Statistical Voice Conversion using GA-based Informative Feature

Kohei Sawada, Yoji Tagami, Satoshi Tamura, Masanori Takehara, and Satoru Hayamizu
Department of Information Science, Gifu University, Gifu, Japan
E-mail: {kouhei@asr.info., tagami@asr.info., tamura@info., takehara@asr.info., hayamizu@}gifu-u.ac.jp

***Abstract*— In order to make voice conversion (VC) robust to noise, we propose VC using GA-based informative feature (GIF), by adding an extraction process of GIF to a conventional VC. GIF is proposed as a feature that can be applied not only in pattern recognition but also in relative tasks. In speech recognition, furthermore, GIF could improve recognition accuracy in noise environment. We evaluated the performances of VC using spectral segmental features (conventional method) and GIF, respectively. Objective experimental result indicates that in noise environments, the proposed method was better than the conventional method. Subjective experiment was also conducted to compare the performances. These results show that application of GIF to VC was effective.**

I. INTRODUCTION

Speech (voice) is one of the crucial methods in human communication; people can easily communicate with others by speech, comparing to the other methods (writing, gesturing, etc.). 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 electrolaryngeal speech (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. These make it difficult to understand speech contents. If original speech data of vocally-disabled person are available, which were recorded before the person lost voice, then it is possible to convert EL speech into natural speech by applying voice conversion (VC) techniques. VC is applied to speaking-aid system which enhances EL speech [4], STRAIGHT-based VC [5] and voice-quality control adaptation [6]. 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. VC is so useful for natural speech communication in such the situations, however, VC has still some issues to be overcome; in real environments, it may happen that background noises are overlapped with EL speech, degrading the quality of converted speech. Therefore, noise-robust speech feature is expected for real application.

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. In order to evaluate the effectiveness of our proposed scheme, in this paper objective and subjective experiments were conducted comparing GIF with conventional acoustic features.

This paper is organized as follows. In Section II, conventional VC used in this paper is overviewed. Section III introduces a proposed feature GIF. In Section IV, the proposed VC method is explained. Experiments and evaluations are described in Section V. Finally, this paper is concluded in Section VI.

II. 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].

A. 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 $\boldsymbol{T}_l = \left[\boldsymbol{t}_l^{\top}, \boldsymbol{\Delta t}_l^{\top}\right]^{\top}$, where $\top$ indicates transposition of a vector. $\boldsymbol{S}_l$ and $\boldsymbol{T}_l$ are time-aligned features determined by dynamic time warping (DTW). Using the source and target features, a GMM is trained computing a joint probability density $P(\boldsymbol{S}_l, \boldsymbol{T}_l|\lambda)$ as follows:

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

where λ denotes model parameters. The joint probability density is written as

$$P(\boldsymbol{S}_l, \boldsymbol{T}_l|\lambda) = \sum_{i} w_i N\left(\boldsymbol{U}_l; \boldsymbol{\mu}_i^{(U)}, \boldsymbol{\Sigma}_i^{(UU)}\right) \tag{2}$$

$$\boldsymbol{U}_l = \begin{bmatrix}\boldsymbol{S}_l \\ \boldsymbol{T}_l\end{bmatrix},\ \boldsymbol{\mu}_i^{(U)} = \begin{bmatrix}\boldsymbol{\mu}_i^{(S)} \\ \boldsymbol{\mu}_i^{(T)}\end{bmatrix},\ \boldsymbol{\Sigma}_i^{(UU)} = \begin{bmatrix}\boldsymbol{\Sigma}_i^{(SS)} & \boldsymbol{\Sigma}_i^{(ST)} \\ \boldsymbol{\Sigma}_i^{(TS)} & \boldsymbol{\Sigma}_i^{(TT)}\end{bmatrix} \tag{3}$$

where $N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ shows a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. A weight of an i-th mixture component is w_i. $\boldsymbol{\mu}_i^{(S)}$ and $\boldsymbol{\mu}_i^{(T)}$ are mean vectors of an i-th mixture component for the source and for the target,

respectively. Matrices $\boldsymbol{\Sigma}_i^{(SS)}$ and $\boldsymbol{\Sigma}_i^{(TT)}$ are covariance matrices, and $\boldsymbol{\Sigma}_i^{(ST)}$ and $\boldsymbol{\Sigma}_i^{(TS)}$ are cross-covariance matrices of an i-th mixture component.

