E-026

Noise Robust Voice Conversion using GA-based Informative Feature

澤田 耕平† 田上 陽嗣† 田村 哲嗣† 竹原 正矩† 速水 悟†
Kohei Sawada Yoji Tagami Satoshi Tamura Masanori Takehara Satoru Hayamizu

1. Introduction

Speech (voice) is one of the crucial methods in human communication. It is therefore essential for people who cannot speak, e.g. laryngectomees who have had an operation for laryngectomy due to laryngeal cancer, to use an alternative way to produce speech. One of the ways to do so is using an electrolarynx (EL) which provides EL speech. By using EL, such the people can substitute their own vocal cord and communicate in the natural manner. However, there are some problems in vocalization using EL; speech generated by EL is quite artificial and does not contain any acoustic property about speaker. If original speech data of vocally-disabled person are available, which were recorded before the person lost voice, then it is possible to converted EL speech into natural speech by applying voice conversion (VC) techniques. In this paper, VC based on maximum-likelihood estimation using Gaussian mixture models (GMMs) is chosen [1]. The VC method can convert EL speech with high quality by conducting soft clustering. In this paper, we propose a VC technique using noise-robust acoustic features: GA-based informative feature (GIF). GIF is designed to improve the performance of various pattern recognitions, and as a result, it is found that GIF has robustness against noise [2]. It is thus expected to increase the robustness of VC technique and to improve the quality of converted speech, by applying GIF to VC.

2. Voice Conversion

This section summarizes a VC method used in this paper, proposed in [1]. A GMM is at first built using training data, and secondly, output features are converted from input features. Details of the VC method should be referred to [1].

2.1 Training

Acoustic features are extracted from training data consisting of utterance pairs made by source and target speakers. Let us denote a source feature by $\boldsymbol{S}_l$ and a target feature having static and dynamic parameters by $\boldsymbol{T}_l = [\boldsymbol{t}_l^\top, \boldsymbol{\Delta t}_l^\top]^\top$, where $\top$ indicates transposition of a vector. Using the source and target features, a GMM is trained computing a joint probability density $P(\boldsymbol{S}_l, \boldsymbol{T}_l|\lambda)$ as followds:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_l P(\boldsymbol{S}_l, \boldsymbol{T}_l|\lambda) \tag{1}$$

where λ denotes model parameters.

2.2 Conversion

Using a trained GMM, output features can be obtained based on maximum likelihood estimation. A source feature sequence $\boldsymbol{S} = [\boldsymbol{S}_1^\top, \dots, \boldsymbol{S}_L^\top]^\top$, a target sequence $\boldsymbol{T} = [\boldsymbol{T}_1^\top, \dots, \boldsymbol{T}_L^\top]^\top$, and a target static feature sequence $\boldsymbol{t} = [\boldsymbol{t}_1^\top, \dots, \boldsymbol{t}_L^\top]^\top$ are utilized, where L donates the number of frames. The conversion is then performed maximizing the likelihood function $P(\boldsymbol{T}|\boldsymbol{S}, \lambda)$. A converted feature sequence is determined as follows:

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t}}{\operatorname{argmax}}\, P(\boldsymbol{T}|\boldsymbol{S}, \lambda) = \underset{\boldsymbol{t}}{\operatorname{argmax}}\, P(\boldsymbol{W}\boldsymbol{t}|\boldsymbol{S}, \lambda) \tag{2}$$

†Department of Information Science, Gifu University,
1-1 Yanagido, Gifu, Gifu, 501-1193, Japan.
kouhei@asr.info.gifu-u.ac.jp

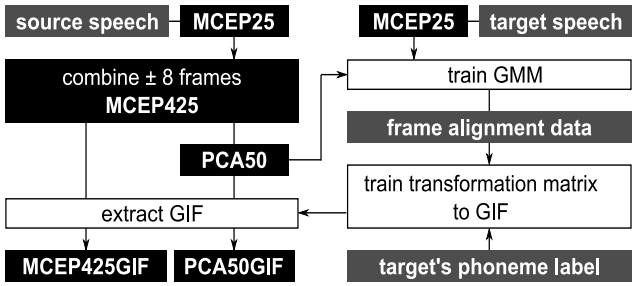

Fig. 1 *Features employed in this paper. PCA50 is used in conventional VC, and MCEP425GIF and PCA50GIF are proposed features.*

where $\boldsymbol{W}$ is a transformation matrix to extend a static feature sequence to static and dynamic feature sequence.

