Arbitrary speaker conversion based on speaker space bases constructed by deep neural networks

Tetsuya Hashimoto*, Daisuke Saito† and Nobuaki Minematsu*
* Graduate School of Engineering, The University of Tokyo, Japan
† Graduate School of Information Science and Technology, The University of Tokyo, Japan
E-mail: {hashib, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—This paper proposes a novel approach to construct a Deep Neural Network (DNN) based voice conversion (VC) system, where DNNs are integrated with speaker eigenspace. The proposed network consists of multiple DNNs and each of them converts input features to features corresponding to a base of eigenspace. Training of these DNNs is achieved with the assistance of Eigenvoice GMM (EVGMM). Experimental evaluations using one-to-many VC tasks show that the proposed method achieved better performance compared with that of EVGMM.**

I. Introduction

Voice conversion (VC) is a technique to modify an input utterance of a speaker so that it sounds as if it is generated by another speaker while its linguistic content is preserved. This technique has been directly applied to postprocessing of Text-to-Speech [1] and indirectly applied to speech enhancement [2], and so on.

In VC studies, statistical approaches have been often used for mapping features of a source speaker to those of a target one. Recently, approaches based on Gaussian mixture models (GMM) or neural networks (NN) have been widely investigated [3], [4]. To construct a conversion model on these approaches, a parallel speech corpus, which consists of the same sentences read by the source and target speakers, is required. However, the constructed model can be used only to modify the speaker identity of that source speaker's arbitrary utterances to that of that target speaker. Namely, the model can be applied only to that specific speaker pair. Hence, researchers pay special attention to performance improvement of conversion from/to open speakers. For this purpose, parameter adaptation techniques have been investigated and applied to GMM-based approaches [5], [6].

Although VC based on deep neural networks (DNN) achieves some performance improvement compared to GMM-based VC, since functions of each layer and node in DNN are difficult to interpret, the constructed DNN is generally low in its flexibility and can be applied only to a specific speaker pair. Then, the flexible control of speaker identities in DNN-based approaches is still a problem to solve.

In VC based on GMM, eigenvoice conversion (EVC) [7] and tensor-based VC [8] achieve some improvements of its ability to control speaker identities by using pre-stored data. Also in VC based on DNN, using pre-stored data for training achives some performance improvements [10]. In the DNN-based VC, however, as we noted before, a flexible control of speaker identity is not yet realize. Then in this paper, we propose a DNN-based VC method using eigenspace feature. Inspired by the EVC method, we expect that speech features of any input speaker can be divided into and characterized by eignespace components. To realize this concept, in this paper, we propose an architecture which consists of multiple DNNs to convert input features of a speaker into their eigenspace components. Once these DNNs are trained, target features are represented by the weighted sum of their outputs. These weight parameters can be estimated in an unsupervised manner.

The remainder of this paper is organized as follows. Section 2 describes EVC. Then, section 3 shows our proposed method using multiple DNNs and EVGMM. In section 4, experimental evaluation about one-to-many VC is described. Finally, section 5 concludes this paper.

II. Eigenvoice conversion (EVC)

In this section, one-to-many EVC is described. Let D, M, and s be the dimension of input and output features, the number of mixture components and the index of a target speaker, respectively. Feature vectors of a source speaker $\mathbf{X}_t$ and those of the s-th target speaker $\mathbf{Y}_t^{(s)}$ are modeled by EVGMM. The joint probability density is described as follows:

$$P(\mathbf{X}_t, \mathbf{Y}_t^{(s)} \mid \lambda^{(EV)}, \mathbf{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}([\mathbf{X}_t^\top, \mathbf{Y}_t^{(s)\top}]^\top; \mu_m^{(Z)}(\mathbf{w}^{(s)}), \mathbf{\Sigma}_m^{(Z)}) \quad (1)$$

$$\mu_m^{(Z)}(\mathbf{w}^{(s)}) = \begin{bmatrix} \mu_m^{(X)} \\ \mathbf{B}_m \mathbf{w}^{(s)} + \mathbf{b}_m^{(0)} \end{bmatrix} \quad (2)$$

$$\mathbf{\Sigma}_m^{(Z)} = \begin{bmatrix} \mathbf{\Sigma}_m^{(XX)} & \mathbf{\Sigma}_m^{(XY)} \\ \mathbf{\Sigma}_m^{(YX)} & \mathbf{\Sigma}_m^{(YY)} \end{bmatrix} \quad (3)$$

$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ is a Gaussian distribution with a mean vector μ and a covariance matrix $\mathbf{\Sigma}$. α_m means the weight for the m-th component. $\lambda^{(EV)}$ denotes the parameters of EVGMM which are independent of the target speakers.

