IEICE TRANS. INF. & SYST., VOL.E102–D, NO.2 FEBRUARY 2019

PAPER

Discriminative Learning of Filterbank Layer within Deep Neural Network Based Speech Recognition for Speaker Adaptation

Hiroshi SEKI[†a)], ***Nonmember***, **Kazumasa YAMAMOTO**[††], **Tomoyosi AKIBA**[†], ***Members***, ***and*** **Seiichi NAKAGAWA**[†,††], ***Fellow***

SUMMARY Deep neural networks (DNNs) have achieved significant success in the field of automatic speech recognition. One main advantage of DNNs is automatic feature extraction without human intervention. However, adaptation under limited available data remains a major challenge for DNN-based systems because of their enormous free parameters. In this paper, we propose a filterbank-incorporated DNN that incorporates a filterbank layer that presents the filter shape/center frequency and a DNN-based acoustic model. The filterbank layer and the following networks of the proposed model are trained jointly by exploiting the advantages of the hierarchical feature extraction, while most systems use predefined mel-scale filterbank features as input acoustic features to DNNs. Filters in the filterbank layer are parameterized to represent speaker characteristics while minimizing a number of parameters. The optimization of one type of parameters corresponds to the Vocal Tract Length Normalization (VTLN), and another type corresponds to feature-space Maximum Linear Likelihood Regression (fMLLR) and feature-space Discriminative Linear Regression (fDLR). Since the filterbank layer consists of just a few parameters, it is advantageous in adaptation under limited available data. In the experiment, filterbank-incorporated DNNs showed effectiveness in speaker/gender adaptations under limited adaptation data. Experimental results on CSJ task demonstrate that the adaptation of proposed model showed 5.8% word error reduction ratio with 10 utterances against the un-adapted model.

key words: *speech recognition, deep neural network, acoustic model, speaker adaptation, filterbank learning*

1. Introduction

In recent years, deep neural networks (DNNs) have been applied to automatic speech recognition (DNN-HMM; deep-neural-network hidden Markov models) and have outperformed conventional Gaussian mixture model (GMM) based methods [1]. One main advantage of DNNs is a hierarchical non-linear feature extraction under a simple objective function. Exploiting this property, some recent novel approaches focus on front-end learning based on DNNs that take low-level acoustic features [2]–[6]. Sainath et al. [2] and Sailor et al. [3] proposed an end-to-end model that uses waveform and performs frequency analysis. These studies reported that some of the learned characteristics showed a similarity with human auditory characteristics and traditional refined hand-crafted feature extractors [2], [3]. In addition, Sailor et al. [3] investigated the difference of center frequencies among models that were trained by both clean and noisy speech. They reported that the center frequency of learned filters do not show consistency between clean speech and noisy speech, suggesting that the optimal properties of filterbanks depend on the task and target environments. Zhu et al. [4] also presented a model to learn features directly from waveforms and performed convolution operations with several types of window sizes and stride parameters to push past the inherent trade-off between temporal and frequency resolutions. These DNN-based systems eliminate the feature extraction stage and significantly improve the recognition performance.

Earlier works reported the difference of filter characteristics caused by the condition of training data. However, since a system can not identify varying test speaker and test environment in advance, there is a mismatch between input test speech and learned model which causes performance reduction. Therefore, adaptation remains a major challenge for DNN-based systems, which must alleviate the mismatch and recover recognition performance. In practical use, it is preferable for low-level feature extractor to track various test conditions. To adapt DNNs, model adaptation techniques re-estimate the model parameters based on the test data. In this scenario, the trade-off between the size of the adaptation data and the number of parameters becomes a critical problem. In other words, too many parameters cause poor generalization and overfitting to the given data if the available adaptation data are limited.

In contrast to DNNs, a physiologically motivated model is composed of a small number of parameters. Therefore, a physiologically motivated model is advantageous in model adaptation under limited adaptation data. Furthermore, introducing restrictions resulting from a physiologically motivated model protects the (introduced) filterbank layer from extreme deterioration. Previously, we proposed a filterbank-incorporated DNN with a Gaussian filterbank layer at DNN's bottom and evaluated the effectiveness of our proposed model as a data-driven filterbank learning technique [7].

