PAPER

Mapping Articulatory-Features to Vocal-Tract Parameters for Voice Conversion

Narpendyah Wisjnu ARIWARDHANI[†a)], ***Nonmember***, **Masashi KIMURA**[†], **Yurie IRIBE**[††], **Kouichi KATSURADA**[†], ***and*** **Tsuneo NITTA**[†,†††], ***Members***

SUMMARY In this paper, we propose voice conversion (VC) based on articulatory features (AF) to vocal-tract parameters (VTP) mapping. An artificial neural network (ANN) is applied to map AF to VTP and to convert a speaker's voice to a target-speaker's voice. The proposed system is not only text-independent VC, in which it does not need parallel utterances between source and target-speakers, but can also be used for an arbitrary source-speaker. This means that our approach does not require source-speaker data to build the VC model. We are also focusing on a small number of target-speaker training data. For comparison, a baseline system based on Gaussian mixture model (GMM) approach is conducted. The experimental results for a small number of training data show that the converted voice of our approach is intelligible and has speaker individuality of the target-speaker.

key words: *voice conversion, articulatory feature, neural network, arbitrary speaker*

1. Introduction

Voice conversion (VC) is one of the important technologies in the field of speech processing. VC transforms the voice from the source-speaker onto the target-speaker. When a source-speaker utters a certain sentence, the converted speech will sound as if a target-speaker is speaking the same sentence. There are several potential applications for VC, e.g., voice restoration in old documents/movies, dubbing television program, and speech-to-speech translation. Moreover, the result of VC can be applied to speech synthesizers in which we can expand the variety of speakers and make the synthesizer more flexible and cost-efficient.

One of the most widely used VC methods is the statistical parametric approach, Gaussian mixture model (GMM)-based algorithm [1]–[3]. While this Gaussian system is recognized as effective in individuality conversion, the speech quality of conventional GMM-based VC is not satisfactory, particularly in small number of training data. This might be owing to two main limitations of the conventional GMM-based VC, i.e., discontinuity and over smoothing. The first limitation comes from the fact that conventional GMM-based VC is conducted as a frame by frame operation, while the second limitation occurs because the system can only capture gross detail of the converted spectra. Therefore, most research on GMM-based VC were conducted to overcome these limitations, e.g., by combining dynamic features and incorporating global variance (GV) into the system. The newest improvement in this approach is the implementation of real-time GMM-based VC [4].

From a different perspective, another transformation paradigm was also conducted, namely frequency warping. This transformation function maps significant positions of the frequency axis (e.g., central frequency of formants) from the source-speaker to the target-speaker. As this method does not modify the fine spectral details of the source spectrum, it preserves very well the quality of the converted speech [5]. However, it is less accurate than that of GMM-based VC.

On the other hand, there exists other issues in typical VC systems, that is, they are text-dependent and need parallel training utterances of source and target-speaker. Because such parallel data may not always be feasible, there have been some approaches proposed in [6]–[9], which do not need parallel data. However, even though these text-independent VC approaches do not need parallel data, they still require speech data from source-speakers to build the VC model.

Regarding this issue, some research on VC application for arbitrary speakers have been proposed [10], [11]. These approaches do not require any speech data from a source-speaker in building the VC model, and hence can be used to transform an arbitrary speaker voice into a predefined target-speaker voice.

Another approach to solve this issue is introduced by mapping speaker-independent representation of a speech signal onto speaker-specific representation of a speech signal. The speaker-independent representation is expected to bring only linguistic information, while the speaker-specific representation is expected to bring both linguistic and speaker information. The study in [11] has an idea similar to our approach. It uses the lower order of linear prediction (LP) spectrum to capture the linguistic information of the signal, and mel-cepstrum (MCEP) to capture both the linguistic/message and speaker information. Meanwhile, we use articulatory features (AF) as the speaker-independent representation and vocal-tract parameter (VTP), represented by LPC coefficients, as the

Manuscript received August 19, 2013.
Manuscript revised November 17, 2013.

[†]The authors are with the Graduate School of Engineering, Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
[††]The author is with Information Science and Technology, Aichi Prefectural University, Nagakute-shi, 480–1198 Japan.
[†††]The author is with Green Computing Research Organization, Waseda University, Tokyo, 162–0042 Japan.
a) E-mail: ariwardhani@gmail.com

DOI: 10.1587/transinf.E97.D.911

speaker-specific representation [12].

While the previous works of VC use spectrum origin features that include various factors, such as speakers, phoneme contexts, ambient noise, etc., our proposed VC is based on the sparse representation of articulatory features. This also underlines our different perspective of addressing VC problems from previous papers. We also do not need manual efforts to carefully prepare training data.

In this paper, we not only avoid the training process for source speaker, but also focus on making VC application with a small number of target-speaker training data. For this purpose, speaker adaptation technique was conducted. Because this approach requires a small number of target-speaker training data, the proposed VC process is expected to be more user-friendly.

This paper is organized as follows. First, an outline of GMM-based VC is reviewed in Sect. 2. Next, the proposed approach articulatory feature-artificial neural networks (AF-ANNs)-based VC is given in Sect. 3. An evaluation of VC system, in which we present and discuss experimental results, is explained in Sect. 4. Finally, we summarize our findings in Sect. 5.