In EVGMM, S pre-stored speakers are used to derive K base speakers ($K < S$), by using a linear combination of whom, the mean vector of any target speaker can be represented. In EVGMM, individualities of the target speaker are controlled by K-dimensional weight vector $\mathbf{w}^{(s)}$. It means, speaker space is represented by K eigenspace supervector

$\mathbf{B} = [\mathbf{B}_1^\top, \mathbf{B}_2^\top, \dots, \mathbf{B}_K^\top]^\top \in \mathcal{R}^{DM \times K}$ and bias supervector $\mathbf{b} = [\mathbf{b}_1^{(0)\top}, \mathbf{b}_2^{(0)\top}, \dots, \mathbf{b}_K^{(0)\top}]^\top \in \mathcal{R}^{DM \times 1}$. $\mathbf{B}$ can be considered as eigenspace bases.

In a training step, first, target-independent (TI) GMM is trained using parallel corpora between an anchor speaker and all pre-stored speakers. Then, using each of the corpora between the anchor speaker and a pre-stored speaker, target-dependent (TD) GMMs are trained by updating only their mean vectors. In this process, the TIGMM is used as an initial model for TDGMMs. The principal components analysis is applied to the mean vectors extracted from the constructed TDGMMs .

Finally, bias supervector $\mathbf{b}$, eigenspace supervector $\mathbf{B}$ and the weight for speaker s, $\mathbf{w}^{(s)}$ are determined.

When adapting the EVGMM to arbitrary target speaker, we estimate the weight vector $\mathbf{w}$ by maximum likelihood criterion. This adaptation means estimating the projection weights to each eigenspace super vector. This process is carried out in an unsupervised manner, and it can be realized with a small amount of data. Because the EVGMM is modeled as joint probability density, the many-to-many conversion system can be constructed using the many-to-one EVC and one-to-many EVC [9].

III. Construction of the speaker space based on DNN

A. Architecture

In EVC, from another viewpoint, the converted features are a linear combination of generated features which correspond to eigenbase parameters. Based on this idea, in the proposed method, DNN is utilized to convert the features of an source speaker to a set of base speaker features corresponding to that source speaker. Once the weight corresponding to the target is determined, the target's features are obtained as a weighted sum of outputs from the DNNs. To construct the above DNNs, the parallel data of the source and the base are required. By using EVGMM, these parallel data can be prepared rather easily. Finally, training of these DNNs is achieved by utilizing the prepared data.

B. Parallel data preparation based on EVGMM

In this section, preparation of parallel data using EVGMM is described. In EVC, the conversion from the feature from the source speaker $\mathbf{X}_t$ to that of the target speaker $\mathbf{Y}_t$ is denoted by equation (4);

$$F(\mathbf{X}_t) = \sum_{m=1}^{M} \gamma_{m,t} (\mathbf{B}_m \mathbf{w}^{(Y)} + \mathbf{b}_m^{(0)} + \mathbf{A}_m (\mathbf{X}_t - \mu_m^{(X)})). \quad (4)$$

$\gamma_{m,t}$ is a posterior probability of the m-th component given the input features, and $\mathbf{A}_m$ contains the variance and covariance matrices as shown below.

$$\gamma_{m,t} = P(m \mid \mathbf{X}_t, \lambda^{(EV)}), \quad (5)$$
$$\mathbf{A}_m = \mathbf{\Sigma}_m^{(YX)} \mathbf{\Sigma}_m^{(XX)-1} \quad (6)$$

Equation (4) can be modified as below:

$$F(\mathbf{X}_t, \lambda^{(EV)}) = \mathrm{TD}(\mathbf{X}_t, \lambda^{(EV)}) + \mathrm{SD}(\mathbf{X}_t, \lambda^{(EV)}), \quad (7)$$
$$\mathrm{TD}(\mathbf{X}_t, \lambda^{(EV)}) = \sum_{k=1}^{K} \mathbf{w}_k^{(Y)} \sum_{m=1}^{M} \gamma_{m,t} \mathbf{B}_{m,k},$$
$$\mathrm{SD}(\mathbf{X}_t, \lambda^{(EV)}) = \sum_{m=1}^{M} \gamma_{m,t} \{\mathbf{b}_m^{(0)} + \mathbf{A}_m (\mathbf{X}_t - \mu_m^{(X)})\},$$

