Unsupervised Speaker Adaptation for DNN-based Speech Synthesis using Input Codes

Shinji Takaki*, Yoshikazu Nishimura†, Junichi Yamagishi*
* National Institute of Informatics, Tokyo, Japan
† alt Inc., Tokyo, Japan

***Abstract*—A new speaker-adaptation technique for deep neural network (DNN)-based speech synthesis – which requires only speech data without orthographic transcriptions – is proposed. This technique is based on a DNN-based speech-synthesis model that takes speaker, gender, and age into consideration as additional inputs and outputs acoustic parameters of corresponding voices from text in order to construct a multi-speaker model and perform speaker adaptation. It uses a new input code that represents acoustic similarity to each of the training speakers in a probability. The new input code, called "speaker-similarity vector," is obtained by concatenating posterior probabilities calculated from each model of the training speakers. GMM-UBM or i-vector/PLDA, which are widely used in text-independent speaker verification, are used to represent the speaker models, since they can be used without text information. Text and the speaker-similarity vectors of the training speakers are used as input to first train a multi-speaker speech-synthesis model, which outputs acoustic parameters of the training speakers. A new speaker-similarity vector is then estimated by using a small amount of speech data uttered by an unknown target speaker on the basis of the separately trained speaker models. It is expected that inputting the estimated speaker-similarity vector into the multi-speaker speech-synthesis model can generate synthetic speech that resembles the target speaker's voice. In objective and subjective experiments, adaptation performance of the proposed technique was evaluated using not only studio-quality adaptation data but also low-quality (i.e., noisy and reverberant) data. The results of the experiments indicate that the proposed technique makes it possible to rapidly construct a voice for the target speaker in DNN-based speech synthesis.**

I. Introduction

The flexibility and controllability of speech-synthesis systems are as important as naturalness of speech in some applications; hence, constructing such a flexible speech-synthesis system is an interesting research topic in the field of DNN-based speech synthesis. A variety of multi-speaker modeling and speaker-adaptation techniques for DNN-based speech synthesis have been proposed recently. Multi-speaker modeling is a technique for synthesizing voices of various speakers by using a common model, and speaker adaptation is a technique for estimating a new acoustic model by using a small amount of speech data uttered by a new target speaker or in a new speaking style (e.g., a different emotion). To give a few examples of the multi-speaker modeling in the field of DNN-based speech synthesis, using speaker codes that represent a speaker's identity for multi-speaker modeling, in which additional inputs are used to distinguish speakers, has been proposed [1], [2], [3]. A speaker-adaptation technique using i-vectors as an additional input, one using an adaptation method for speech recognition called "learning hidden-unit contributions" [4], one using linear transforms defined by Gaussian mixture models (GMMs), and combinations of those methods, was proposed [5]. In another study, it was assumed that the output layer in a DNN captures most speaker differences, and under that assumption, it was attempted to estimate a speaker-dependent output layer by using individual speaker's data while keeping the hidden network layers shared across all speakers [6].

Prior to the present study, a DNN-based acoustic model using auxiliary features referred to as input codes was proposed [7]. In that model, to more-effectively retain speaker voice characteristics and allow speaker adaptation, a speaker's identity, gender, and age classes were additionally used. Speaker adaptation was performed by estimating a new speaker code based on back-propagation (BP) using a small amount of the target speaker's speech data and associated linguistic features obtained from text. Almost all other adaptation techniques proposed for DNN synthesis are based on BP [1], [5], [6], [8], [9]; hence, not only speech data but also linguistic features are always required.

In this study, a new speaker-adaptation technique for DNN-based speech synthesis – which requires only speech data without orthographic transcriptions – is proposed. This technique is traditionally called unsupervised speaker adaptation for speech synthesis [10]. A naive dirty way is to obtain transcriptions using external automatic speech recognition and use conventional speaker adaptation based on speech and automatically generated transcriptions [11]. However, this procedure may result in issues when the outputs of speech recognition have severe errors.

The proposed technique uses a new input code designed for speaker adaptation without using text. The new input code, called "speaker-similarity vector", represents acoustic similarity of a target speaker to each of several training speakers in terms of probability, and it is obtained by concatenating posterior probabilities calculated from each of the models of the training speakers. Intuitively, this process may be viewed as replacing a conventional binary hard speaker code with continuous soft codes according to speaker similarity. Therefore, if a multi-speaker speech-synthesis model is trained using text, and speaker-similarity vectors are used as input, it can be expected that speaker characteristics of synthetic speech generated from the trained multi-speaker model will depend on the speaker-similarity vectors, and the synthetic speech will

vary if the speaker similarity vectors change. Furthermore, if a new speaker-similarity vector is estimated by using a small amount of speech data uttered by an unknown target speaker (on the basis of separately trained speaker models), and the estimated speaker-similarity vector is input into a multi-speaker speech synthesis model, the resulting synthetic speech will probably resemble the voice of the new target speaker. More importantly, the speaker-similarity vector can be computed by using widely used text-independent automatic-speaker-verification models, such as the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [12] and i-vector probabilistic linear discriminant analysis (PLDA) [13], without the need for text information; hence, unsupervised speaker adaptation can be achieved.

In this paper, also, we train robust speaker verification models to calculate appropriate posterior probabilities from low-quality (i.e., noisy and reverberant) speech to synthesize the target speaker's voice even if low-quality speech are given as adaptation data. An issue to be addressed is mismatch between recording conditions for the training data and adaptation data fed into speaker-verification models, in which posterior probabilities are calculated from low-quality speech via speaker-verification models trained by using studio-quality speech data. To alleviate this conditional mismatch between training data and adaptation data, low-quality speech data are artificially created, and speaker-verification models are trained using the created data instead of studio-quality speech data.