2. GMM-Based Voice Conversion

The outline of a GMM-based VC system, comprising training and testing module, is shown in Fig. 1. VC can be defined as mapping the source feature vector $\boldsymbol{x}_t$ into the target feature vector $\boldsymbol{y}_t$, at each time t. At the training module, acoustic feature vectors from both the source and target speakers are extracted and aligned by dynamic time warping (DTW). The source vectors are augmented with the corresponding target features as $\boldsymbol{z}_y = [\boldsymbol{x}_t^T \boldsymbol{y}_y^T]^T$ and the GMM is estimated for the augmented vectors.

The means and covariances of the GMM of the augmented vectors are given as

$$\boldsymbol{\mu}_n^z = \begin{bmatrix} \boldsymbol{\mu}_n^x \\ \boldsymbol{\mu}_n^y \end{bmatrix} \tag{1}$$

$$\boldsymbol{\Sigma}_n^z = \begin{bmatrix} \boldsymbol{\Sigma}_n^{xx} & \boldsymbol{\Sigma}_n^{xy} \\ \boldsymbol{\Sigma}_n^{yx} & \boldsymbol{\Sigma}_n^{yy} \end{bmatrix} \tag{2}$$

where vectors $\boldsymbol{\mu}_n^x$ and $\boldsymbol{\mu}_n^y$ denote the mean of the source and target entries of the augmented vector in Gaussian n, respectively, and the superscripts of the covariance matrices denote their respective covariances and cross-covariances. In the conversion, for M-component Gaussian mixture model, the mapped target vector $\hat{\mathbf{y}}_t$ is formed from the source vector $\boldsymbol{x}_t$ as

$$\hat{\mathbf{y}}_t = \sum_{n=1}^{M} \omega_{n,t} \left[\boldsymbol{\mu}_n^y + \boldsymbol{\Sigma}_n^{yx} (\boldsymbol{\Sigma}_n^{xx})^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_n^x) \right] \tag{3}$$

where $\omega_{n,t}$ is the posterior probability that the n-th Gaussian has produced the t-th observation, calculated using the source vector $\boldsymbol{x}_t$ and mean $\boldsymbol{\mu}_n^x$ and covariance $\boldsymbol{\Sigma}_n^{xx}$ as

$$\omega_{n,t} = \frac{\propto_n N(\boldsymbol{x}_t; \boldsymbol{\mu}_n^x, \boldsymbol{\Sigma}_n^{xx})}{\sum_{m=1}^{M} \propto_m N(\boldsymbol{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^{xx})} \tag{4}$$

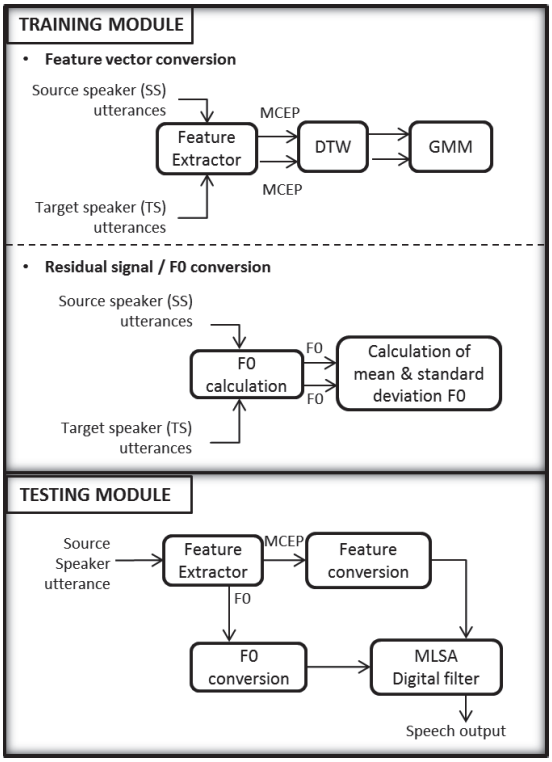

Fig. 1 Outline of GMM-based VC.

The joint density mapping Eq. (4) is the maximum-likelihood estimate of the target vectors given the source vectors. In this paper, we have conducted GMM-based VC experiments on the VC setup built in FestVox distribution [13]. This VC setup is based on the study in [14], and supports the conversion considering the correlation between frames (MLPG) and GV of spectral trajectory.

3. AF Based Voice Conversion

3.1 Articulatory Feature Representation and Vocal Tract Parameter

Our approach maps speech signal onto speaker-independent representation of an AF sequence first, then the AF is converted to speaker-specific representation of a speech signal. Because the AF sequence is expected to bring only linguistic information, source-speaker training data is not required during the training process. In our proposed approach, we use AM as the speaker-independent representation and VTP as the speaker-specific representation.

3.1.1 AF Representation

AF describes the articulatory manners and places in human speech production at given time t, and is combined with its preceding and following time. In our system, this AF sequence is represented by three time frames of a current frame, previous frame (t – 3), and following frame (t + 3).