B. Conversion

Using a trained GMM, output features can be obtained based on maximum likelihood estimation. In the conversion, 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:

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t}}{\operatorname{argmax}}\, P(\boldsymbol{T}|\boldsymbol{S}, \lambda) = \underset{\boldsymbol{t}}{\operatorname{argmax}}\, P(\boldsymbol{W}\boldsymbol{t}|\boldsymbol{S}, \lambda) \tag{4}$$

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

III. GA-BASED INFORMATIVE FEATURE

In this section, our proposed feature GIF is introduced. This feature can be utilized in various pattern recognition tasks and related works [2], [3].

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

$$\boldsymbol{y} = \boldsymbol{A}\,(\boldsymbol{x}^\top\ 1)^\top \tag{5}$$

In Eq.(5), $\boldsymbol{A}$ is a $C \times (N+1)$ transformation matrix, where C is the number of classes that should be classified. This process is called "Stage 1." In the next process "Stage 2," the vector $\boldsymbol{y}$ is further converted into an M-dimensional output feature vector (GIF) $\boldsymbol{z}$ as:

$$\boldsymbol{z} = \boldsymbol{B}\boldsymbol{y} \tag{6}$$

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

A. Stage 1: Getting a first transformation matrix

A binary classifier which distinguishes an input vector into a certain class or its complementary class is focused on. The transformation matrix $\boldsymbol{A}$ is eventually obtained by the following process for all classes.

1) Step 1: Building a candidate classifier:

For an i-th class $(1 \le i \le C)$, assume a linear classifier f for an input vector $\boldsymbol{x} = (x_j)$**:**

$$f(\boldsymbol{x};\ \boldsymbol{a}_i) = \left(\sum_{j=1}^{N} a_{i,j} x_j\right) + a_{i,N+1} \tag{7}$$

where $\boldsymbol{a}_i = (a_{i,1}, \cdots, a_{i,N}, a_{i,N+1})$ includes classifier parameters corresponding to a part of $\boldsymbol{A}$. The classifier is designed to return a positive value if $\boldsymbol{x}$ is in the class or a negative value otherwise. The parameters are computed using a training set $R = \{\boldsymbol{r}_n\}$ and the standard GA:

(i) Initialization
An initial population G_0 including K individuals is created. An individual has $(N+1)$ chromosomes, each of which encodes a classifier parameter.

(ii) Fitness function
For a k-th individual $\boldsymbol{v}_k$ in an h-th generation G_h, a fitness function $E(\boldsymbol{v}_k)$ is calculated as:

$$E(\boldsymbol{v}_k) = \sum_{n=1}^{|R|} q_n \cdot sgn\big(f_i(\boldsymbol{r}_n;\ \boldsymbol{a}_i)\big) \tag{8}$$

where $\boldsymbol{a}$ is a parameter set obtained by decoding $\boldsymbol{v}_k$, and q_n is a transcribed label that corresponds to 1 if $\boldsymbol{r}_n$ belongs to the class (positive data) or -1 otherwise (negative data). The minimum value of $E(\boldsymbol{v}_k)$ is set to 1.

(iii) Elitist selection, inheritance, mutation and crossover
Conventional GA operations are employed to form a next generation G_{h+1} from a current generation G_h; elitist selection and inheritance are applied to copy a certain individual to G_{h+1}; for genetic diversity, mutation and crossover are also conducted to generate a new individual that is different from its parent(s).

