IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 6, Dec-Jan, 2014
ISSN: 2320 - 8791
www.ijreat.org

Enhancement Of Esophageal Speech By Statistical Voice Conversion

Ada Christa[1], Arthi.A[2]

[1]**A.P, ECE Department, Anna University,Prathyusha Institute of Technology and Management, Trivallur, Tamilnadu,India**

[2]**ECE Department, Anna University,Prathyusha Institute of Technology and Management, Trivallur, Tamilnadu,India**

ABSTRACT: People who have undergone a total laryngectomy due to an accident or laryngeal cancer cannot produce speech sounds in a usual way because of the removal of vocal folds. Esophageal speech is one of the alternative speaking methods for laryngectomees. Although it does not require any external devices, generated voices sound unnatural. Esophageal speech is characterized by low naturalness and intelligibility. To improve the intelligibility and naturalness a voice conversion from esophageal speech into normal speech is performed. The conversion of esophageal speech to normal speech is done in a probabilistic manner by training Gaussian Mixture Models of the joint probability densities between acoustic features of esophageal speech and those of normal speech in advance. Based on the statistics of normal speech esophageal speech is converted and hence the quality of esophageal speech can be enhanced.

Keywords — *Esophageal speech, Speech Enhancement, Speech conversion.*

I INTRODUCTION

Laryngectomees who have undergone a total laryngectomy due toan accident or laryngeal cancer cannot produce speech sounds in a usual way because their vocal folds have been removed. They need alternative speaking methods for producing speech sounds. The produced speech is called alaryngeal speech and esophageal speech is a type of alaryngeal speech. In producing esophageal speech, excitation signals are produced by releasing gases from or through the esophagus, and then they are articulated. Esophageal speech sounds more natural than other types of alaryngeal speech, such as electrolaryngeal speech. However, the severe degradation of naturalness and intelligibility of esophageal speech caused by its specific production mechanism is observed compared with normal speech. Moreover, its voice quality is similar even if different laryngectomees speak. Consequently, esophageal speech also suffers from the severe degradation

As one of the statistical parametric speech synthesis techniques capable of manually controlling voice quality of synthetic speech, a multiple regression approach has been proposed in speech synthesis based on hidden Markov model (HMM). This regression approach has also been applied to one-to-many EVC. In this method, voice quality of various speakers is described by a few voice quality control parameters based on primitive word pairs expressing specific voice quality factors. To manually control converted voice quality without any target voice samples, a subspace spanned by a few representative vectors capturing the specific voice quality factors is formed in a statistical conversion model.

II ESOPHAGEAL SPEECH

Esophageal speech often includes some specific noisy sounds. These noises are produced through a process of generating excitation sounds, i.e., pumping air into the esophagus and the stomach and releasing air from them. Waveform envelope and spectral components of esophageal speech don't vary over an utterance as smoothly as those of normal speech. These unstable and unnatural variations cause the unnatural sounds of esophageal speech. Moreover, the pitch of esophageal speech is generally lower and less stable than that of normal speech. Consequently, a usual *F*0 analysis process for normal speech often fails at *F*0 extraction and the unvoiced/voiced decision in esophageal speech. These characteristics of esophageal speech cause severe degradation of analysis-synthesized speech quality.The intelligibility and naturalness of esophageal speech strongly depend on the skill of individual laryngectomees in

producing esophageal speech. However, some specific noises are essentially difficult to remove because they are caused by the production mechanism of esophageal speech.

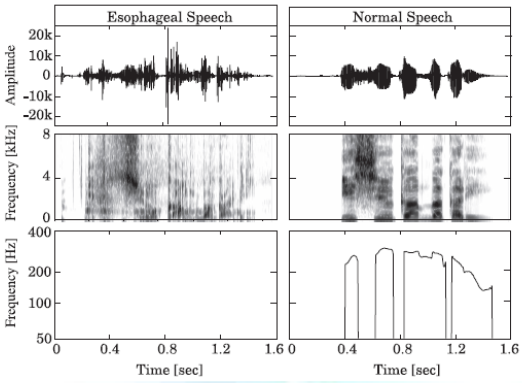


Fig1- Example of waveforms, spectrograms and *F*0 contours of both esophageal and normal speech.

III VOICE CONVERSION PROCESS

The conversion method based on maximum likelihood estimation of speech parameter trajectories considering a global variance (GV) as one of the state-of-the-art statistical VC methods. This method consists of a training and conversion process

A .TRAINING PROCESS

Using multiple parallel datasets between esophageal speech of a laryngectomee and normal speech of many pre-stored target speakers, the joint probability density function (*p.d.f.*) of a source feature vector of esophageal speech X_t and a target feature vector of the S^{th} pre-stored target speaker's normal speech, $Y_t^{(s)}$,at frame *t* is modeled by a one-to-many eigen voice GMM (EV-GMM) as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{w}^{(s)}, \boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix}; \begin{bmatrix}\mu_m^{(X)} \\ \mu_m^{(Y,s)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right)$$

$$\mu_m^{(Y,s)} = \boldsymbol{B}_m^{(Y)} \boldsymbol{w}^{(s)} + \boldsymbol{b}_{m,0}^{(Y)},$$

where $N(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is *M*. The mixture-component weight $\propto_m$, the source mean vector $\mu_m^{(X)}$, and the covariance matrices $\sum_m(XX)$, $\sum_m(XY)$, $\sum_m(YX)$, $\sum_m(YY)$, are tied over every target speaker.