To generate AF from the speech signal, two stages of

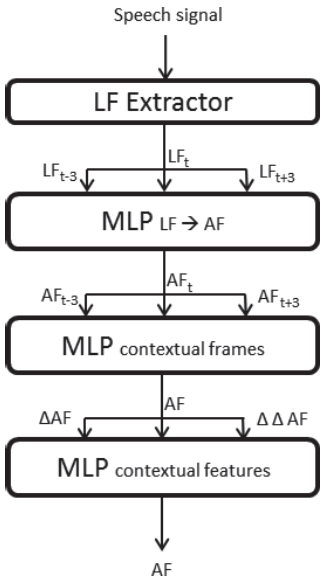

Fig. 2 Four stages of an AF extractor.

signal processing are needed (Fig. 2). The first stage employs the local feature (LF) extractor [15]. Speech signal is sampled at 16 kHz and framed using a 25-ms Hamming window for every 10 ms. Subsequently, a 512-point FFT is applied. Power and delta power is calculated from the resultant FFT power spectrum. Moreover, a 24-ch band pass filter (BPF) with mel-scaled center frequencies is applied to the resultant FFT. The BPF output undergoes three-point linear regression along the time and frequency axes [15]–[17]. Subsequently, discrete cosine transform (DCT) is applied to the output of linear regression. Then, with the delta power been previously calculated, a 25-dimension LF is generated. LFs are acoustic features that represent variation in a spectrum pattern along time and frequency axes. We use LFs for the input of multi-layer perceptron (MLP), because our previous study showed that LFs provides better performance than MFCC as input to this MLP [17].

The second stage of AF extractor comprises three MLPs. The first MLP requires a 75 dimension LF as input and generates 45-dimension discrete AF. The second MLP reduces misclassification at phoneme boundaries by constraining the AF context. It requires 135-dimension AF and its contextual frames as input, and generates a 45-dimension AF. The third MLP uses delta and delta-delta AF as input and generates a 45-dimension final AF.

3.1.2 VTP

VTPs are represented by partial autocorrelation (PARCOR) parameters, associated with linear predictive coding (LPC) coefficients. LPC based vocoders are designed to emulate the human speech production mechanism. In the LPC analysis, the short-term correlations between speech samples (formants) are modeled and removed by an LPC digital filter. These LPC coefficients describe the transfer function of human vocal tract on the excitation signal. The excitation

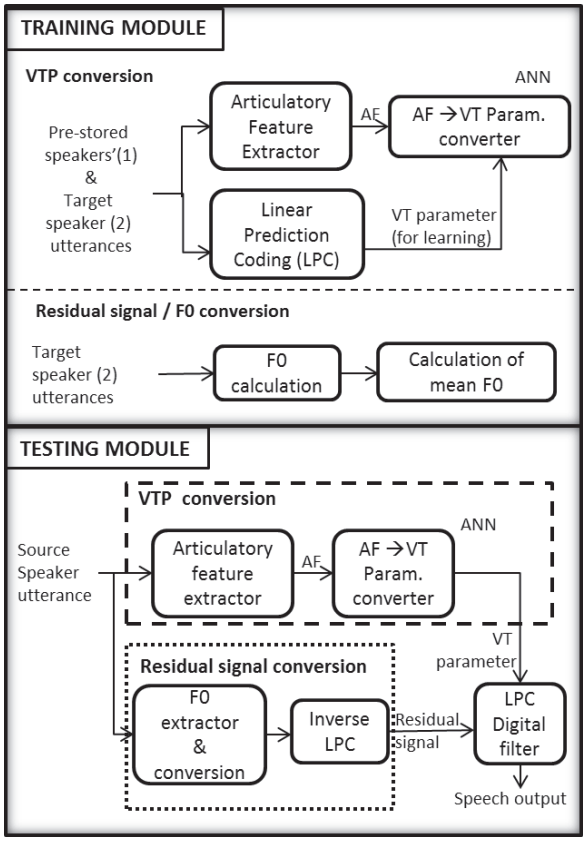

Fig. 3 Training and testing modules of proposed VC system.

signal models the glottal pulses and turbulent air flow at the glottis. This excitation signal, called as LPC residual signal, can be calculated by filtering speech signal with the inverse transfer function from LPC estimation. As the vocal tract shape differs from person to person, LPC parameters are also speaker dependent. Hence, LPC parameters possess the characteristic of a specific speaker. To allow an easy filter stability check, the LPC coefficients are transformed into PARCOR parameters [18]. The LPC digital filter is stable if PARCOR parameters have a magnitude of less than unity. It means that their values range from -1 to $+1$ so that they do not need to be amplitude normalized as the input of ANN. As we solve the LPC estimation using the Levinson-Durbin algorithm. This algorithm guarantees that PARCOR coefficients are bounded by ± 1 [19].

3.2 VTP and Residual Signal Conversion

The VC system consists of a training module and a testing module. The training module can be divided into VTP conversion and residual signal/F0 conversion, while the testing module can be divided into VTP conversion and residual signal conversion (Fig. 3).

3.2.1 VTP Conversion

The mapping of AF to VTP is conducted using an ANN

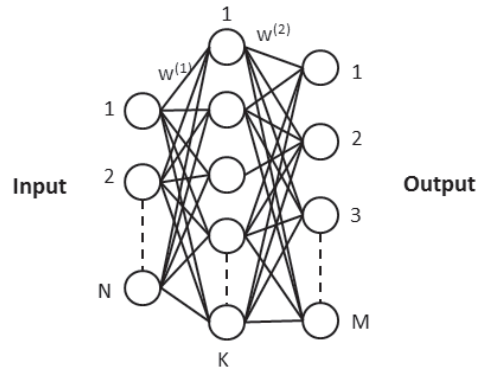

Fig. 4 Architecture of a three layered ANN with N input nodes, M output nodes, and K nodes in the hidden layers.