where $\mathbf{w}_k^{(Y)}$ denotes the k-th dimensional element of $\mathbf{w}^{(Y)}$ and $\mathbf{B}_{m,k}$ denotes the k-th row vector of $\mathbf{B}_m$. In EVC, $\mathbf{b}_m^{(0)}$ means the average vector of all the pre-stored speakers about the m-th component. Hence, $\mathrm{SD}(\mathbf{X}_t, \lambda^{(EV)})$ can be regarded as features that depend on the average speaker, $\mathrm{TD}(\mathbf{X}_t, \lambda^{(EV)})$ represents the residuals of the target after subtracting the average speaker, and converted features $F(\mathbf{X}_t, \lambda^{(EV)})$ becomes a linear combination of them. $\mathrm{TD}(\mathbf{X}_t, \lambda^{(EV)})$ and $\mathrm{SD}(\mathbf{X}_t, \lambda^{(EV)})$ can be considered as target-dependent and source-dependent feature components.

As described in Section 2, target speaker individualities are controlled by K-dimensional weight vector $\mathbf{w}^{(s)}$, and $\mathbf{w}^{(s)}$ represents target-dependent weights for each eigenspace basis. Thereby, source speaker's feature can be converted to each eigenspace feature component by using 1-of-K coding instead of $\mathbf{w}^{(Y)}$. Let a feature of the k-th eigenbase corresnponding to $\mathbf{X}_t$ be $\mathbf{E}_t^k$. It can be represented as below:

$$\mathbf{E}_t^k = \sum_{m=1}^{M} \gamma_{m,t} \mathbf{B}_{m,k} \qquad (k = 1, 2, ..., K), \quad (8)$$
$$\mathbf{E}_t^0 = \sum_{m=1}^{M} \gamma_{m,t} \{\mathbf{b}_m^{(0)} + \mathbf{A}_m (\mathbf{X}_t - \mu_m^{(X)})\} \quad (9)$$

$\mathbf{E}_t^0$ denotes a bias component corresponding to the averaged speaker. Finally, The paired data of $[\mathbf{X}_t, \mathbf{E}_t^k]$ are utilized as the parallel data to train the DNN for the k-th eigenbase.

C. Adaptation for a new target speaker

Once the multiple DNNs are trained utilizing the parallel data prepared in the previous section, features of the source speaker can be converted to those of arbitrary target speakers by adapting the weights. Let $\mathrm{DNN}^{(k)}$ be the DNN that functions as converter to the k-th base of eigenspace. Then, converted features $f(\mathbf{X}_t)$ can be represented as below:

$$f(\mathbf{X}_t) = \sum_{k=1}^{K} \mathbf{w}_k^{(s)} \mathrm{DNN}^{(k)}(\mathbf{X}_t) + \mathrm{DNN}^{(0)}(\mathbf{X}_t), \quad (10)$$

where $\mathrm{DNN}^{(0)}$ converts the input feature to the bias feature. There are several approaches to determine the weight corresponding to the target. One is to use the weight estimated from EVGMM directly. Another is the weight is estimated in a supervised manner using features from the source and the target.

IV. Experiments

A. *Experimental conditions*

Experimental evaluations of one-to-many VC were carried out to investigate the effectiveness of the proposed architecture. A male speaker from ATR Japanese speech database B-set [11] was selected for a source speaker. 450 utterances were used for training, and 21 utterances included in neither training nor adaptation data were used for evaluation. As pre-stored speakers, we used 96 speakers including 48 male and 48 female speakers from a speech corpus called JNAS (the Japanese Newspaper Article Sentences) [12]. The utterances of each pre-stored speaker correspond to the 50 sentences of the 450 training sentences. In adaptation, namely, for target speakers, 10 speakers of 5 males and 5 females were used, and each of them uttered 32 sentences. We used 24-dimensional mel-cepstrum vectors for spectrum representation (D=24). These were derived by STRAIGHT analysis [13]. Aperiodic components, which are needed to generate mixed excitation in STRAIGHT, were not converted in this study, and they were fixed to -30dB at all the frequencies. The power coefficients and the fundamental frequencies were converted in a simple manner such that only the mean and the standard deviation were considered.

Three types of the methods were compared: conventional GMM, EVC, and the proposed approach using multiple DNNs. In the conventional GMM method, to achieve the best performance, the number of mixtures was varied from 1 to 256 and the optimal number was selected for each condition of the number of adaptation sentences.