The remainder of the paper is as follows. Section 2 describes multi-speaker modeling and the conventional speaker-adaptation technique using input codes. Section 3 explains the proposed speaker-adaptation technique that does not require text information. Section 4 describes how to artificially create noisy and reverberant speech. In Section 5, the proposed approaches are evaluated by using studio-quality speech data, and in Section 6, they are evaluated by using low-quality speech data. Section 7 concludes the paper.

II. Multi-speaker speech synthesis and speaker adaptation using input codes

The previously proposed multi-speaker speech-synthesis model [7] was trained using simply pooled data of multiple speakers. Identity, gender, and age of the speaker are represented using one-hot vectors, binary values (0: Female; 1: Male), and raw age values, respectively, and added as a part of the input to the neural network.

To adapt the above multi-speaker speech-synthesis models to a new speaker, the BP algorithm is used to minimize the mean-square prediction error over a small amount of data uttered by the target speaker according to the study by Bridle and Cox [14]. Note that the BP algorithm only updates the speaker codes, without changing the DNN weights, in contrast to algorithms developed in other studies, e.g., [1], that used fixed codes but added new weights. The BP algorithm starts from the average speaker code and continues until it estimates a new speaker code and a stopping criterion is satisfied.

III. Unsupervised speaker adaptation using a speaker-similarity vector

A. Flow of the proposed unsupervised speaker-adaptation technique

The proposed unsupervised-speaker-adaptation technique using speaker-similarity vectors is explained. The procedure for training a multi-speaker model and performing speaker adaptation based on the speaker similarity vectors is explained as the following steps.

1) First, text-independent speaker verification models are constructed for each of the training speakers included in a speech database, which is also used for training the multi-speaker speech synthesis model. GMM-UBM [12] or i-vector/PLDA [13] is used as a text-independent speaker verification model.
2) Then, the posterior probability of each training speaker given by one of the multiple text-independent speaker verification models in Step 1 is computed. The obtained posterior probabilities are concatenated to form a speaker similarity vector for each of the training speakers. 112-dimensional speaker similarity vectors are obtained (since the number of training speakers was 112).
3) Next, the speaker similarity vectors computed in Step 2 are used to replace the one-hot-vector based speaker code, and a DNN-based multi-speaker speech synthesis model is constructed. Linguistic features, gender, and age codes are the same as those used in the systems described in Section 2[1]
4) Speaker adaptation is performed as follows. A speaker-similarity vector of an unknown target speaker is estimated in a similar way as in Step 2: the posterior probabilities of the target speaker given by the multiple text-independent speaker-verification models are computed, and the obtained posterior probabilities are concatenated to form a speaker-similarity vector.
5) The estimated speaker-similarity vector of the target speaker is used as a new speaker code of the above multi-speaker speech-synthesis model, thereby changing the speaker characteristics of synthetic speech.

B. Speaker-verification models

For text-independent speaker verification, the GMM-UBM [12] and i-vector/PLDA [13] approaches are widely used. The proposed technique also uses these approaches. As for the GMM-UBM approach, a speaker model is obtained by adapting parameters of GMMs trained using speech data of many speakers [12]. As for i-vector/PLDA, i-vectors are first computed from sufficient statistics of speech data and are regarded as observations for a Gaussian PLDA model given as

$$\boldsymbol{w}_u = \bar{\boldsymbol{w}} + \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\alpha}_u + \boldsymbol{\epsilon}_u, \quad (1)$$

[1]The same technique may also be used to estimate age and gender codes. However, this option is not explored in this paper due to space limitation.

where $\bar{\boldsymbol{w}}$ is a speaker-independent supervector. $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ represent eigenvoice matrices for speaker- and channel-dependent components, respectively. Speaker and channel factors, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_u$, are assumed to have a standard Gaussian distribution as a prior distribution. In this study, the third term in Eq. (1) was not used.

C. Advantage of proposed framework

As mentioned earlier, several techniques for speaker adaptation using i-vectors [5] or d-vectors [15] have been developed. As for the former, i-vectors are directly used as inputs for DNN-based speech synthesis. On the other hand, as for the proposed framework, GMM-UBM or i-vector/PLDA is used only to calculate posterior probabilities for each training speaker. That is, the proposed multi-speaker speech-synthesis model does not depend on any acoustic parameterization or dimensions of i-vectors and has weaker dependency on acoustic features.

An unsupervised speaker-adaptation technique using a bottle-neck layer of a DNN-based speaker-recognition model for DNN-based speech synthesis was proposed by Doddipatla et al. [15]. As for this technique, PCA is applied to the bottle-neck features of the DNN-based speaker recognition, and the first eigenvector is interpolated on the basis of the posterior probabilities of the speaker-recognition model. In the following, we argue that the proposed technique is much simpler and more intuitive for constructing a flexible multi-speaker speech-synthesis model.

IV. Speaker-verification models robust against low-quality adaptation speech data

Speech used as adaptation data is usually low quality because recording studio-quality speech incurs high cost. In this study, robust speaker verification models are trained to perform the proposed unsupervised speaker adaptation without significant degradation of speech quality even if low-quality speech is given as adaptation data. To train speaker-verification models robust against low-quality speech data, the mismatch between recording conditions for training and adaptation speech data fed into the models needs to be alleviated. Hence, low-quality speech data is artificially created by adding noise and reverberation to studio-quality speech, and the created data is used for training the speaker-verification models.