(iv) Generation change
The above processing ((ii) and (iii)) is repeated from $h = 0$ to $h = F - 1$. A final population G_F is then generated.

2) Step 2: Completing a binary classifier:

From G_F, individuals having the (K/I) highest fitness value are extracted and added to a candidate population G_C. By repeating the step I times, the population is completed. Step 1 is then applied again where G_C is used as an initial generation, then the best-fit individual $\hat{\boldsymbol{v}}$ is obtained. The transformation parameter set $\hat{\boldsymbol{a}}$ is consequently acquired by decoding the solution $\hat{\boldsymbol{v}}$. It is often pointed out that the solution is not stable since GA highly depends on an initial population and operations randomly determined. The two-step approach enables us to compute a stable GA solution.

B. Stage 2: Getting a second transformation matrix

To enhance discriminative and recognition abilities, and to reduce the dimension of feature vectors, we employ a second-stage procedure explained below:

(i) For an i-th class, a mean vector $\bar{\boldsymbol{\mu}}_i$ is calculated as:

$$\bar{\boldsymbol{\mu}}_i = \frac{1}{|R_i|} \sum_{\boldsymbol{r} \in R_i} \boldsymbol{A}\,(\boldsymbol{r}^\top\ 1)^\top \tag{9}$$

where R_i is a subset of the training data, in which all vectors belong to the i-th class.

(ii) Let us denote a linear transformation g, for a vector $\boldsymbol{y} = (y_i)$ obtained in the first stage, by the following equation:

$$g(\boldsymbol{y};\ \boldsymbol{b}_m) = \sum_{j=1}^{C} b_{m,j} y_j \tag{10}$$

where $\boldsymbol{b}_m = (b_{m,1}, \cdots, b_{m,C})$ indicates classifier parameters and a part of $\boldsymbol{B}$.

(iii) For $m = 1$, the projection g is determined so that a variance of transformed mean vectors would be maximized. The parameter set $\boldsymbol{b}_1$ is optimized by applying GA explained previously, where the fitness function is modified as:

$$E(\boldsymbol{v}_k) = var(w_1, \cdots, w_C) \quad \text{where } w_i = g(\bar{\boldsymbol{\mu}}_i;\ \boldsymbol{b}) \tag{11}$$

In Eq.(11), $\boldsymbol{b}$ is obtained by decoding $\boldsymbol{v}_k$.

(iv) For $m = 2$, $\boldsymbol{b}_2$ is optimized so as to maximize a variance just as same as $\boldsymbol{b}_1$, under the constraint that an inner product between $\boldsymbol{b}_1{}^\top$ and $\boldsymbol{b}_2{}^\top$ should be zero**:**

$$\langle \boldsymbol{b}_1{}^\top, \boldsymbol{b}_2{}^\top \rangle = \sum_{j=1}^{C} \boldsymbol{b}_{1,j} \boldsymbol{b}_{2,j} = 0 \tag{12}$$

(v) For any $m\ (2 \le m \le M)$, an m-th parameter set $\boldsymbol{b}_m$ is calculated in the same way, under the restriction that all inner products should be zero.

C. Feature vector computation

Once the first projection $\boldsymbol{A}$ and the second projection $\boldsymbol{B}$ are determined, a feature vector $\boldsymbol{z}$ can be computed by applying Eqs.(5) and (6). Before applying the second projection, a bias vector $\boldsymbol{\mu}$ is calculated beforehand as:

$$\boldsymbol{\mu} = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{y}_l = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{A}\,(\boldsymbol{x}_l{}^\top\ 1)^\top \tag{13}$$

where $\boldsymbol{X} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x}_L)$ is a sequence of input vectors. Subsequently, each intermediate vector is normalized by suppressing the bias vector:

$$\hat{\boldsymbol{y}}_l = \boldsymbol{y}_l - \boldsymbol{\mu} \tag{14}$$

IV. VC USING GIF

In speech recognition, GIF can greatly improve accuracy in noise environments [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. Thus, in this paper GIF is applied to VC to overcome noise distortion.