In this paper, we evaluate the adaptation of a filterbank-incorporated DNN. Since the filterbank layer that presents the filter shape/center frequency consists of a small number of parameters, a filterbank-incorporated DNN is effective in regards to adaptation with limited available data. The fol-

Manuscript received July 13, 2018.
Manuscript revised October 2, 2018.
Manuscript publicized November 7, 2018.

[†]The authors are with the Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
[††]The author is with the Department of Computer Science, Chubu University, Kasugai-shi, 487–8501 Japan.
a) E-mail: seki@nlp.cs.tut.ac.jp

DOI: 10.1587/transinf.2018EDP7252

lowings are the contributions of this work:

(i) proposed a filterbank-incorporated DNN;
(ii) evaluated our proposed model for speaker/gender adaptation and compared various filter types;
(iii) discussed the relation between the physical characteristics of vocal tracts and optimal filterbanks from an engineering viewpoint.

The rest of the paper is organized as follows. We first review related works on adaptation techniques and data-driven filterbank learning techniques in Sect. 2. Next, we introduce our framework of a filterbank-incorporated DNN in Sect. 3. Experiments are conducted in Sect. 4, followed by a conclusion and future work in Sect. 5.

2. Review of Adaptation Technique and Related Works

In this section, we first summarize feature adaptation and model adaptation techniques in Sect. 2.1. We also discuss the introduced constraints for some adaptation techniques from the viewpoint of matrix multiplication and present data-driven filterbank learning methods in Sect. 2.2.

2.1 Adaptation Techniques and Introduced Constraints

Adaptation techniques can be roughly classified into three types: feature adaptation, model adaptation, and addition of auxiliary features [8]–[10]†. In this section, we summarize feature adaptation and model adaptation techniques that are related to our proposed filterbank-incorporated DNN.

Feature adaptation techniques are used to adapt the input acoustic features to the DNN-HMM system independent from the DNN. fMLLR [11], VTLN [12], and Maximum Likelihood Linear Transforms (MLLT) + Linear Discriminant Analysis (LDA) [13] are used to adapt the acoustic features that support the performance improvement of DNN-based systems, and Seide et al. [14] concluded that DNNs can subsume much of the VTLN gain. Feature adaptation methods factorize speech recognition tasks into several independent sub-tasks. Therefore, the feature adaptation criterion is different from other parts of DNN.

Model adaptation is a promising method to adapt a DNN that updates its parameters given the adaptation data. For model adaptation, structural changes and parameter restrictions are introduced to robustly learn a speaker specific transformation using short samples from the target speaker. Neto et al. and Bo et al. presented a Linear Input Network (LIN) that restricts the adapting layer to the input layer [15], [16]. The same idea is also applicable to other layers: Linear Hidden Network (LHN) and Linear Output Network (LON). The computation of each layer consists of a matrix multiplication and a non-linear transformation. The singular value decomposition (SVD) replaces the matrix to the product of two low ranked matrices. The SVD-based parameter reduction showed effective adaptation [17]. Swietojanski et al. [18] presented an approach that adapted hidden units called learning hidden unit contributions (LHUCs), which directly rescale the amplitude of the hidden units. Zhao et al. [19] presented an adaptation method to adapt the node activation function. LHUC and the adaptation of the node activation function also resemble a matrix multiplication of a diagonal matrix. In the particular case of feed-forward DNNs, the neighboring frames of the acoustic features are concatenated to take the context information into account that contributes to the senone classification. Focusing on such stacked frames, Seide et al. [14] inserted a linear layer that is tied across neighboring frames (fDLR; feature-space discriminative linear regression). In the model adaptation, a large number of parameters must be trained without causing any overfitting. Yu et al. [20] presented a Kullback-Leibler divergence-based regularization to address this concern. Such model adaptation techniques only focus on the reduction of free parameters. However, we must consider expressiveness against the total free parameters for adaptation under limited available data.