The low-quality speech data was artificially created by using the Demand noise database [16] and the ACE Challenge reverberant database [17]. The low-quality speech was created by adding noise from the Demand database and reverberation from the ACE Challenge database to studio-quality speech waveforms. It is assumed that adaptation speech data used for speech synthesis is recorded in indoor rooms, so noise and room impulse responses recorded in an office or meeting room were used. The first channel of office-and-meeting-room noise recordings at 48 kHz sampling frequency was selected from the Demand database, and room impulse responses recorded in an office or meeting room (Office 1 and Meeting Room 1) were selected from the ACE database. Noise and reverberation were added to studio-quality speech in the same way as used in [18] as follows,

$$y = x * h_1 + \alpha(n * h_2), \tag{2}$$

where x and n represent a studio-quality speech waveform and a noise waveform, respectively, h_1 and h_2 represent the room impulse responses, $*$ is a convolution operator, and α is used for adjusting signal-to-noise ratio (SNR). Room impulse responses h_1 and h_2 are recorded using microphones located in positions 1 and 2.

In our experiments, adaptation speech data was also artificially degraded by using the same way. Using speech waveforms recorded under real conditions as adaptation data is future work.

V. Experiments using studio-quality speech data

The proposed technique for unsupervised speaker adaptation using studio-quality speech data as adaptation data was evaluated as described below.

A. Experimental conditions

Speech database: For our experiments, the Japanese Voice Bank corpus, containing studio-quality native Japanese speech uttered by 65 males and 70 females aged between 10 and 89, was used. The speech from 56 males and 56 females was used to train the speaker-verification models and the multi-speaker speech-synthesis models. The speech from the remaining speakers (9 males and 14 females) was saved for speaker adaptation. With approximately 100 utterances per speaker, this dataset yielded a total of 11,154 training-data utterances. For the adaptation experiments, either 10, 50, or 100 utterances from each of the 23 speakers not included in the training set were used as adaptation materials. The sampling frequency of the speech-signal waveform was 48 kHz. Speaker adaptation was evaluated by using 10 utterances per speaker not included in either the training or adaptation sets.

Speaker-verification models: To train the speaker verification models, an open-source toolkit called SIDEKIT [19] was used. The acoustic features used for training these models are listed in Table I. Since spectral features (20-dimensional MFCCs) and fundamental frequency/F0 (1-dimensional) are dimensionally significantly different, 20-dimensional F0 features were also obtained by applying a discrete cosine transform (DCT) to fundamental frequency values of the current, next, and previous 32 frames. Moreover, instead of the standard MFCC, spectral features used for speech synthesis models, referred to as MGC, were investigated. Then, GMMs with 64 mixtures were trained to extract 400-dimensional i-vectors. The size of the eigenvoice matrices for the speaker dependent components in Gaussian PLDA was 20.

Speech-synthesis models: For extracting the acoustic features for the speech-synthesis model, WORLD analysis [20], [21] was used to obtain 259-dimensional acoustic feature vectors every 5 ms (each feature comprising 59-dimensional mel-spectral coefficients, a linearly interpolated fundamental frequency on the mel scale, and 25-dimensional band aperiodicities, along with their delta and delta-delta). The 259th feature

TABLE I
ACOUSTIC FEATURES USED FOR SPEAKER VERIFICATION MODELS.

MFCC	19-dim MFCCs (plus energy), Δ, Δ^2
MGC	19-dim WORLD mel-cepstrum (plus 0th), Δ, Δ^2
F0	20-dim features derived from F0, Δ, Δ^2

TABLE II
FOUR MULTI-SPEAKER SPEECH-SYNTHESIS MODELS USED FOR SPEAKER ADAPTATION EXPERIMENTS. *g* AND *i* DENOTE *GMM* AND *i-vector*, RESPECTIVELY.

Systems	Multi-speaker model	Adaptation
averaged	one-hot vector	–
supervised	one-hot vector	vector estimated by BP
unsupervised (*g*)	speaker similarity vec. obtained from GMM-UBM	
unsupervised (*i*)	speaker similarity vec. obtained from i-vector/PLDA	

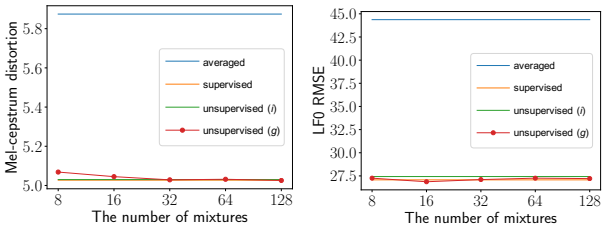
was a binary voiced/unvoiced flag. 389-dimensional linguistic features were used as an input vector. This input vector was augmented with speaker, gender, and age codes. The oracle duration was used since it makes it possible to easily compute objective measures such as mel-cepstral distortion. All multi-speaker speech synthesis models were feedforward DNNs with five hidden layers of 1024 nodes each. Sigmoid activation functions were used for all units in the hidden and output layers. The models were initialized randomly and trained to minimize the mean square error by stochastic gradient descent.

Speaker adaptation: The proposed unsupervised speaker-adaptation technique was compared with a supervised speaker-adaptation technique using speaker codes. Systems constructed for the experiments are listed in Table. II. An averaged system is a reference system which uses one-hot vectors to train a multi-speaker model and replaces all one-hot vector elements with their average value during synthesis, since it can be viewed as the average voice system. In a supervised system, the multi-speaker model is the same as that used in the averaged system, but the speaker code for the target speaker is estimated on the basis of BP. Unsupervised systems (*GMM* and *i-vector*) are proposed unsupervised speaker-adaptation systems, in which speaker similarity vectors are estimated by using either GMM-UBM or i-vector/PLDA, respectively.

