controlling voice quality of enhanced speech to effectively recover speaker individuality. We have developed three AL speech enhancement systems for 1) esophageal speech, 2) electrolaryngeal speech, and 3) body-conducted silent electrolaryngeal speech. Our experimental results have demonstrated that these systems yield significant improvements in naturalness and speaker individuality of each type of AL speech.

Keywords: laryngectomees, alaryngeal speech, speech enhancement, statistical approach, voice conversion.

1. INTRODUCTION

Patients who suffer from laryngeal cancer require total laryngectomy, which is a surgical operation to remove the larynx and tissues around the larynx such as the vocal folds. People who have undergone total laryngectomy, called laryngectomees, cannot speak in the usual manner owing to the removal of their vocal cords. Therefore, they need alternative speaking methods to produce speech sounds using residual organs or medical devices instead of vocal cords. Speech sounds generated by alternative speaking methods without vocal fold vibration are called alaryngeal speech (AL speech).

There are several alternative speaking methods, such as speaking methods for esophageal speech (ES speech), electrolaryngeal speech (EL speech), and body-conducted silent EL speech. ES speech is produced by modulating alternative excitation sounds that are generated by releasing gases from or through the esophagus by articulatory movement. EL speech is produced by articulating alternative excitation sounds generated from an electrolarynx, which is a medical device to mechanically generate the sound source signals. Body-conducted silent EL speech is produced by a new speaking method [1] using two devices, 1) a small sound source unit to generate less audible sound source signals while keeping them from emitting outside as noise and 2) a special body-conductive microphone, called nonaudible murmur (NAM) microphone [2], to detect extremely soft speech.

Although these types of AL speech allow laryngectomees to speak again, their sound quality, listenability, and intelligibility are severely degraded compared with those of normal speech. Moreover, AL speech sounds are of similar voice quality regardless of the speaker differences because the generation mechanism of the excitation sounds in each type of AL speech strongly affects the voice quality of the produced speech. Consequently, AL speech also suffers from the degradation of speaker individuality (identity).

To address these issues of AL speech, we have recently proposed a statistical approach to enhancement of AL speech, called AL-to-Speech [3], inspired by voice conversion (VC) techniques [4, 5], which have been mainly studied as speaker conversion techniques to convert the source speaker's voice into target speaker's voice while keeping linguistic information unchanged. This position paper gives a brief overview of our developed technologies for enhancing several types of AL speech by converting them into normal speech.

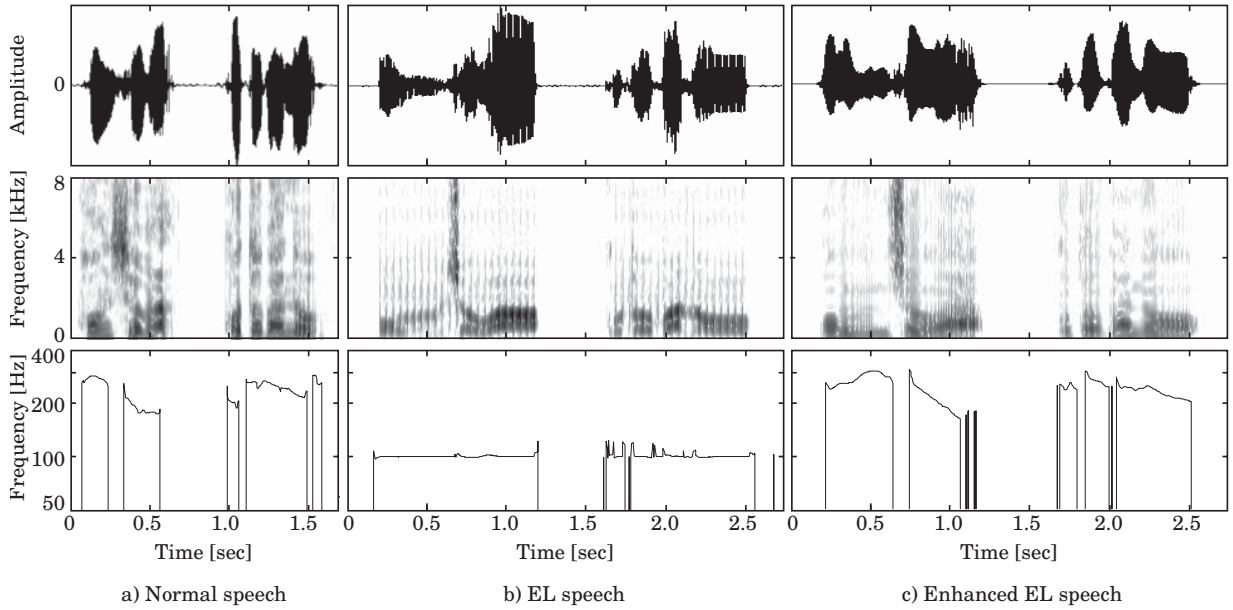
2. ALARYNGEAL SPEECH TO SPEECH (AL-TO-SPEECH)