Time-aligned phoneme labels of source speakers are needed in the GIF training. It is difficult to estimate the phoneme labels of source speakers, but it is easier those of target speakers, since in this case speech data of source speakers are EL speech, whereas target data are natural speech. Therefore, we make training data using frame alignment results between source and target speakers as well as forced alignment results of target speakers in order to train GIF transformation. The process to obtain the converted features $\boldsymbol{s}$ from spectral segmental features of the EL speech (input feature) is shown in the following (i) to (iv).

(i) 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.

(ii) 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.

(iii) GIF matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are obtained using the positive and negative data, according to Section III.

(iv) Converted features $\boldsymbol{z}$ are calculated from spectral segmental features $\boldsymbol{x}$ using the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

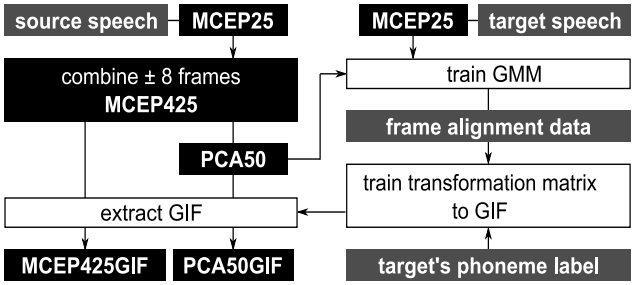

Fig. 1 *Features employed in this paper. PCA50 is used in conventional VC, and MCEP425GIF and PCA50GIF are proposed features.*

V. EXPERIMENTS

A. Experimental condition

We collected speech utterances of one male laryngectomee speaking with EL as source speech and those of one male non-laryngectomee speaker as target speech. Speech data of each person include 50 phoneme-balanced sentences. 40 sentence pairs were used for training and the rest 10 sentences were used for testing.

Regarding features, the 0th to 24th 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-dimentional 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 ratio was 95.40%.

For the other experimental setups, the numbers of mixture components of the GMM to estimate spectrum, F_0, and aperiodic components [1] were 32, 64, and 32, respectively. Parameters in GIF were employed as follows: N = 50 (for PCA50) or 425 (for MCEP425), C = 27, M = 50, K = 1000, F = 30, and I = 50. The number of phoneme classes used in GIF was 27. The phonemes that had small numbers of samples were integrated into the similar phonemes. The acoustic features of target speech were MCEP25.

We conducted experiments in three environments: (1) clean, (2) white noise 15dB (SNR15dB) and (3) white noise 10dB (SNR10dB). We compared three VC methods using the following source features respectively:

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

These features used in this experiment are illustrated in Figure 1. Note that a source feature and a target feature were assumed to be clean data when training a GMM. GIF matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ obtained from the clean features were commonly used in the three environments.

B. Objective evaluation

First, 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 an objective

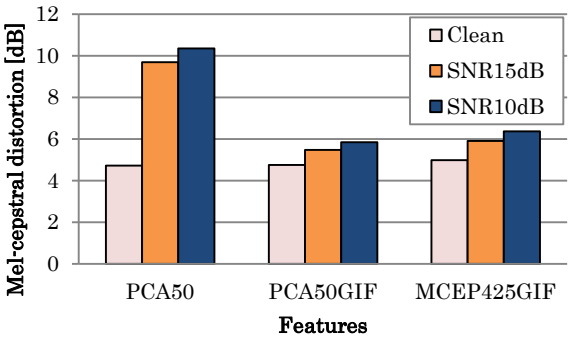

Fig. 2 *Mel-CDs of three kinds of feature (PCA50, PCA50GIF and MCEP425GIF) in three environments (clean, SNR15dB and SNR10dB).*