B. Objective evaluation of multi-speaker modeling

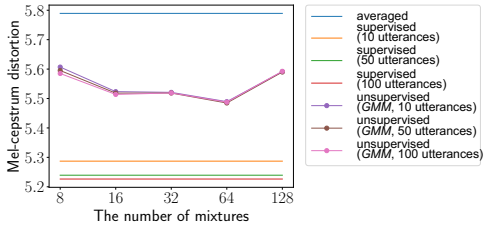
Performance of multi-speaker modeling using the proposed technique was evaluated. Speaker codes used in training the multi-speaker speech synthesis model were used to synthesize voices of training speakers. Objective results in terms of mel-cepstrum distortion and root mean square error (RMSE) of log $F0$ (in short, LF0 RMSE) are shown in Fig. 1. The number of mixtures for the unsupervised system (*GMM*) was 8, 16, 32, 64 or 128. Only MFCCs were used as features to train the speaker verification models.

It can be seen from Fig. 1 that all the other supervised and unsupervised systems were significantly accurate than the averaged system. This result indicates the multi-speaker speech-synthesis model using the proposed speaker-similarity vectors as well as one-hot vectors was successful at approximating the many speakers in the training corpus. Next, as for the

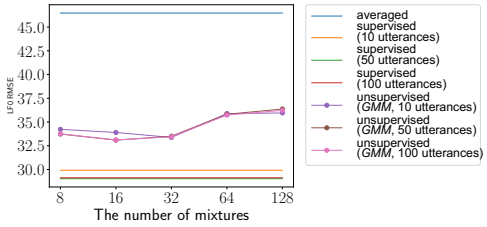


(a) Mel-cepstrum distortion (b) LF0 RMSE

Fig. 1. Objective results (Mel-cepstrum distortion and LF0 RMSE) of multi-speaker speech synthesis models.



(a) Mel-cepstrum distortion



(b) LF0 RMSE

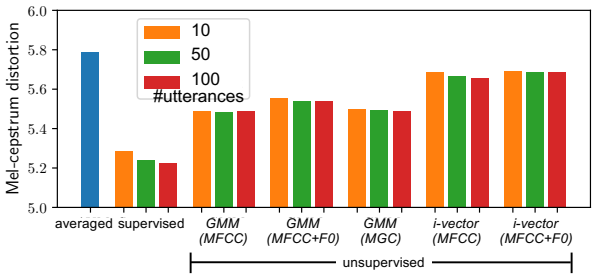
Fig. 2. Objective results of supervised and proposed unsupervised adaptation techniques. The number of mixtures for GMM was 8, 16, 32, 64 or 128. The numbers included in the labels represent the number of adaptation utterances (10, 50 or 100 utterances).

supervised and the proposed unsupervised (*GMM* and *i-vector*) systems, their performances do not significantly differ.

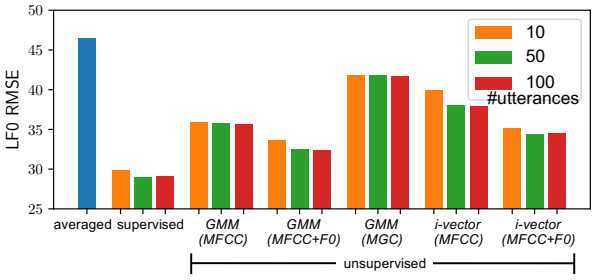
C. Objective evaluation of speaker-adaptation performance

Supervised and unsupervised adaptation: Objective results of supervised and proposed unsupervised speaker-adaptation systems based on GMM-UBM are shown in Fig. 2 in terms of mel-cepstrum distortion and LF0 RMSE. Objective results of the averaged system are also shown. First, it can be seen that the unsupervised system based on GMM-UBM (*GMM*) produces smaller errors than the averaged system, showing that the proposed technique successfully performed speaker adaptation. It can also be seen that the results of the unsupervised system (*GMM*) are worse than those of the supervised systems, as expected.

Second, in terms of the number of mixtures for the unsupervised system (*GMM*), it can be seen that the lowest mel-cepstrum distortion and LF0 RMSE are obtained by using GMMs with 32 and 64 mixtures, respectively. The number of mixtures of GMMs may be smaller than the number of mixtures generally used in speaker-verification tasks. Since the final aim is to perform speaker adaptation rather than verifi-



(a) Mel-cepstrum distortion



(b) LF0 RMSE

Fig. 3. Objective results of the averaged, supervised and unsupervised systems (*GMM* and *i-vec*). Acoustic features used for training speaker-verification models are shown inside brackets.

cation, the appropriate number of mixtures would be different from the number of mixtures used for speaker verification.

It was also noticed that when a larger number of mixtures (e.g., 128) was used, the speaker-similarity vectors for training speakers became closer to the one-hot vectors due to over-fitting to them; hence, GMM-UBM did not provide appropriate similarity vectors of unknown speakers. Furthermore, it can be seen that the performance of the systems using the smaller number of mixtures (e.g., 8) is worse than the systems using 32 or 64 mixtures. This result is due to the fact that the speaker-verification models are too simple to represent all training speakers. In the following experiments, GMMs with 64 mixtures were used as the speaker-verification models.

Evaluation of the impacts of different speaker-verification models for the proposed technique: Objective results of the averaged, supervised and unsupervised systems are shown in Fig. 3. The acoustic features used for training each of the speaker-verification models are shown inside brackets. First, it can be seen that although mel-cepstrum distortion of the unsupervised system using f0 features (*GMM(MFCC+F0)*) is slightly increased compared to that of the unsupervised system using only MFCCs (*GMM(MFCC)*), LF0 RMSE of the unsupervised system using f0 features (*GMM(MFCC+F0)*) decreased compared to that of the unsupervised system using only MFCCs (*GMM(MFCC)*). This result means that the speaker-similarity vectors considering log F0 made log F0 of synthetic speech closer to that of the target speakers. A similar tendency was observed for the unsupervised systems based on i-vector/PLDA (*i-vector(MFCC)* and *i-vector(MFCC+F0)*).