The proposed approach consists of training and conversion processes. In the training process, a conversion function from acoustic features of AL speech into those of target normal speech is modeled using training data including utterance pairs of AL speech and normal speech. In the conversion process, any utterance of AL speech is converted to that of target normal speech on the basis of the conversion function without any text information. This data-driven approach is capable of complex acoustic modifications to compensate for the large acoustic differences between AL speech and normal speech.

As one of the state-of-the-art VC algorithms, a trajectory-wise conversion method using Gaussian mixture models (GMMs) [6] has been successfully applied to AL-to-Speech [7, 8]. Furthermore, to flexibly change the converted voice quality for helping laryngectomees to speak in their favorite voices or in their own voices that have already been lost but a few recorded samples are available, one-to-many eigenvoice conversion (EVC) [9], which is a technique for flexibly converting a specific source speaker's voice into an arbitrary target speaker's voice, has also been applied to AL-to-Speech [3].

We have developed AL-to-Speech systems for ES speech, EL speech and body-conducted silent EL speech. Note that the conversion process can be performed in real

PR0001/13/0000-1602 ¥400



a) Normal speech b) EL speech c) Enhanced EL speech

Fig. 1 Example of acoustic features, *i.e.*, waveforms, spectrograms, and fundamental frequency patterns, of a) normal speech, b) EL speech, and c) EL speech enhanced by AL-to-Speech based on EVC.

time by further implementing a real-time VC technique [10] for AL-to-Speech. Therefore, it is possible that laryngectomees use the developed AL-to-Speech systems in their conversations.

3. EFFECTIVENESS OF AL-TO-SPEECH

The effectiveness of the developed AL-to-Speech systems has been evaluated from various perspectives [3]. The experimental results have demonstrated that

1) every AL-to-Speech system significantly improves the speech quality of each type of AL speech,
2) the listenability of ES speech and body-conducted silent EL speech is significantly improved by AL-to-Speech,
3) the intelligibility of body-conducted silent EL speech is significantly improved by AL-to-Speech,
4) the AL-to-Speech systems based on EVC are capable of effectively adjusting the voice quality of the enhanced speech to the target voice quality using only one arbitrary utterance of the target voice.

Figure 1 shows an example of acoustic features of normal speech, EL speech, and EL speech enhanced by the AL-to-Speech system based on EVC. It can be observed that acoustic features of EL speech are very different from those of normal speech but these acoustic differences are effectively reduced by AL-to-Speech. Note that speech duration is not changed in AL-to-Speech as its conversion is essentially difficult in the real-time processing.

4. CONCLUSIONS

In this position paper, we have presented a brief overview of our developed technologies for enhancing alaryngeal speech uttered by laryngectomees. More details are in [3, 7, 8].

Acknowledgment: This work was supported in part by MEXT Grant-in-Aid for Young Scientists (A).

REFERENCES

[1] K. Nakamura *et al.*, "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, 2010.
[2] Y. Nakajima *et al.*, Non-Audible Murmur (NAM) Recognition. *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006.
[3] H. Doi *et al.*, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques", *Proc. ICASSP*, pp. 5136–5139, Prague, Czech Republic, May. 2011.
[4] M. Abe *et al.*, "Voice conversion through vector quantization", *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.
[5] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion", *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.
[6] T. Toda *et al.*, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory", *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[7] K. Nakamura *et al.*, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech", *Speech Communication*, Vol. 54, No. 1, pp. 134–146, 2012.
[8] H. Doi *et al.*, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models", *IEICE Trans. on Inf. and Syst.*, Vol. E93-D, No. 9, pp. 2472–2482, 2010.
[9] T. Toda *et al.*, "One-to-many and many-to-one voice conversion based on eigenvoices", *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.
[10] T. Toda *et al.*, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, Portland, USA, Sep. 2012.