3. VC using GA-based informative Feature

In speech recognition, GIF can greatly improve accuracy in noise environment [2]. Since GIF provides robustness against noise in speech recognition, it is expected that VC using GIF can convert input speech data more robust and precisely than conventional VC, even in noise conditions. Details of GIF should be referred to [2].

At first, a N-dimensional input vector $\boldsymbol{x}$ is converted into a C-dimensional intermediate vector $\boldsymbol{y}$ as:

$$\boldsymbol{y} = \boldsymbol{A}\,(\boldsymbol{x}^\top\ 1)^\top \tag{3}$$

In Eq.(3), $\boldsymbol{A}$ is a $C \times (N+1)$ transformation matrix, where C is the number of classes that should be classified. In the next process, the vector $\boldsymbol{y}$ is further converted into an M-dimensional output feature vector (GIF) $\boldsymbol{z}$ as:

$$\boldsymbol{z} = \boldsymbol{B}\boldsymbol{y} \tag{4}$$

where $\boldsymbol{B}$ is an $M \times C$ transformation matrix. These matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are computed by Genetic Algorithm (GA).

The process to obtain the converted features from spectral segmental features of the EL speech (input feature) is shown in the following (1) to (4).

(1) The training data which consist of spectral segmental features for the source EL speech with phoneme labels are created using the forced alignment results and the frame alignment results.
(2) Positive data for each phoneme are extracted from the training data. Negative data are extracted from the data of the other phonemes so that the number of negative data is equivalent to that of positive data.
(3) GIF training matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are obtained using the positive and negative data.
(4) Converted features $\boldsymbol{z}$ are calculated from spectral segmental features $\boldsymbol{x}$ using the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

4. Experiment

4.1 Experimental condition

We collected speech utterances of one male laryngectomee speaking with EL as a source speaker and those of one male non-laryngectomee speaker as target speech. Speech data of each person include 50 phoneme-balanced sentences. 40 sentence

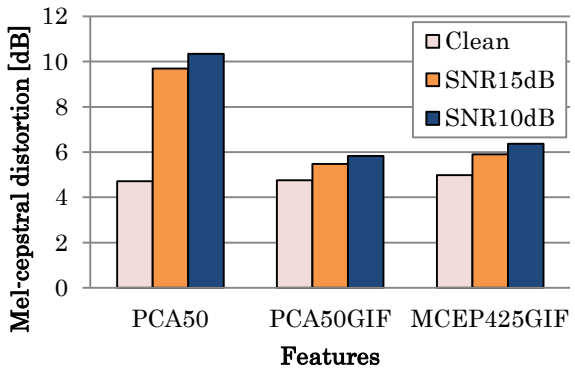

Fig. 2 *Mel-CDs of three kinds of feature (PCA50, PCA50GIF, and MCEP425GIF) in clean and white noises environments.*

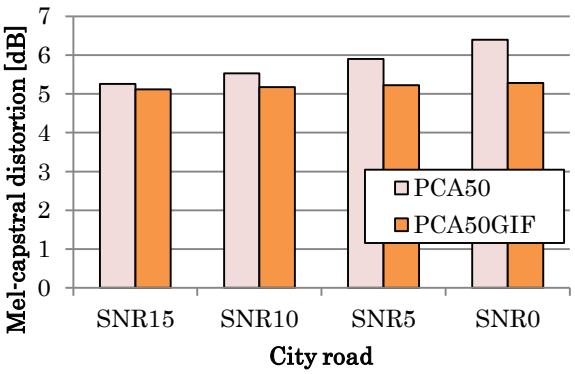

Fig. 3 *Mel-CDs of PCA50 and PCA50GIF in real environments (in-car noises on city road).*

pairs were used for training and the rest 10 sentences were used for testing.