Several model adaptation techniques can be regarded as matrix multiplications. Figure 1 summarizes the relation among adaptation methods and introduced restrictions to the matrix. LIN inserts a matrix at the bottom of the DNN without any restrictions that resembles a fully connected layer. fDLR introduces a restriction where the matrix is a block-diagonal type and the block is shared across the diagonal. Therefore, frame-based transformation is carried out for each frame. Likewise, VTLN is regarded as a transformation by a tri-diagonal matrix, even though it is not adapted under the backpropagation framework [21]. LHUC is also regarded as a matrix multiplication by introducing a restriction under which the matrix takes a diagonal matrix. From these results, LHUC's expressiveness is included in

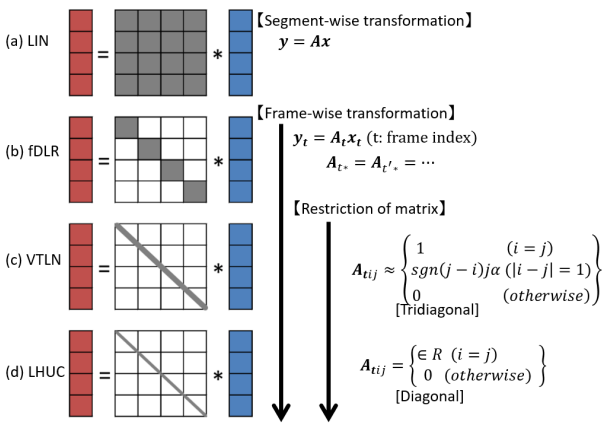


Fig. 1 Relations among adaptation methods and introduced restrictions: LIN inserts a matrix **A** without adaptation. Here, the input is composed of four frames. fDLR introduces a restriction with a block-diagonal matrix and the block is shared across the diagonal ($\mathbf{A}_t = \mathbf{A}_{t*} = \cdots$). VTLN and LHUC are regarded as a matrix multiplication with tridiagonal and diagonal matrices, respectively.

†The distinction between model adaptation and feature adaptation is getting blurry. Adaptation of denoising auto-encoder can be regarded as both feature adaptation and model adaptation.

the fDLR, which is again included in the LIN. The categorization of the feature transformations, which are based on the considerations of the Spectro-Temporal domain, was previously discussed [22].

Some studies reported an acoustic model based on convolutional neural networks (CNNs) [23]. A convolution operation focuses on the small localized regions of the input speech, unlike the fully connected layer. In addition, weight sharing significantly reduces the number of parameters. Kaneyama et al. [24] proposed a method to apply convolutional filters that follows a Gabor function for image texture classification. CNN's success shows that the introduction of structural restrictions are critical to capture locally invariant features and further improve the performance even though fully connected neural networks include CNN capability.

Other studies reported methods that represent speaker characteristics as a combination of the components of *bases*. Cluster adaptive training (CAT) combines multiple weight matrices using an interpolation vector to form one final DNN layer [25]. In the adaptation stage, the interpolation vector is updated while maintaining the weight matrices. The Factorized Hidden Layer (FHL) approach resembles CAT [26]. In FHL, the interpolation vector is shared among several layers and initialized by an i-vector. By introducing an interpolation vector, these studies separate speaker and phone spaces for efficient adaptations. CAT and FHL only adapt the interpolation vector, and robust adaptation is guaranteed only within the range that covered by training data.

2.2 Filterbank Learning

Finding an optimal filterbank is an important topic not only for speech recognition but also for speaker recognition, dialization, and event detection. Several studies proposed methods based on heuristic search algorithms. Pinheiro et al. [27] proposed a scheme to find the best filterbank configuration using an Artificial Bee Colony (ABC) algorithm for speaker verification. Charbuillet et al. [28] proposed an method to search for optimal center frequency and bandwidth based on genetic algorithms. These heuristic search algorithms independently repeat both the *selection* and *evaluation* stages. Several studies proposed methods that introduce objective functions. Kobayashi et al. [29] and Burget et al. [30] proposed methods based on a dimensionality reduction technique, and Suh et al. [31] proposed a method that measures filterbank properties derived from the Kullback-Leibler (KL)-divergence among filters. Recently, hierarchical feature extraction based on deep neural networks has become a topic of interest in classification tasks [6]. Sainath et al. [2] presented a method to apply convolution over a raw time-domain waveform. Sailor et al. [3] proposed a method based on a convolutional Restricted Boltzmann Machine that uses a raw time-domain waveform. Tokozume et al. [32] presented an end-to-end convolutional neural network for environmental sound classification. Su et al. [33] further introduced an event-specific Gaussian filterbank layer to handle different temporal properties of audio events. In this paper, we propose a novel approach to train and adapt filterbank based on DNN.