Second, as for comparing the unsupervised systems using MFCCs and MGCs as acoustic features (*GMM(MFCC)* and *GMM(MGC)*), the unsupervised system using MGCs (*GMM(MGC)*) was significantly less accurate than the unsupervised system using MFCCs (*GMM(MFCC)*) in terms of LF0 RMSE, although the mel-cepstrum prediction performance was almost the same. This result is thought to be due to the fact that MFCCs include more F0 information than MGCs.

Third, as shown in Fig. 3, the unsupervised system based on i-vector/PLDA (*i-vector*) was significantly less accurate than the unsupervised system based on GMM-UBM (*GMM*) regardless of the acoustic features used for training the speaker verification models. To understand why the unsupervised system based on i-vector/PLDA was less accurate than the unsupervised system based on GMM-UBM, the estimated speaker-similarity vectors were investigated. Estimated speaker similarity-values of the training speaker themselves and accumulated speaker-similarity values of the top-N training speakers used in speaker adaptation are listed in Tables III and IV, respectively. From these tables, the following tendencies are clear:

- In case of the unsupervised system (*i-vector*), speaker-similarity vectors for training speakers were representations close to one-hot vectors.
- In case of the unsupervised system (*i-vector*), speaker-similarity values for a few training speakers were very large.

That is, the speaker-similarity vectors were not appropriately estimated when i-vector/PLDA was used with current parameter settings.

TABLE III
ESTIMATED SPEAKER-SIMILARITY VALUES OF TRAINING SPEAKERS THEMSELVES.

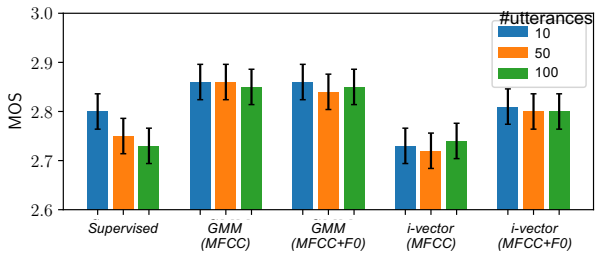
unsupervised (*GMM(MFCC)*)	0.15
unsupervised (*GMM(MFCC+F0)*)	0.087
unsupervised (*i-vector(MFCC)*)	0.99
unsupervised (*i-vector(MFCC+F0)*)	0.98

TABLE IV
ACCUMULATED SPEAKER-SIMILARITY VALUES OF TOP-N TRAINING SPEAKERS USED IN SPEAKER ADAPTATION. THE NUMBER OF ADAPTATION UTTERANCES WAS 100.

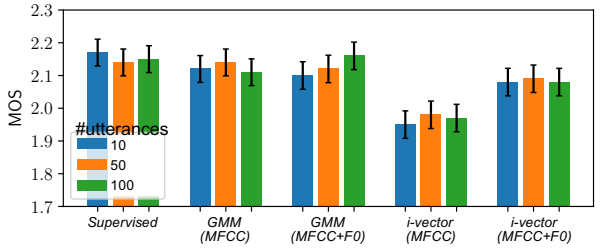
	top 1	top 2	top 3
unsupervised (*GMM(MFCC)*)	0.039	0.072	0.10
unsupervised (*GMM(MFCC+F0)*)	0.041	0.075	0.11
unsupervised (*i-vector(MFCC)*)	0.83	0.96	0.99
unsupervised (*i-vector(MFCC+F0)*)	0.58	0.75	0.84

D. Subjective evaluation

The subjective evaluation was conducted involving 180 crowd-sourced Japanese native listeners. Fifteen conditions (five systems: the supervised system and the four unsupervised systems (*GMM(MFCC)*, *GMM(MFCC+F0)*, *i-vector(MFCC)*, and *i-vector(MFCC+F0)*) $\times$ three amounts of data (10, 50, and 100 adaptation utterances) were evaluated. Number of synthetic samples was $3,450$ (15 conditions $\times$ 23 target speakers $\times$ 10 test sentences). Participants evaluated speech naturalness and speaker similarity compared with a reference natural speech utterance on a five-point mean opinion score



(a) Naturalness



(b) Speaker similarity

Fig. 4. Subjective results. Participants evaluated speech naturalness and speaker similarity compared with a reference natural speech utterance on a five-point mean opinion score (MOS) scale.

(MOS) scale. Each synthetic sample was evaluated 10 times, giving a total of 34,500 evaluation scores.

The subjective results are shown in Fig. 4. In terms of speech naturalness, first, it can be seen from Fig. 4(a) that MOS of the supervised system decreased when a larger amount of adaptation data was used, although the objective results improved when a larger amount of adaptation data was used. This result may be due to the fact that the representation of speaker codes estimated by BP significantly differed from the one-hot vectors used in training the multi-speaker speech synthesis model. Second, among the supervised and the proposed unsupervised systems based on GMM-UBM (*GMM(MFCC)* and *GMM(MFCC+F0)*), the proposed systems (*GMM(MFCC)* and *GMM(MFCC+F0)*) outperformed the supervised system. Third, the proposed systems using the speaker verification models based on i-vector/PLDA (*i-vector(MFCC)* and *i-vector(MFCC+F0)*) obtained lower scores than those obtained by the unsupervised systems based on GMM-UBM (*GMM(MFCC)* and *GMM(MFCC+F0)*).