The conventional GMM was trained in a supervised manner using adaptation data. In the EVC method, the number of mixtures was fixed to be 256. In the proposed method, each DNN which includes 5 layers with 256 units were constructed. In our method, rectified linear units were used as activation functions [15], and the DNNs were trained with dropout [14]. In both the EVC and the proposed method, the number of eigenspace bases was fixed to 96. This is equivalent to the number of pre-stored speakers.

The conversion performance was evaluated objectively using mel-cepstral distortion between the converted vectors and the vectors of the targets. Mel-cepstral distortion is denoted as follows,

$$\mathrm{MelCD[dB]} = \frac{10}{\ln 10}\sqrt{2\Sigma_{d=1}^{D}(mc_d - \bar{mc}_d)^2} \tag{11}$$

where mc_d, $\bar{mc}_d$ are the converted feature vectors and those from the target speaker. In the experiment, the number of target speakers' utterances to be used for adapting the models were varied, then the effectiveness of the methods was investigated.

B. *Objective evaluations of one-to-many VC*

In this evaluation the conversion performance for the 10 target speakers was evaluated. Fig. 1 shows the results of the four methods in terms of average mel-cepstral distortion for the test data as a function of the number of adaptation, or training sentences (the conversion to 10 target speakers).

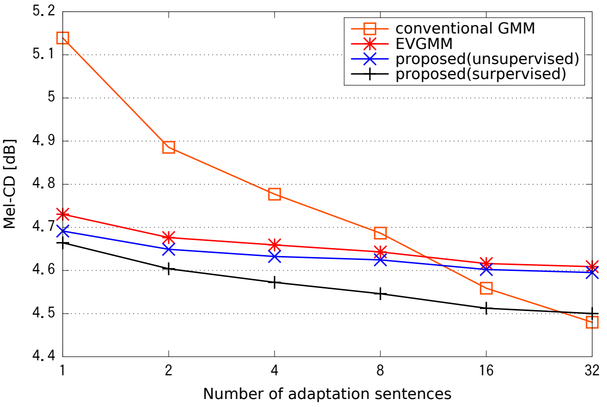

Fig. 1. Results of objective evaluations for 10 target speakers by mel-cepstral distortion (MCD).

In the proposed method, two conditions of weights were tested. One was to use the weight estimated from EVGMM directly ("proposed (unsupervised)" in Fig. 1). Another was to use the weight estimated by a supervised manner ("proposed (supervised)" in Fig. 1). The proposed methods and the "EVGMM" significantly outperform "conventional GMM" when using a small amount of adaptation data less than 8. This means that prior knowledge underlying the pre-stored data set is effectively utilized for improvement of the performance. Compared with "proposed (unsupervised)" and "EVGMM," our proposed method in all the conditions of adaptation outperforms the EVGMM method. Note that the two methods share the same target-dependent weights. This means that the proposed method works well to convert input feature to eigenbases.

On the other hand, compared with "proposed (supervised)" and "conventional GMM," "conventional GMM" outperforms "proposed (supervised)" when 32 adaptation data are available. It might be due to the low complexity of the proposed method. Although our proposed method updates only weights, conventional GMM updates other parameters such as mean vectors and covariance matrices. Then, to improve the conversion performance, it might be effective that target-dependent weights are extended to weight vectors when using a large amount of adaptation data.

C. *Subjective evaluations of one-to-many VC*

A listening test was carried out to evaluate the naturalness and the speaker individuality of converted speech. The test was conducted with subjects who are native Japanese with normal hearing. To evaluate the naturalness, a paired comparison was carried out. In this test, pairs of two different types of the converted samples were presented to the subjects, and then they judged which sample sounded more natural as native spoken Japanese. To evaluate speaker individuality, an RAB test was performed, where pairs of two different types of the samples were presented after presenting the reference sample of the target speech. In the tests, from 4 target speakers 2 male and 2 female, 5 sentences per a speaker were used. To reduce

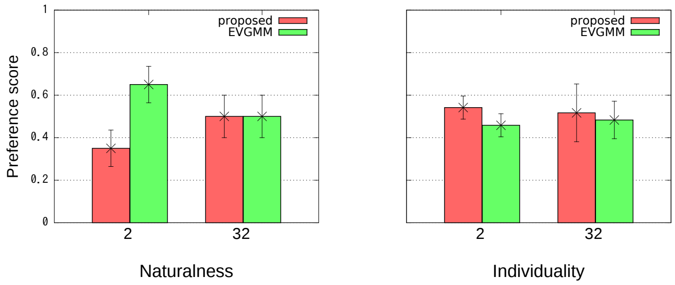

Fig. 2. Results of subjective evaluations between the proposed method and EVGMM. The number in x axis is the number of adaptation (or training) sentences.