3. Discriminative Learning of a Filterbank Llayer

In this section, we first introduce neural network-based filterbank weighting, Gaussian filter and Gammatone filter, and compare with a conventional method, triangular filter, in Sect. 3.1.1. Next, we present a procedure to train our proposed model in Sect. 3.2, and summarize the advantages of our proposed methods in Sect. 3.3. Finally, we present a procedure to adapt our proposed model and compared methods (LHUC and SVD) in Sect. 3.4.

3.1 Incorporation of a Filterbank Layer

3.1.1 Triangular Filterbank

Filterbank feature is calculated by weighing spectra of speech waveform using triangular filterbank. The vertex of triangles is configured according to the mel-scale which models non-linear sensitivity of human perception. The mel-scale is defined as follows:

$$p(f) = 1127.0 \ln\left(\frac{f}{700.0} + 1.0\right) \tag{1}$$

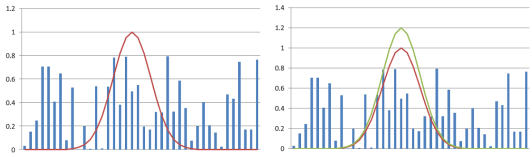
where f is linear frequency and $p(f)$ is mel-scale frequency. The pre-defined configuration of filterbank is unchanged at all times.

3.1.2 Gaussian Filterbank

Log mel-scale filterbank features are computed by applying filterbank weighting to the spectra. In general, hand-crafted triangular filters are used as filter shapes. However, this triangular filter is not differentiable and cannot be incorporated into a scheme of a backpropagation algorithm. To parametarize the filter, Biem et al. modeled its shape as a Gaussian function [34]:

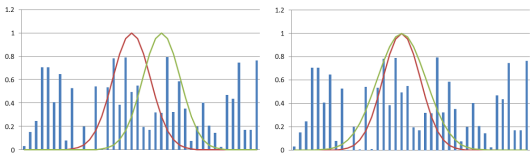
$$\theta_n(f) = \varphi_n \exp\left\{-\beta_n(p(\gamma_n) - p(f))^2\right\}, \tag{2}$$

where $\theta_n(f)$ is an n-th filter at frequency f. φ_n is a gain parameter, β_n is a bandwidth parameter, and γ_n is a center frequency. Function $p(\cdot)$ maps linear frequency f to the mel-scale. Three trainable parameters, φ_n, β_n, and γ_n, control the filter shape. Figure 2 visualizes the role of three parameters. A change of the gain parameter scales the magnitude of the filterbank features. This function is also realized by the adjusted weight in the following layer. A change of the center frequency parameter shifts the region of the power spectra on which the filter is focused. A change of the bandwidth parameter enlarges the power spectra region on which the filter is focused. A set of Gaussian filters can be regarded as a neural network layer that maintains a function of frequency domain smoothing.



(a) Initial Gaussian filter and input power spectra

(b) Change of gain parameters



(c) Change of center frequency

(d) Change of bandwidth

Fig. 2 Roles of parameter changes, gain, center frequency, and bandwidth. The x- and y-axis of each subfigure are power spectra and corresponding amplitude. The red line represents an initial filter shape, and the green line represents the filter shape after adaptation.

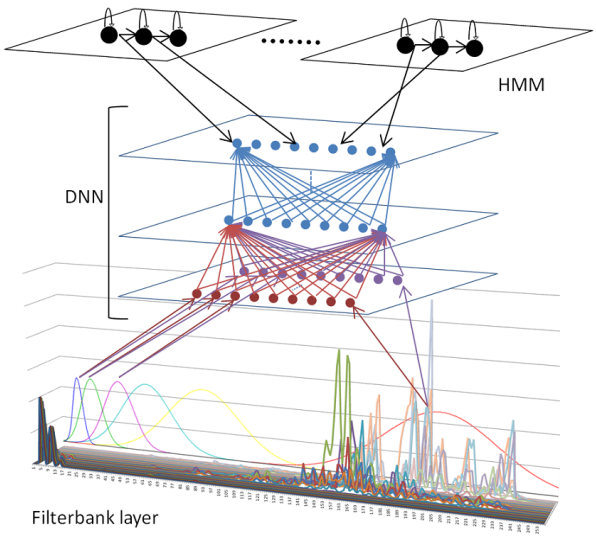


Fig. 3 Overview of Gaussian-filterbank-incorporated DNN: Filterbank weighting is performed at DNN's bottom. Horizontal axis is for the frequency bin, and vertical axis is for the power spectrum. In the experiment, input power spectra are concatenated from several consecutive frames (depth).