Regarding features, the 0th to24th mel-cepstral coefficients were at first extracted (**MCEP25**) as a basic feature of source speech in which the 0th coefficient captured power information, where a frame size and a frame shift were 5 msec. For the source speech, a 425-dimensional feature (**MCEP425**) consisting of a current frame vector as well as previous and incoming 8 vectors was obtained at every frames. Afterwards, 50-dimensional components were extracted from MCEP425 by principal component analysis (PCA), which is the feature used in the conventional VC, and we call this feature **PCA50**. A cumulative contribution was 95.40%.

For the other experimental setups, the number of mixture components of the GMM to estimate output features was 32 or 64. The number of phoneme class used in GIF was 27. The phonemes that had small numbers of samples were integrated into the similar phonemes. The features of target speech included 0th to 24th mel-cepstral coefficients.

We conducted experiments in three environments: (1) clean, (2) white noise (SNR15dB and SNR10dB), and (3) real environmental noise (city road and expressway). We compared three VC methods using the following features respectively.

(i) The conventional feature (PCA50)
(ii) GIF extracted from PCA50 (**PCA50GIF**)
(iii) GIF extracted from MCEP425 (**MCEP425GIF**)

These features (PCA50, PCA50GIF, and MCEP425GIF) used in this experiment are illustrated in Figure 1.

4.2 Objective evaluation

We conducted objective evaluations for the three features. A mel-cepstral distortion (Mel-CD) [dB] between the target and converted mel-cepstra was used as the objective evaluation measure.

Figure 2 shows Mel-CDs of PCA50, PCA50GIF, and MCEP425GIF in clean and white-noise environments (SNR15dB and SNR10dB). In the clean environment, the results show that the performance of PCA50GIF was roughly equivalent to that of PCA50. However, Mel-CD of MCEP425 was slightly worse than PCA50 and PCA50GIF. This is because that the scale of MCEP425 increased by orthogonalization in PCA. In addition, from these results in the white noise (SNR15dB and SNR10dB), the proposed features were better than the conventional feature. Figure 3 illustrates Mel-CDs of PCA50 and PCA50GIF in real environments. According to Figure 3, the larger noise became, the larger Mel-CDs of conventional feature were. Furthermore, the Mel-CD of the proposed feature was almost constant even if the noise level became large. Similar results were obtained when the noise was expressway. Note that PCA50GIF was adopted as a proposed feature because the performance was better than MCEP425GIF. As shown, GIF can reduce the degradation of the performance in different environments. Since it is expected that EL is used in real environments, the robustness of EL is essential. Our proposed method is consequently effective in practical use.

5. Conclusion

This paper proposed a statistical voice conversion (VC) method using GA-based informative feature (GIF), in order to accomplish noise-robust VC from electrolaryngeal speech (EL speech) to natural speech. We evaluated the proposed method with the mel-cepstral distortion (Mel-CD) between the acoustic feature of converted speech and target speech. In the clean environment, the difference of Mel-CD between the conventional method and the proposed method is not significant. But in the noise environment the difference becomes larger, indicating the effectiveness of the proposed method.

Our feature works includes: (1) further investigation of improvements in our method, (2) comparison of other features that are robust against noise, and (3) combination of boosting silent EL speech technique using non-audible murmur (NAM) microphone [4][5].

Acknowledgment

The authors are deeply grateful to Prof. Tomoki Toda of Nara Institute of Science and Technology, Japan, for giving valuable advices and providing techniques of voice conversion.

References

[1] T.Toda et al., “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory” IEEE Trans.Audio,Speech and Language, vol.15, no.8, pp.2222-2235, 2007

[2] 田村哲嗣ら, “GIF-SP: 汎用・識別的な特徴量を用いた音声認識性能の改善” 電子情報通信学会 技術研究報告, SP2011-92, vol.111, no.365, pp.119-124 ,2011

[3] 大谷大和ら, “STRAIGHT 混合励振源を用いた混合正規分布モデルに基づく最ゆう変換法” 信学論(D), vol.J-91-D, no.4, pp.1082-1091, 2008

[4] 中島淑貴ら, “非可聴つぶやき認識” 信学論(D-II), vol.J87-D-II, no.9, pp.1757-1764, 2006

[5] 中島淑貴ら, “無音声認識(NAM 認識)におけるセンシング方法の改善” 音響講論集, 3-Q-1, pp.145-146, 2004