Next, in terms of speaker similarity, it can be seen from Fig. 4(b) that the supervised system and the unsupervised systems based on GMM-UBM (*GMM(MFCC)*, and *GMM(MFCC+F0)*) obtained almost the same scores. And the unsupervised system based on i-vector/PLDA (*i-vector(MFCC)* and *i-vector(MFCC+F0)*) obtained lower scores than those obtained by the unsupervised systems based on GMM-UBM (*GMM(MFCC)* and *GMM(MFCC+F0)*). These results indicate that the new speaker-similarity vectors estimated by using speaker verification based on GMM-UBM were effectively used to construct a speaker-adapted system.

TABLE V
QUALITY TYPES OF TRAINING AND ADAPTATION DATA FED INTO SPEAKER-VERIFICATION MODELS

Training data	Adaptation data	Quality condition
CLEAN	CLEAN	ideal
CLEAN	OFFICE	mismatched
CLEAN	MEETING	mismatched
OFFICE	OFFICE	matched
MEETING	MEETING	matched

VI. EXPERIMENTS USING LOW-QUALITY SPEECH DATA

The proposed unsupervised-speaker-adaptation technique was evaluated by using low-quality speech data as adaptation data.

A. Experimental conditions

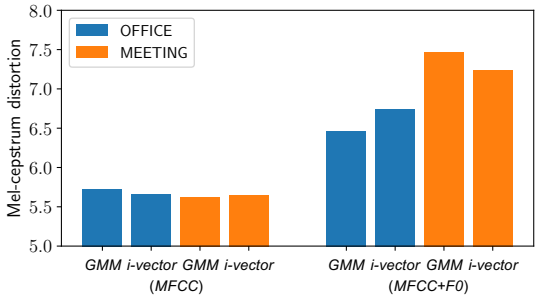
The Japanese Voice Bank corpus was also used for the experiments using low-quality speech data. The same utterances and speakers used in the experiments using only studio-quality speech data were used for training and adaptation, although 100 utterances from each of target speakers were used as adaptation data.

In contrast to the experiments using only studio-quality data, noise and reverberation were added to studio-quality speech for artificially creating low-quality data. Low-quality speech data created by adding noise and reverberation simulating an office or meeting room (OFFICE/MEETING) as well as studio-quality speech data (CLEAN) were used to train the speaker-verification models used for the proposed technique. To add noise to studio-quality speech, noise segments were randomly chosen. Also, signal to noise ratio (SNR) of low-quality speech used for training the speaker-verification models was adjusted by using α in Eq. (2). As SNR, 2.5-, 7.5-, 12.5- or 17.5-dB utterances were randomly selected to train the robust speaker-verification models against various noise strengths. On the other hand, 0.0-, 5.0-, 10.0- or 15.0-dB utterances were selected as SNR to create low-quality adaptation speech data. Eight sets of low-quality adaptation data (2 OFFICE/MEETING $\times$ 4 SNR types) were eventually created for evaluation. As listed in Table V, performance of the proposed unsupervised speaker-adaptation technique for five conditions (one ideal, two mismatched, and two matched adverse or low-quality conditions) was evaluated.

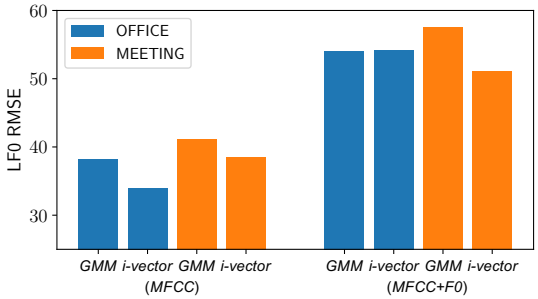
GMMs with 64 mixtures were trained for systems based on GMM-UBM and i-vector/PLDA. The network architecture of the speech-synthesis models was the same as that used in the experiments using only studio-quality speech data. Multi-speaker speech-synthesis models of the proposed technique were trained by using speaker-similarity vectors obtained from studio-quality speech.

B. Objective evaluation

Evaluation of impacts of different speaker-verification models used for the proposed technique: Objective results of the proposed technique for unsupervised speaker adaptation using low-quality adaptation data are shown in Fig. 5. Results

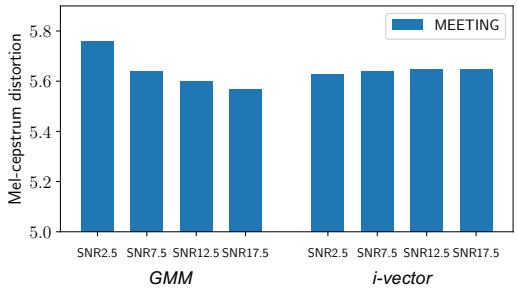


(a) Mel-cepstrum distortion

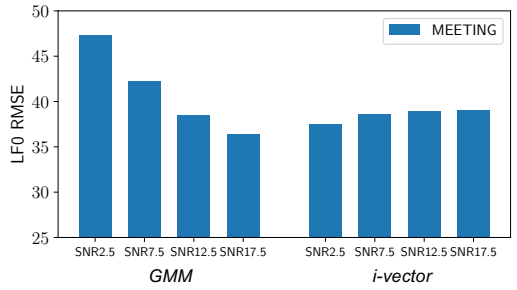


(b) LF0 RMSE

Fig. 5. Objective results of the proposed unsupervised speaker adaptation technique using low-quality adaptation data. Speaker verification models based on GMM-UBM (*GMM*) and i-vector/PLDA (*i-vector*) were trained. Systems with *(MFCC+F0)* used f0 features as well as MFCC to train the speaker-verification models. Labels show the speech quality types of noise and reverberation used for the speaker-verification models. In this figure, objective results in matched conditions are only shown.



(a) Mel-cepstrum distortion



(b) LF0 RMSE

Fig. 6. Objective results of the proposed unsupervised-speaker adaptation-technique using low-quality adaptation data. The results vary according to SNR of the adaptation data.