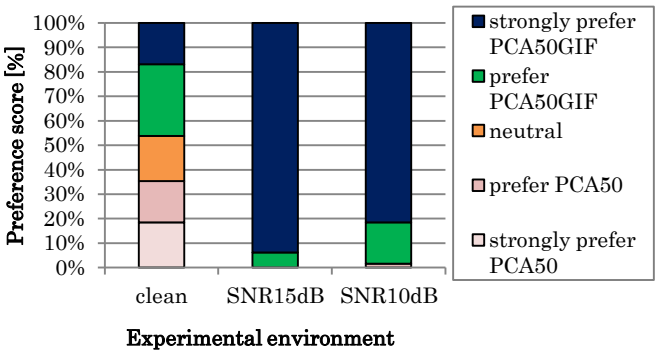

Fig. 3 *The intelligibility of converted speech of conventional and proposed VCs in three experimental environments (clean, SNR15dB, and SNR10dB).*

evaluation measure.

Figure 2 shows Mel-CDs of PCA50, PCA50GIF, and MCEP425GIF in the three experimental environments (clean, SNR15dB and SNR10dB). In the clean environment, Mel-CD of PCA50 was 4.71dB and that of PCA50GIF was 4.75dB. Those results show that the performance of PCA50GIF was roughly equivalent to that of PCA50. However, Mel-CD of MCEP425 was 4.98dB and it 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 SNR15dB and SNR10dB, the proposed features were better than the conventional feature. As shown, GIF is designed to be discriminative, and it can reduce the degradation of the performance in different environments. Therefore, the proposed method could estimate the output feature better than conventional method. 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.

C. Subjective evaluation

Next, we conducted subjective evaluations to estimate the intelligibility of converted speech. Five pairs each consisting of converted speech waveforms obtained from the conventional VC (using PCA50) and the proposed VC (using PCA50GIF) were prepared in every three environments. Note that PCA50GIF was adopted as a proposed feature because the performance was better than MCEP425GIF. In the evaluation, a subject listened to each pair (speech A and speech B) and afterwards gave a relative assessment by choosing five-scale scores. We carefully prepared every pairs in order to avoid the order effect; two speech data in a pair were randomly shuffled. As a result, in this experiment, we collected 13 subjects each who made 15 assessments.

Figure 3 shows the result of subjective evaluations in the three environments. In the clean environment, the intelligibility of the proposed VC is slightly better than the conventional VC. The difference is obvious in noisy environments (SNR15dB and SNR10dB). These results indicate that the proposed VC is superior to the conventional VC particularly in noisy conditions. Note that the result of the proposed method in SNR10dB was a little degraded compared to that in SNR15dB. Converted speech obtained by the proposed method was also damaged in heavily noisy environments, as a result, the difference between the conventional and proposed VCs in SNR10dB became smaller.

VI. CONCLUSION

This paper proposes a statistical VC method using GIF, in order to accomplish noise-robust VC from EL speech to natural speech. We evaluated the proposed method with the Mel-CD distortion 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. Subjective results also show the application of GIF to VC is effective. As our future work, comparison of our method with the other VC methods and source speech is expected.

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] S. Tamura et al., "GIF-SP: GA-based Informative Feature for Noisy Speech Recognition", Proc. APSIPA ASC 2012, 2012.

[3] N. Ukai et al., "GIF-LR: GA-based Informative Feature for Lipreading", Proc. APSIPA ASC 2012, 2012.

[4] K. Nakamura et al., "The Use of Air-Pressure Sensor in Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion", Proc. INTERSPEECH2010, pp.1628-1631, 2010.

[5] Y. Ohtani et al., "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation", Proc. INTERSPEECH2006, pp.2266-2269, 2006.

[6] K. Ohta et al., "Adaptive Voice-Quality Control Based on One-to-Many Eigenvoice Conversion", Proc. INTERSPEECH2010, pp.2158-2161, 2010.