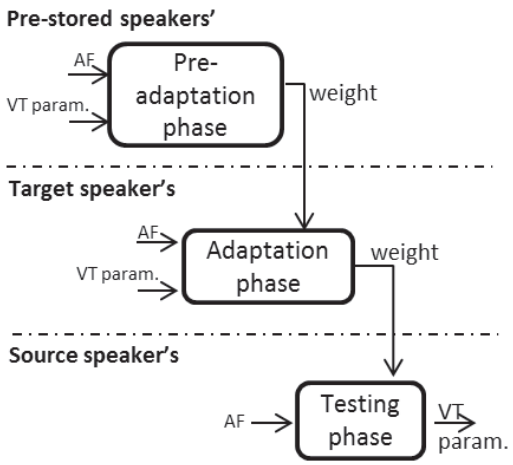

Fig. 5 Adapting ANN with target-speaker's voice.

model. The ANN consists of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnections among nodes have weights associated with them.

A multi-layer feed forward neural network with one or two hidden layers is used in the experiment. The ANN is trained to map AF onto the target speaker VTP. The back-propagation learning law is used to adjust the weights of the neural network to get the minimum mean squared error between the desired and the actual output values. Figure 4 shows the ANN architecture used to obtain the transformation function to map speaker-independent AF onto target speaker VTP. The adjusted weight on every interconnection among nodes represents the mapping function between speaker-independent AF and target speaker VTP.

As can be seen in Fig. 5, there are three phases in the AF to VTP converter neural network, pre-adaptation, adaptation, and testing. This adaptation technique enables VTP to use only a small number of target-speaker training data. While training phase requires a large amount of utterances from pre-stored voices, adaptation phase requires only several utterances from the target-speaker. In the testing phase, one utterance of an arbitrary source-speaker can be input to produce the converted VTP, which later will be synthesized

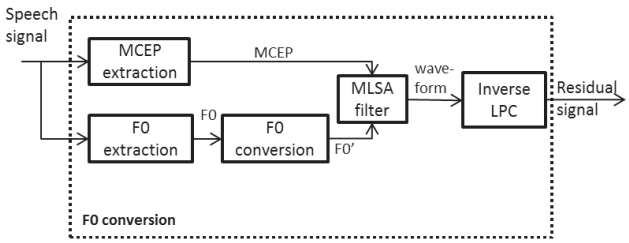

Fig. 6 Residual signal conversion module.

into converted speech. After AF is converted into target-speaker VTPs, then with the residual signal, it will be resynthesized using the LPC digital filter.

3.2.2 Residual Signal Conversion

As explained in Sect. 3.2, the excitation signal is represented by the LPC residual signal. The residual signal has speaker individuality, especially in terms of fundamental frequency (F0). Therefore, in the testing phase, it is important to manipulate source-speaker residual signal so that the converted speech will have similar F0 contours with the target-speaker fundamental frequency (F0). Therefore, in the testing phase, it is important to manipulate source-speaker residual signal so that the converted speech will have similar F0 contours with the target-speaker.

In this paper, we use the traditional approach of F0 transformation, as used in a GMM-based transformation. Figure 6 describes the detail of residual signal conversion module depicted in Fig. 3. Subsequent to F0 extraction, a logarithmic Gaussian transformation is used to transform the F0 of a source-speaker to that of a target-speaker, as indicated in the following equation:

$$\log(F_{0\,conv}) = \mu_{target} + \frac{\sigma_{target}}{\sigma_{source}}(\log(F_{0\,source}) - \mu_{source}) \quad (5)$$

where μ_{source} and σ_{source} are the mean and variance, respectively, of the F0 in logarithmic domain for the source-speaker, μ_{target} and σ_{target} are the mean and variance, respectively, of the F0 in logarithmic domain for the target-speaker, $F_{0\,source}$ is the F0 of the source-speaker and $F_{0\,conv}$ is the converted F0. Because our system uses an LPC digital filter, the converted F0 has to be processed into LPC residual signal before it can be resynthesized with the converted VTP into speech output.

4. Evaluation of VC

4.1 Speech Database

Speech data used in the experiment is sampled with 16 kHz. We used three speech databases for three phases of AF to VTP conversion, i.e., pre-stored speakers for the pre-adaptation phase, target-speakers for the adaptation phase, and source-speakers for the testing phase. These three speech databases utter different utterances (Table 1).

Table 1 Speech database for VC.

Type of speakers	Database	Number of speakers	Utterances /speaker
Pre-stored speakers	ATR PB	6 (male)	50
Target speaker	Labmate1	2 (male)	20
Source speaker	Labmate2	3 (male)	5

When people have to differentiate or identify some speakers, they will find it easier if they already know the speakers. Therefore, because we aim to have subjective evaluation respondents from our lab member and surroundings, we recorded "Labmate database", instead of using the existing database. There were five persons for the overall Labmate database, three persons (END, NIS, and IRI) as source-speakers, and two persons (KZH and SUG) as target-speakers. In total, there were six pairs of speakers available from the Labmate database.