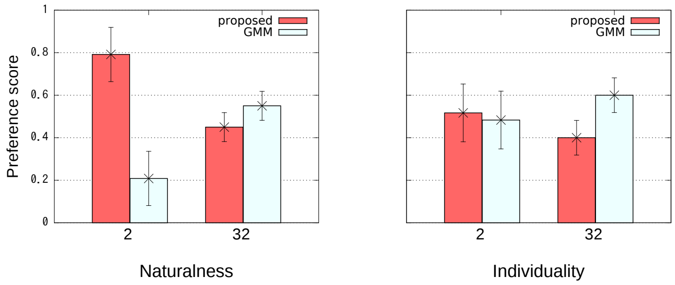

Fig. 3. Results of subjective evaluations between the proposed method and GMM. The number in x axis is the number of adaptation (or training) sentences.

the workload of the subjects, the pairs included the proposed method and those including the conventional methods were used in the tests. Then, the number of sample pairs evaluated by each subject was 80 in each test.

Fig. 2 and Fig. 3 shows the results of the three methods evaluated about the naturalness and individuality. In this tests, the number of adaptation, or training sentences were fixed to 2 or 32. When using 2 sentences, the "Proposed" outperforms "GMM" in both naturalness and speaker individuality. When using 32 adaptation data, "GMM" has the best performance in both naturalness and speaker individuality. Compared with the EVC method, the performance of the proposed method is comparable or slightly better to that of the EVC except in naturalness when using 2 adaptation data. Because the performance of the proposed method is comparable or slightly better to that of the EVC when using 32 adaptation data, the proposed method works well to convert to eigenbases. So this result might be caused by lowness of the weight estimation performance.

V. Conclusion

In this paper, we have proposed a new architecture for the arbitrary speaker conversion based on DNN that is inspired by EVC. The proposed network consists of multiple DNNs and each of them converts input features to features corresponding to a base of eigenspace. Training of these DNNs is achieved with the assistance of EVGMM. In the adaptation step, the weights of a new target speaker for summing up multiple outputs from the proposed network are estimated. Experimental evaluation demonstrates the effectiveness of the proposed method.

We are also planning to apply our method to many-to-many VC. In previous research[10], a DNN trained by using many-to-one parallel data shows some improvements of its generalization ability. Then, if DNNs to divide source features into eigenspace features are trained by using multi speakers such as pre-stored speakers, the proposed method can deal with open source speakers. In addition, if DNN's parameters corresponding to bases of eigenspace was fixed, target-dependent weights can be estimated by using target speaker's features as both the source and target. This estimation is carried out by an unsupervised manner such as an auto-encoder.

Acknowledgment

This work was supported by KAKENHI Grant-in-Aid for Young Scientists (B) (Number 25730105).

I wish to thank Dr. Yosuke Kashiwagi, Sony Corporation, for helpful discussions.

References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," ICASSP, vol. 1, pp. 285–288, 1998.

[2] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," ICASSP, pp. 301–304, 2001.

[3] D. Saito, H. Doi, N. Minematsu, and K. Hirose, "Application of matrix variate Gaussian mixuture model to statistical voice conversion," INTERSPEECH, pp. 2504–2508, 2014.

[4] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A.W. Black, and K. Prahallad,"Voice conversion using artificial neural networks," ICASSP, pp. 3893–3896, 2009.

[5] C.H. Lee, and C.H. Wu, "Map-Based Adaptation for Speech Conversion Using Adaptation Data Selection and Non-Parallel Training," INTERSPEECH, pp. 2254-2257, 2006.

[6] T. Toda, Y. Ohtani, and K. Shikano, "EigenVoice Conversion Based on Gaussian Mixture Model," INTERSPEECH, pp. 2446–2449, 2006.

[7] T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," IEICE TRANS. INF. &SYST., VOL.E93-D, NO.6, pp. 1589–1598, 2010

[8] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space," INTERSPEECH, pp. 653–656, 2011.

[9] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano "Non-parallel training for many-to-many eigenvoice conversion," ICASSP, pp. 4822–4825, 2010.

[10] L.J. Liu, L.H. Chen, Z.H. Ling, and L.R. Dai, "Spectral Conversion Using Deep Neural Netwotks Trained With Multi-Source Speakers," ICASSP, pp. 4849–4853, 2015.

[11] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speechdatabase as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357-363, 1990.

[12] "Jnas: Japanese newspaper article sentences," http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[13] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999.

[14] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[15] Vinod Nair, and Geoffrey E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010.