Figure 3 shows an overview of the Gaussian filterbank-incorporated DNN. Power spectra at frame t, $x_t(f)$, are concatenated from several consecutive frames and fed into the filterbank layer. These features are multiplied by the corresponding filter gain by Eq. (2) and summed across the frequency bin. Then applying a log-compression gives the following neural-network-based log mel-scale filterbank features:

$$h_{t,n} = \log\left(\sum_f \theta_n(f) x_t(f)\right) \tag{3}$$

$$\mathbf{h}_t = [h_{t,1}, h_{t,2}, \ldots, h_{t,n}, \ldots, h_{t,N}] \tag{4}$$

where N is the number of filters and t is the frame index. For the training of feedforward DNN, $\mathbf{h}_t$ with consecutive $\pm c$ frame features, $[\mathbf{h}_{t-c}, \ldots, \mathbf{h}_t, \ldots, \mathbf{h}_{t+c}]$, are fed into the following layer to compute the posterior probability of the triphone states [7]. We call this architecture Gaussian filterbank incorporated DNN (GFDNN).

3.1.3 Gammatone Filterbank

In this framework, arbitrary differentiable filter functions can be used as a filter shape. To compare the recognition performance among filter types, we also used a Gammatone filter, which is a widely used model as an auditory filter [35]. A Gammatone filter is modeled as

$$g_n(t) = c_n t^{a-1} \exp(-2\pi b_n t)\cos(2\pi f_0(n)t + \zeta_n), \tag{5}$$

where c_n is a constant value, ζ_n is a phase, a is an order, b_n is a temporal decay, and $f_0(n)$ is a center frequency. Equations (6) and (7) are obtained by applying the Fourier transform to Eq. (5):

$$H_n(f) = \frac{c_n}{2}(a-1)!(2\pi b_n)^{-a} \left\{ e^{(i\zeta_n)}\left[1 + \frac{i(f - f_0(n))}{b_n}\right]^{-a} + e^{(-i\zeta_n)}\left[1 + \frac{i(f + f_0(n))}{b_n}\right]^{-a}\right\} \tag{6}$$

$$\theta_n(f) = |H_n(f)|^2 \sim k_n^2 \left\{\left[1 + \frac{(f - f_0(n))^2}{b_n^2}\right]^{-a} + \left[1 + \frac{(f + f_0(n))^2}{b_n^2}\right]^{-a}\right\}, \tag{7}$$

where

$$a = 4 \tag{8}$$

$$k_n = \frac{c_n}{2}(a-1)!(2\pi b_n)^{-a} \tag{9}$$

$$\zeta_n(f) = \tan^{-1}\left\{\frac{-2f_0(n)b_n}{b_n^2 + (f^2 - f_0(n)^2)}\right\}. \tag{10}$$

The followings are the trainable parameters: k_n (gain), $f_0(n)$ (center frequency), and b_n (temporal decay). In the experiment, the initial values of $f_0(n)$ and b_n are set [36], [37]:

$$f_0(n) = -\eta + (f_{max} + \eta)\exp\left\{\frac{n \log\frac{f_{min}+\eta}{f_{max}+\eta}}{N}\right\} \tag{11}$$

$$b_n = 1.019 \times 24.7 \times \left(f_0(n) \times \frac{4.37}{100} + 1\right) \tag{12}$$

$$\eta = 228.83, \tag{13}$$

where n is an index of the filters, N is the total number of filters, and f_{min} (in Hz) and f_{max} (in Hz) are the lowest and highest cutoff frequencies of the filterbank, respectively. As

seen in Eq. (7), the Gammatone filter takes a line asymmetric curve. Mitra et al. reported the effectiveness of Gammatone filterbank features for DNN acoustic model in noisy condition (Note that our experimental condition is clean environment) [38]. The Gammatone filterbank incorporated DNN (GtFDNN) without update of filterbank corresponds to the DNN with Gammatone filterbank features.