The same database (source and target-speakers) is also used for GMM-based VC experiments. For comparison, we also asked target-speakers to utter the same sentences as those in Labmate2. However, this recording will be used only for subjective evaluation and for calculating spectral distortion (SD) during the objective evaluation.

4.2 Experimental Setup

For our proposed approach, we use a 45-dimension AF vector, comprising a 15-dimension preceding context, 15 dimensions of current frame, and 15-dimension following context of AF patterns for each input frame as AF representation. Moreover, several orders of LPC analysis were conducted to produce PARCOR parameters as VTP. For the feature vectors of GMM-based VC, we use MCEP extracted using FestVox distribution [15].

Two evaluations are performed, objective and subjective. For objective evaluations, spectrum distortion (SD) is calculated on the speech segment (excluding silence part) to measure the distance between target-speaker spectrum and converted spectrum. We use this measure to check the performance of mapping obtained by an ANN or a GMM model. SD is computed as follows:

$$SD = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(W_{\text{conv}} - W_{\text{target}})^2} \quad (6)$$

where L is the number of frames, K is the number of frequency bins, and W_{conv} and W_{target} are the log amplitude of converted and target-speaker spectra, respectively.

For subjective evaluations, three tests are conducted, similarity test, XAB test, and MOS test, with nine listeners. In the similarity test, we present the listeners with the source-speaker utterance, target-speaker utterance, and each converted utterance from AF-ANN and MCEP-GMM models. The listeners would be asked to provide a score indicating how similar the converted speech with either the source-speaker or target-speaker. The range of similarity score is from 1 to 5, where a score of 1 indicates that the converted speech sounds very similar to source-speaker and score 5 indicates that the converted speech sounds very similar to the target-speaker.

For the XAB test, we present the listeners with X, a natural utterance of the target-speaker, to be compared against an AF-ANN converted speech and an MCEP-GMM converted speech. To ensure that the listener is not biased, we shuffle the position of the AF-ANN/MCEP-GMM converted speech, i.e., A and B, with X always given at the beginning of the test. The listeners would be asked to select what they perceive to be closer, A or B, to the target utterance X. The last subjective test is MOS test where listeners evaluate the speech quality of the converted voices using a 5-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent).

4.3 Objective Evaluation of ANN-Based VC System

LPC analysis is dependent upon its filter order, i.e., the number of LPC coefficients. The order of LPC filter is typically estimated by starting with a heuristic value according to the sampling frequency. This heuristic value is equal to the sampling rate in kHz, with 4 or 5 additional coefficients [20]. Since our speech data is sampled with 16 kHz, a 20-order of LPC analysis is chosen for VTP. We aim to investigate the effect of ANN architecture and different VTP orders on the performance of AF-ANN based VC. Six ANN architectures are compared, all with 45 nodes in the input layer, representing a 45-dimension AF.

The first architecture uses only one hidden layer and x output layers, where the value of x represents the number of LPC order to generate VTP. For example, for the VTP 40, the ANN architecture would be 45 input nodes, 450 hidden layer nodes, and 20 output nodes. From the second to the sixth architecture, we considered augmenting VTP with contextual frames, i.e., appending VTP from previous and next frames to the current frame of VTP. Hence, the number of output nodes is three times that of the VTP order, i.e., 60 output nodes for VTP 20, 120 output nodes for VTP 40, and 180 output nodes for VTP 60. In this paper, we investigate three-layer and four-layer ANNs, i.e., one input layer (IL), one or two hidden layers (HL), and one output layer (OL).

From the second to the sixth architecture, we considered augmented VTP, i.e., appending VTP from previous and next frames to the current frame of VTP. Hence, the number of output nodes was three times that of the VTP order, i.e., 60 output nodes for VTP 20, 120 output nodes for VTP 40, and 180 output nodes for VTP 60. In this paper, we experimented with three-layer and four-layer ANNs, i.e., one input layer (IL), one or two hidden layers (HL), and one output layer (OL).

Table 2 provides SD scores of END-KZH for three VTP orders and six ANN–model architectures. From this table, we see that three-layered architecture 45(IL) 450(HL) 3x(OL) for VTP 20 provides a better result when compared with other architectures. We also confirmed this result by listening to the resultant speech. Hence, for the remaining experiments reported in this paper, the three-layered

Table 2 SD obtained on one-utterance END-KZH for different architectures of an ANN model.

No	ANN architecture	SD (dB)		
		VTP 20	VTP 40	VTP 60
1	45(IL) 450(HL) x(OL)	7.944	7.534	7.198
2	45(IL) 450(HL) 3x(OL)	**6.925**	7.394	7.297
3	45(IL) 3x(HL) 3x(OL)	7.22	7.231	7.382
4	45(IL) 6x(HL) 3x(OL)	7.727	7.216	7.425
5	45(IL) 45(HL) 3x(HL) 3x(OL)	8.107	7.877	7.971
6	45(IL) 90(HL) 6x(HL) 3x(OL)	8.107	7.604	7.425

Table 3 Averaged SD obtained for six pairs-of-speakers.