To convert esophageal speech into normal speech, some parameters of normal speech, such as spectrum, aperiodic components, and *F*0, are estimated from a spectral parameter of esophageal speech. For the estimation of spectrum and aperiodic components, we independently train two EV-GMMs modeling the joint *p.d.f.*s of the spectral segment feature of esophageal speech and two features of the normal speech parameters, spectrum and aperiodic components, using corresponding joint feature vector sets.

B.ADAPTATION AND CONVERSION PROCESS

As for the adaptation of spectrum and aperiodic components, each EV-GMM is separately adapted to given target speech samples in an unsupervised manner. An optimum value of the adaptive vector $\boldsymbol{w}$ is determined by maximizing a marginal likelihood $P(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{\lambda})$ of the EV-GMM for the given target speech features $\boldsymbol{Y}$. For the *F*0 adaptation, global mean and standard deviation values are extracted from the given target speech samples.

In conversion, spectrum and aperiodic components are separately estimated from the spectral segment feature of esophageal speech using the corresponding adapted EV-GMMs. *F*0 is estimated from the spectral segment feature using the standard GMM. The maximum likelihood estimation method considering dynamic features and the global variance is used in these estimation processes. To adapt global *F*0 characteristics to those of the given target speech samples, the estimated *F*0 pattern is linearly transformed so as to its mean and standard deviation values over an utterance is equivalent to the target values.

C. VOICE QUALITY CONTROL IN ES-TO-SPEECH

voice quality control methods in ES-to-Speech to make it possible to manually control converted voice quality. It is essential in voice quality control to design an intuitively controllable parameter to be manipulated. One promising approach is to use perceptual scores expressing specific voice quality factors. In the literature, several primitive word pairs to efficiently represent voice quality of various speakers, such as male/female for gender or

elder/younger for age, have been extracted through a large-sized perceptual evaluation using normal speech of a lot of speakers.

IV FEATURE EXTRACTION

The spectral components of esophageal speech vary unstably. To alleviate this issue, spectral segment feature extracted from multiple frames is used. At each frame, a spectral parameter vector at the current frame and those at several preceding and succeeding frames are concatenated. Although it is difficult to extract *F*0 from esophageal speech we usually perceive pitch information of esophageal speech. The spectral segment feature as an input feature for estimating *F*0 in the conversion process. Moreover, in order to make the estimated *F*0 correspond to the perceived pitch information of esophageal speech, as an output feature *F*0 values extracted from normal speech uttered by a non-laryngectomee is used so as to make its prosody similar to that of esophageal speech.

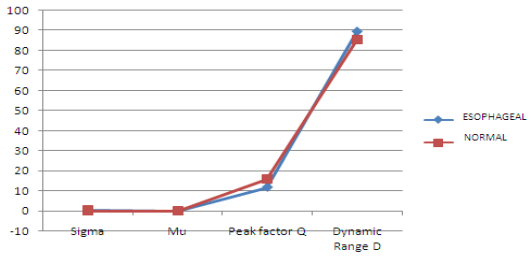
A.NORMAL SPEECH

Sigma	Mean	Peak factor	Dynamic range
0.25671	3.15E-	11.8113	89.6159

B.ESOPHAGEAL SPEECH

Sigma	Mean	Peak factor	Dynamic range
0.16064	-0.0068564	15.875	85.473

C.DEVIATIONS FOUND



Based on the deviations found the esophageal speech is trained according to the values of normal speech such that converted speech has an enhanced quality when compared to the esophageal speech.

V CONCLUSION

This paper presents a novel method for enhancing esophageal speech using statistical voice conversion. The proposed method (ES-to-Speech) converts a spectral segment feature of esophageal speech into spectrum, *F*0, and aperiodic components of normal speech independently using three different GMMs. The experimental results have demonstrated that ES-to-Speech yields significant improvements in intelligibility and naturalness of esophageal speech.

VI REFERENCES

[1] 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. & Syst.*, Vol. E93-D, No. 9, pp. 2472–2482, 2010.
[2] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.
[3] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *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, 2007.
[5] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans. Inf. & Syst.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.
[6] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Regression approaches to voice quality control based on one-to-many eigenvoice conversion. *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101– 106, Bonn, Germany, Aug. 2007.
[7] H. Kido and H. Kasuya. Everyday expressions associated with voice quality of normal utterance —Extraction by perceptual evaluation—. *J. Acoust. Soc. Jpn.*, Vol. 57, No. 5, pp. 337-344, 2001 [in Japanese].
[8] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. SAP*, Vol. 8, No. 6, pp. 695–707, 2000.
[9] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveign´e. 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.