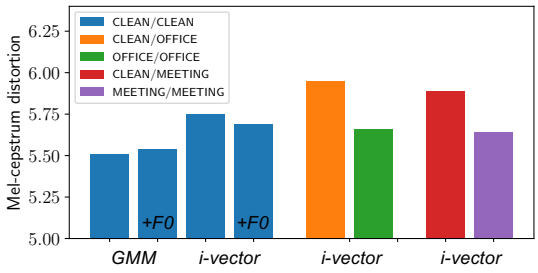
obtained under matched conditions, in which the same type of lower quality speech data was fed into the speaker-verification models in the training and adaptation phases, are shown in Fig. 5. First, it can be seen from Fig. 5 that using F0 features for the speaker-verification models increases mel-cepstrum distortion and LF0 RMSE. This is because performance of F0 extraction from a low-quality speech waveform was problematic, and the speaker-verification models using F0 features cannot output the appropriate speaker-similarity vector for speaker adaptation. Second, it can be seen from Fig. 5 that the systems based on i-vector/PLDA outperformed those based on GMM-UBM in the case of *MFCC*. These results are the opposite of those obtained from the experiments using only studio-quality speech data.

To compare the speaker-verification models based on GMM-UBM and i-vector/PLDA in more detail, results separated according to SNR of adaptation data are shown in Fig. 6. As shown in Fig. 6, performance of the system based on i-vector/PLDA was almost the same in all SNR cases; however, performance of that based on GMM-UBM was drastically effected by noise strength. These results indicate that the speaker-verification models based on i-vector/PLDA are more robust against low-quality speech data than the ones based on GMM-UBM for the proposed unsupervised adaptation technique.

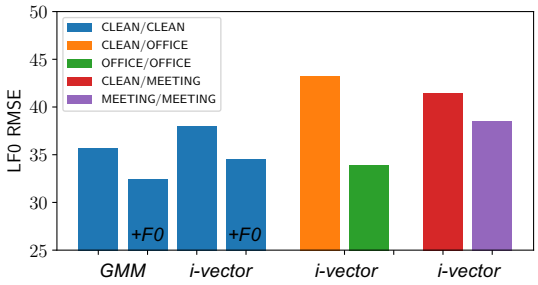
Evaluation of the proposed technique for matched and mismatched speech quality conditions: Objective results of the proposed unsupervised speaker adaptation under matched and mismatched conditions between training and adaptation data fed into the speaker-verification models are shown in Fig. 7. First, it can be seen from Fig. 7 that performance of the proposed unsupervised speaker adaptation using low-quality adaptation data is worse than that of the best systems, i.e., the unsupervised system (*GMM*) for Mel-cepstrum distortion and the unsupervised system (*GMM(F0)*) for LF0 RMSE, under the ideal condition using studio-quality speech as both training and adaptation data, as expected. As for comparing the results of the systems under matched and mismatched conditions in Fig. 7, the systems under matched conditions obtain better performance than the ones under mismatched conditions. This result indicates that training speaker-verification models with speech data whose quality is matched to adaptation data improves performance of the proposed unsupervised-speaker-adaptation technique.

C. Subjective evaluation

The subjective evaluation was conducted involving 153 crowd-sourced Japanese native listeners. Twenty systems shown in Fig. 8 were evaluated. Number of synthetic samples was $4,600$ (20 conditions $\times$ 23 target speakers $\times$ 10 test sentences). Participants evaluated speech naturalness and speaker similarity compared with a reference natural speech utterance on a five-point mean opinion score (MOS) scale. Each synthetic sample was evaluated five times, giving a total of $23,000$ evaluation scores.



(a) Mel-cepstrum distortion



(b) LF0 RMSE

Fig. 7. Objective results of the proposed technique for unsupervised speaker adaptation for matched and mismatched conditions between training and adaptation data fed into speaker verification models. The words before and after the slashes in the labels represent speech quality of the training and adaptation data, respectively.

The subjective results are shown in Fig. 8. Similar tendencies observed in the objective results can be observed from the subjective results. First, using F0 features extracted from low-quality speech data worsened the performance of the proposed unsupervised speaker adaptation. Second, under matched low-quality conditions, the systems based on i-vector/PLDA obtained better subjective scores than the systems based on GMM-UBM.

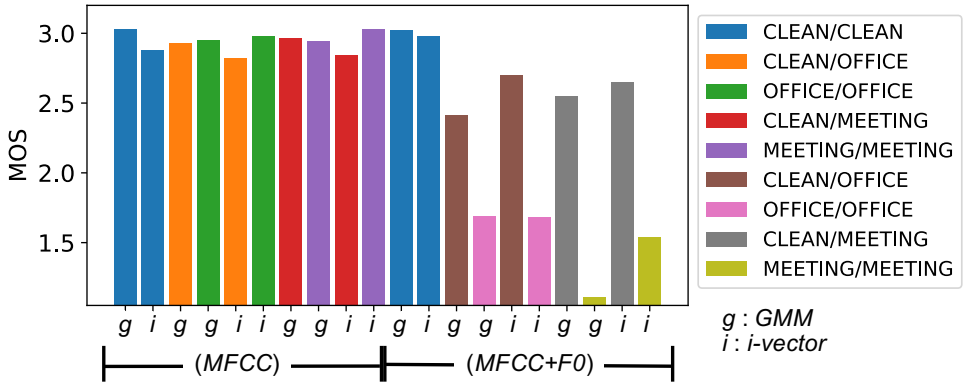
In regard to the subjective results, most importantly, some systems using low-quality speech as adaptation data, i.e., the unsupervised system (*i-vector*, OFFICE/OFFICE and MEETING/MEETING) without F0 features, obtained almost the same scores compared to the systems under the ideal conditions in terms of speech naturalness and speaker similarity.