	ANN	GMM
SD (dB)	10.193	12.628

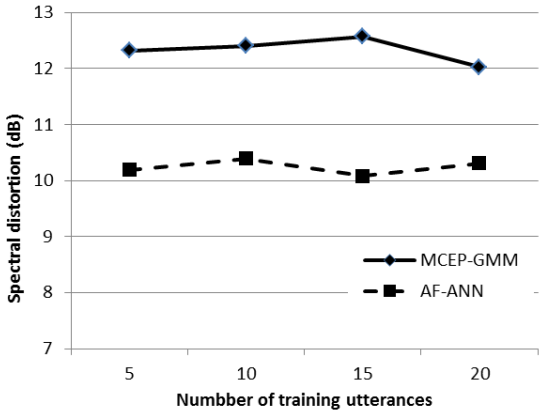

Fig. 7 SD scores of VC based on AF-ANN and MCEP-GMM for six pairs of speakers.

architecture 45(IL) 450(HL) 60(OL) is used. The overall SD scores for six pairs of speakers of both AF-ANN and MCEP-GMM-based VC are shown in Table 3, which indicate that the AF-ANN-based VC is comparable with MCEP-GMM-based VC. From the objective evaluation, SD of GMM-based VC has more than 1 dB difference than that of AF-ANN-based approach.

We conducted the first training of AF to VTP converter using 6 sets of phonetically balanced database. In this step, AF to VTP converter learns to convert any phoneme, represented by AF, into VTP. Subsequently, the adaptation phase is conducted with small number of adaptation data. Based on the analysis in [22], the nasal sounds (e.g., N, n, m, ny, my) and the vowel part has relatively high correlations with the perception (speaker identity). Therefore, in our approach, we can conduct adaptation phase with a small number of target-speaker training data, in which this adaptation data contains nasal sounds and all the needed vowels.

To determine the effect of the number of training utterances for the VC models, we performed the experiments by varying the target-speaker training data from 5 to 20 utterances. Please note that our AF-ANN approach also needed pre-stored data (non-parallel with the target-speaker utterances), while MCEP-GMM approach needed parallel training utterances of source and target-speakers. GMM-based VC performance is expected to improve as the number of training utterances increases [11]. However, since we are focusing in building VC for a small number of target-speaker training data, the experiments were conducted until 20 training utterances. From Fig. 7, we observe that as the number of training utterances increase, the SD scores obtained by MCEP-GMM decreased, especially for 20 parallel training utterances. For AF-ANN, the SD scores seem to be more stable and even have the lowest value for 15 training utterances.

A simple assessment on the computational cost is conducted by measuring the program completion time on both of the approaches. The computational complexity of the proposed method is compared with baseline GMM-based VC. For this computational cost, the program completion time for learning 20 utterances of target-speaker training data is measured on a personal computer (PC) equipped with a quad-core 2.80 GHz CPU and 2G memory. The operating system of the PC is based on Windows 7. The results of the computational complexity tests indicate that the computational time of the proposed method is 5 times slower when compared with GMM-based VC. Note that the comparison only measures the adaptation phase of our approach. Our approach also needs a large number of pre-stored speakers training data for the first ANN training (before adaptation). Moreover, the F0 conversion module during the testing phase is also need to be simplified.

4.4 Subjective Evaluation of GMM and ANN-Based VC System

In voice conversion, typical objective evaluation is done by comparing the converted speech to the ideal target speaker utterance. However, due to inter-speaker variability, a speaker can utter the same utterance in various ways. Therefore, objective measures do not always support subjective evaluations [23]. Currently the most accurate method for evaluating speech quality is through subjective listening tests [24]. Thus, subjective evaluation is needed to confirm the result of objective evaluation.

In this section, we provide subjective evaluation results for AF-ANN and MCEP-GMM-based VC systems. We conducted similarity, XAB, and MOS tests to evaluate the performance of the AF-ANN-based transformation against the MCEP-GMM-based transformation. A total of 9 respondents were asked to participate in the experiments. Figure 8 provides the similarity, XAB, and MOS scores for six pairs of speakers (END-KZH, NIS-KZH, IRI-KZH, END-SUG, NIS-SUG, and IRI-SUG). The testing is done on the test set of 30 utterances. The overall similarity scores indicate that for AF-ANN based VC, the respondents perceived that the converted speech is more similar to the target-speaker than to the source-speaker. The XAB scores indicate that compared with the MCEP-GMM-based VC system, the AF-ANN-based VC system performs better for a small number of target-speaker training data. MOS test is also performed to confirm that the resulting speech of AF-ANN based VC system is intelligible.

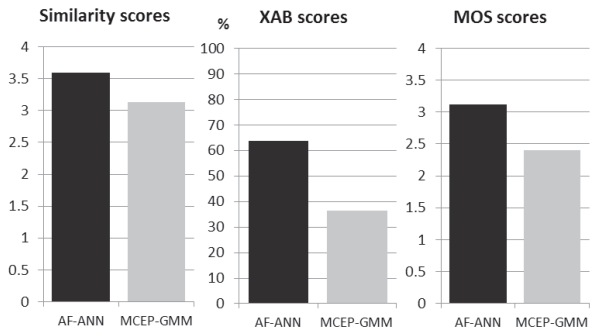

Fig. 8 Similarity, XAB, and MOS scores of VC based on AF-ANN and MCEP-GMM for six pairs of speakers.