3.1.4 Exponential Filterbank

Sainath et al. [39] proposed a method to jointly train a filterbank layer and the following networks under a restriction where the elements of the filters take positive values by introducing the exponential of weights (Exponential filterbank incorporated DNN; ExpFDNN):

$$h_{t,n} = \exp(\mathbf{w}_n)\mathbf{x}_t = \sum_f \exp(w_n(f))x_t(f), \tag{14}$$

where n is a filter index, f is a frequency bin, $\mathbf{w}_n$ is a weight vector of n, and $x(f)$ are power spectra. However, this weak restriction does not explicitly give a smoothing function, which is the original purpose of a hand-crafted triangular filterbank. In other words, the parameters of the filterbank layer overfit the given data and the shape of the filters leads to multiple peaks. Figure 4 shows an example of the actual filter shapes that were fine-tuned in the experiment. This ExpFDNN characteristic could becomes a disadvantage in adaptation. Therefore, we also trained this model for comparison.

3.2 Training Procedure

The filterbank layer parameters were trained by backpropagation. The update rule of φ_n, for example, is as follows:

$$\varphi_n^{new} = \varphi_n^{old} - \eta\frac{\partial L}{\partial \varphi_n} \tag{15}$$

$$= \varphi_n^{old} - \eta\frac{\partial L}{\partial h_n}\frac{\partial h_n}{\partial \varphi_n}, \tag{16}$$

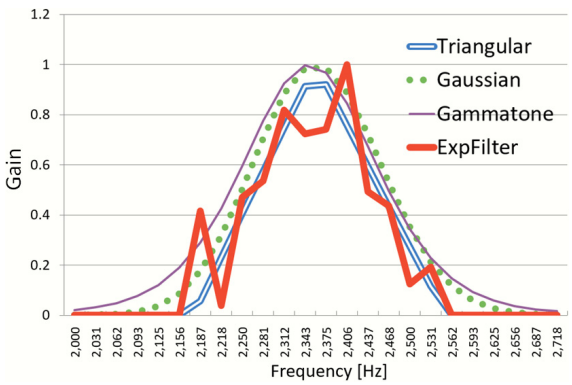


Fig. 4 Example of actual filter shapes that were fine-tuned in the experiment. Blue double line shows conventional triangular filter. Green dotted line is a Gaussian filter, and red bold line is an exponential filter.

where L is an objective function and η is the learning rate. Other parameters, β_n, γ_n (for GFDNN), k, f_0, and b (for GtFDNN), are updated in the same manner. The models are trained in two stages. First, except for the filterbank layer, the DNN is fine-tuned until a convergence criterion is met. Then the filterbank layer and the following DNN are trained jointly with the same initial learning rate.

3.3 Characteristics of Filterbank-Incorporated DNN

The filterbank-incorporated DNN has some advantages compared with earlier studies.

- The proposed method can compute neural network-based log mel-scale filterbank features.
- Unlike the fully connected layer, the proposed system performs framewise transformation. Each filter takes a certain portion of the power spectra. The initial center frequency and bandwidth values are described in Sect. 4.1.2.
- An adjustment of the gain parameter corresponds to fMLLR [40]. An adjustment of the center frequency parameter corresponds to VTLN [41] by regarding the frequency shift as frequency warping. In summary, our proposed system has fMLLR and VTLN capability while minimizing the number of free parameters.
- The shapes of the filters are adapted in a discriminative manner using backpropagation.
- The filterbank layer, which consists of a small number of parameters, is effective for adaptation under limited available data while fully neural network based architecture suffers from the overfitting problem (e.g. time-domain convolution layer in [2] has 16,000 parameters).

We considered whether there is a relation between the learned center frequencies and the vocal tract length. A vocal tract's average length depends on gender and age. The average length of the vocal tracts of Japanese adult males and females is 17.0 cm and 15.0 cm, respectively. The average length of the vocal tracts of children is 9.0 cm [42]. Theoretically, the spectra of female speakers shift to an approximately 11.8% (15.0/17.0) higher frequency domain from that of male speakers due to the differences of vocal tract length. Therefore, we assume that the center frequencies of the filterbank layer shift to a 11.8% higher frequency domain by adapting a filterbank layer of a male-specific DNN using female speech data[†].