In summary, contrasting to the results observed in the experiment using only studio-quality speech, the results of the experiment using low-quality speech data showed that f0 features were not useful and i-vector/PLDA was more effective than GMM-UBM for the proposed unsupervised-speaker-adaptation technique. Also, it was found that training the speaker-verification models using speech data whose quality is matched to adaptation data improves performance of the proposed unsupervised speaker adaptation. Finally, if there is no quality mismatch between the training and adaptation data fed into the speaker-verification models, the proposed unsupervised-speaker-adaptation technique using low-quality adaptation data achieved almost the same performance as that of the systems using studio-quality speech as adaptation data.

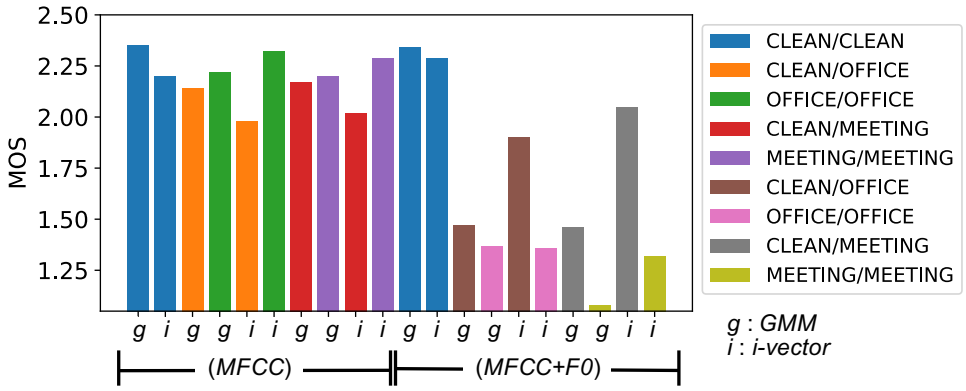
VII. Conclusions

An unsupervised-speaker-adaptation technique for DNN-based speech synthesis with input codes, using only speech data from a target speaker without transcriptions, was proposed. As for the proposed technique, the speaker-similarity vectors obtained using the speaker-verification models were used as the speaker codes instead of conventional one-hot vectors. The results of experiments using only studio-quality speech data demonstrated that the use of the speaker-similarity vectors estimated from the speech of an unknown target speaker as a speaker code appropriately changed the speaker characteristics of synthetic speech. They also showed that compared with a supervised adaptation technique based on BP, the proposed technique using GMM-UBM performed unsupervised speaker adaptation well without speech quality degradation. The results of the experiments using low-quality speech data showed that using speech data whose quality is matched to adaptation data for training the speaker-verification models effectively improved performance of the proposed unsupervised speaker adaptation.

Our future work includes evaluation of the proposed technique using MP3 or AMR codec speech and speech recorded under real conditions as adaptation data.



(a) Naturalness



(b) Speaker similarity

Fig. 8. Subjective results. Participants evaluated speech naturalness and speaker similarity compared with a reference natural speech utterance on a five-point mean opinion score (MOS) scale. The words before and after the slashes in the labels represent speech quality types of training and adaptation data, respectively.

REFERENCES

[1] Nobukatsu Hojo, Yusuke Ijima, and Hideyuki Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. Interspeech*, 2016.

[2] Sercan Ömer Arik, Gregory F. Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *CoRR*, vol. abs/1705.08947, 2017.

[3] Yi Zhao, Daisuke Saito, and Nobuaki Minematsu, "Speaker representations for speaker adaptation in multiple speaker BLSTM-RNN-based speech synthesis," in *Proc. Interspeech*, 2016.

[4] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. SLT*, 2014, pp. 171–176.

[5] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. Interspeech*, 2015.

[6] Yuchen Fan, Yao Qian, Frank K. Soong, and Lei He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[7] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, and Junichi Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," *Proceedings of ICASSP*, pp. 4905–4909, 2017.

[8] Yuchen Fan, Yao Qian, Frank K. Soong, and Lei He, "Speaker and language factorization in dnn-based tts synthesis," *Proc. ICASSP*, pp. 5540–5544, 2016.

[9] S. Takaki, S. Kim, and J. Yamagishi, "Speaker adaptation of various components in deep neural network based speech synthesis," *Proceedings of Speech Synthesis Workshop 9 (SSW9)*, pp. 167–173, 2016.

[10] Simon King, Keiichi Tokuda, Heiga Zen, and Junichi Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," *Proc. Interspeech*, pp. 1869–1872, 2008.

[11] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, "Voice synthesis for in-the-wild speakers via a phonological loop," *ArXiv e-prints*, July 2017.

[12] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2017.

[13] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," *Odyssey 2010*, 2010.

[14] John S. Bridle and Stephen Cox, "RecNorm: Simultaneous normalisation and classification applied to speech recognition," in *Proc. NIPS*, 1990, pp. 234–240.

[15] Rama Doddipatla, Norbert Braunschweiler, and Ranniery Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors," *Proc. Interspeech*, pp. 3404–3408, 2017.

[16] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," vol. 133, pp. 3591, 05 2013.

[17] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2014, pp. 313–317.

[18] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[19] Anthony Larchera, Kong Aik Lee, and Sylvain Meignier, "An extensible speaker identification sidekit in python," *Proceedings of ICASSP*, 2016.
[20] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," *the Stockholm Music Acoustics Conference 2013 (SMAC2013)*, pp. 289–292, 2015.
[21] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,," *Speech Communication*, vol. 67, pp. 1–7, 2015.