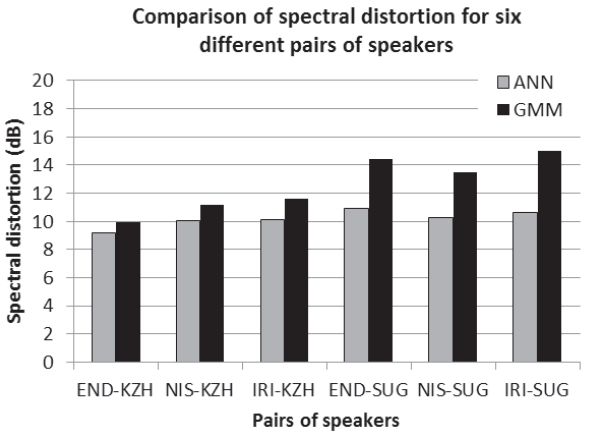

Fig. 9 SD scores of VC based on AF-ANN and MCEP-GMM over six pairs of speakers.

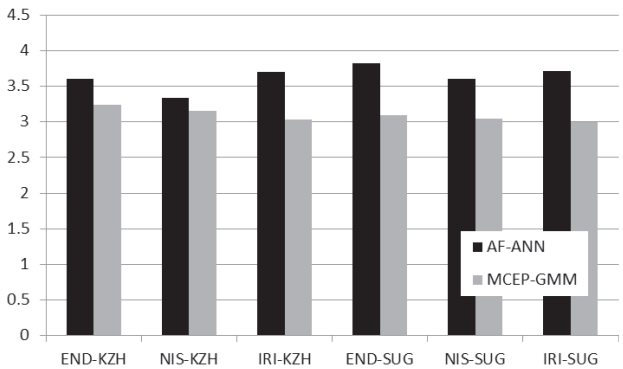

Fig. 10 Similarity scores of VC based on AF-ANN and MCEP-GMM over six different-pairs of speakers.

4.5 Experiments on Multiple Speaker Pairs

To show that the ANN-based transformation can be generalized over different databases, we conducted objective and subjective evaluations for six pairs of speakers. Figure 9 shows SD scores of AF-ANN and MCEP-GMM based VC systems for six pairs of speakers. This figure shows that for all pairs of speakers, SD scores of AF-ANN-based VC are lower than those of MCEP-GMM-based VC system.

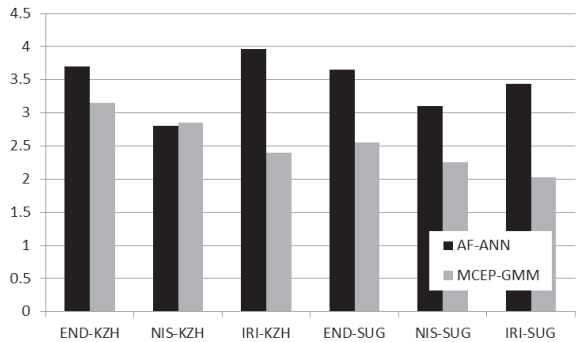

Fig. 11 MOS scores of VC based on AF-ANN and MCEP-GMM over six different-pairs of speakers.

Moreover, Fig. 10 and Fig. 11 show similarity and MOS scores of AF-ANN and MCEP-GMM-based VC systems for different pairs of speakers. While for MOS scores, AF-ANN-based VC system outperforms MCEP-GMM-based VC system in most cases, for similarity scores, AF-ANN-based VC system always outperforms MCEP-GMM-based VC system.

5. Conclusions

We have proposed an articulatory based VC that does not require speech data from source-speakers, and hence can be considered as independent of source-speaker. The experimental results of subjective evaluation tests in VC show that the converted voice is intelligible and has speaker individuality of the target-speaker. For the overall performance, AF-ANN-based VC outperforms MCEP-GMM-based VC for a small number of target-speaker training data. Future studies will be conducted in the AF extractor domain for cross-lingual VC.

Acknowledgements

This work is supported by a Grant-in-Aid for Young Scientists (B) 24700167 2012 and for Scientific Research (B) 22300060 2012 from MEXT, Japan, the Strategic Information and Communications R&D Promotion Programme by MIC, Japan, and the Kayamori Foundation of Information Science Advancement.

References

[1] A. Kain and M.W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction," Proc. ICASSP 2001, vol.2, pp.813–816, Salt Lake City, Utah, USA, 2001.
[2] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," Proc. Eurospeech, pp.447–450, Madrid, Spain, 1995.
[3] T. Toda, A.W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," Proc. ICSLP, 2004, pp.1129–1132, Jeju, South Korea, Oct. 2004.
[4] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," Proc. Interspeech 2012, Portland, USA, 2012.

[5] H. Valbret, E. Moulines, and J.P. Tubach, "Voice transformation using PSOLA technique," Speech Commun., vol.11, pp.175–187, June 1992.
[6] D. Sundermann, A. Bonafonte, H. Hoge, and H. Ney, "Voice conversion using exclusively unaligned training data," Proc. ACL/SEPLN, 42nd Annual Meeting Association for Computational Linguistics, pp.41–48, Barcelona, Spain, July 2004.
[7] D. Sundermann, H. Ney, and H. Hoge, "VTLN based cross-language conversion," Proc. 8th IEEE Automatic Speech Recognition and Understanding (ASRU), pp.676–681, Virgin Islands, USA, Dec. 2003.
[8] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and S. Narayanan, "Text-independent voice conversion based on unit selection," Proc. IEEE ICASSP 2006, Toulouse, France, pp.81–84, May 2006.
[9] A. Mouchtaris, J.V. Spiegal, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. Audio, Speech, and Language Processing, vol.14, no.3, pp.952–963, May 2006.
[10] H. Ye and S. Young, "Voice conversion for unknown speakers," Proc. ICSLP 2004, pp.1161–1164, Jeju, South Korea, 2004.
[11] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," IEEE Trans. Audio, Speech, and Language Processing, vol.18, no.5, pp.954–964, July 2010.
[12] T. Nitta, T. Onoda, M. Kimura, Y. Iribe, and K. Katsurada, "One-model speech recognition and synthesis based on articulatory movement HMMs," Proc. Interspeech 2010, pp.2970–2973, Makuhari, Japan, 2010.
[13] A.W. Black and K. Lenzo, "Building voices in the festival speech synthesis system," Language Technologies Institute, Carnegie Mellon University, http://festvox.org/bsv, accessed 24 March 2013.
[14] T. Toda, A.W. Black, and K. Tokuda, "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, Nov. 2007.
[15] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. ICASSP 1999, Phoenix, Arizona, USA, pp.421–424, 1999.
[16] M.N. Huda, H. Kawashima, and T. Nitta, "Distinctive phonetic feature (DPF) extraction based on MLNs and inhibition/enhancement network," IEICE Trans. Inf. & Syst., vol.E92-D, no.4, pp.671–680, April 2009.
[17] T. Fukuda, W. Yamamoto, and T. Nitta, "Distinctive phonetic feature extraction for robust speech recognition," Proc. ICASSP 2003, Hong Kong, Hong Kong, pp.25–28, 2003.
[18] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood," Proc. 6th International Congress on Acoustics, Los Alamitos, California, USA, pp.C-17–C-20, 1968.
[19] L.R. Rabiner and R.W. Schafer, "Introduction to digital speech processing," Foundations and Trends in Signal Processing, vol.1 nos.1-2, pp.1–194, 2007.
[20] J.D. Markel and A.H. Gray, Linear Prediction of Speech, Springer, Verlag, Berlin, 1976.
[21] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," Proc. Interspeech 2008, Brisbane, Australia, pp.1453–1456, 2008.
[22] K. Amino, T. Sugawara, and T. Arai, "Speaker similarities in human perception and their spectral properties," Proc. WESPAC IX, Seoul, 2006.
[23] H. Benisty, D. Malah, and K. Crammer, "Modular global variance enhancement for voice conversion systems," Proc. EUSIPCO 2012, pp.370–374, Bucharest, Romania, Aug. 2012.
[24] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, and Language Processing, vol.16, no.1, pp.229–238, Jan. 2008.

Narpendyah Wisjnu Ariwardhani received her B. Eng. degree and M. Eng. degree from Bandung Institute of Technology, Bandung, Indonesia. She is currently a doctoral student at the Graduate School of Engineering, Information and Electronic Engineering, Toyohashi University of Technology, since 2010. Her research interest includes speech recognition. She is a member of ASJ.

Masashi Kimura received his B. Eng. degree and M. Eng. degree from Toyohashi University of Technology, Japan. He has finished his coursework and is in the final year of his doctoral study, in-waiting for the oral examination from Graduate School of Engineering, Information and Electronic Engineering, Toyohashi University of Technology. Currently he works as a speech recognition engineer at ATR-Trek Inc., Japan. His research interests include speech synthesis, speech recognition and intelligent agent. He is a member of JSAI, IEICE, ASJ and IPSJ.

Yurie Iribe received her M.S. and Ph.D. degrees from the Graduate School of Human Informatics of Nagoya University, Japan. She is currently an Assistant Professor at School of Information Science and Technology, Aichi Prefectural University, Japan. Her recent research interests include education support and speech recognition. She is a member of ISCA, IEICE, IPSJ, JSAI, and ASJ.

Kouichi Katsurada received his Ph.D. degree from Osaka University in 2000. He has been with Toyohashi University of Technology since 2000. He is currently an Associate Professor at the Center for International Relations and Department of Computer Science and Engineering, Toyohashi University of Technology. His research interests include multimodal interaction, facial image processing, and spoken term detection. He is a member of IEEE, ISCA, IEICE, IPSJ, JSAI, and ASJ.

Tsuneo Nitta received his Ph.D. from Tohoku University, Japan. He had worked at R&D Center and Multimedia Eng. Lab. of Toshiba Corp. from 1970 to 1998. He subsequently joined Toyohashi University of Technology as a Professor in the Graduate School of Eng. In 2012, he became a TUT Professor Emeritus and a visiting professor both at TUT and Waseda University. His current research interest includes speech recognition, speech synthesis, and multi-modal interaction. He is an IPSJ Fellow and a member of IEICE, ASJ, JSAI, and IEEE.