3.4 Adaptation of Proposed and Compared Methods

In the experiment, we conducted supervised adaptation to evaluate the effectiveness of the filterbank-incorporated DNNs. During adaptation, the parameters of the filterbank

[†]The shift (warping) of frequencies in VLTN is also accomplished by the adjustment of channel gains. The function of VTLN is executed by both shift of center frequencies and scale of filter gains.

Table 1 Comparison of number of parameters updated in adaptation.

Target of adaptation	Parameters
GFDNN	120
GtFDNN	120
ExpFDNN	10,240
fDLR	1,600
LHUC	2,048
SVD	420

layer that present the filter shape/center frequency are updated to minimize the cross-entropy loss function at the level of the triphone states. During the adaptation of GFDNN, GtFDNN, fDLR, and ExpFDNN, the bottom layers of the models were updated. Table 1 summarizes the number of parameters that were updated in the model adaptation. The dimension of power spectra and filterbank feature are set to 256 and 40 respectively. The number of parameters for GFDNN and GtFDNN was 120 (40 filters × 3 parameters).

We also used LHUC- and SVD-based adaptation in our experiment for comparison.

3.4.1 LHUC

The LHUC rescales the hidden units of the l-th layer as following equation:

$$h_j^l = 2\sigma(r_j^{l,s}) \cdot \psi(\mathbf{w}_j^l \mathbf{h}^{l-1} + b_j^l), \tag{17}$$

where j is an index of the units, $\mathbf{w}_j^l$ is a weight vector, and $\psi(\cdot^l)$ computes the l-th hidden units. They are rescaled by applying element-wise multiplication with $\sigma(\cdot)$ ranging from 0.0 to 1.0. Variables $r_j^{l,s}$ are optimized by each speaker s.

3.4.2 SVD

The SVD was applied to the matrix of l-th hidden layer:

$$\mathbf{w}_{m,n}^l = U_{m,n}^l \Sigma_{n,n}^l (V_{n,n}^l)^T \tag{18}$$

$$\approx U_{m,k}^l \Sigma_{k,k}^l (V_{n,k}^T)^T \tag{19}$$

where Σ is a diagonal matrix of singular values, and subscript is the size of matrix. Adaptation of decomposed diagonal matrix and further selection of k singular values decrease the number of free parameters. The singular values in the diagonal matrix, $\Sigma_{k,k}^l$, were updated in the experiment.

4. Experiments

4.1 Experimental Setup

4.1.1 Corpora

The details of the Corpus of Spontaneous Japanese (CSJ) [43] are shown in Table 2. It consists of 186 hours of speech of male speakers (SM) and 42 hours of speech of female speakers (SF). We used an attached evaluation set-2 for

Table 2 Details of CSJ corpus.

	Gender	Male (SM)	Female (SF)
Train	Lectures	787	166
	Data	186 hours	42 hours
Test	Lectures	5	5
	Data	1.0 hours	0.9 hours

the evaluation that consists of five male speakers and five female speakers. We used all utterances of the evaluation set-2 as test data for speaker-independent experiment and gender adaptation experiment. In case of speaker-independent experiment, we trained SM-specific, SF-specific, and gender independent models and tested the models using the gender matched test data. In case of gender adaptation experiment, we adapted the SM-specific models using the training data of female speakers, and tested the models using female speakers in the evaluation set-2. In case of speaker adaptation experiment, we assigned 20 utterances to the adaptation data and 40 utterances to the test data. The SM-specific models were trained by male speakers and they were adapted/tested by 5 male speakers in the evaluation set-2.

The speech was analyzed using a 25-ms Hamming window with a pre-emphasis coefficient of 0.97 and shifted with a 10-ms frame advance.

4.1.2 Acoustic Model

We built a hybrid DNN-HMM system. For an experiment of speaker and gender adaptations, we implemented some regular model adaptation techniques. The following is the experimental setup of GMM-HMM and DNN:

GMM-HMM
To obtain the training target labels for DNNs, GMM-HMMs are trained using a corpus of SM, SF, and SM plus SF (mixed). The models are trained on standard MFCC features. The senones of SM, SF, and SM plus SF are 4783, 4860, and 5023, respectively. Corresponding GMM-HMMs are used for forced alignment.

Baseline DNN (Triangle filterbank)
As a baseline system, we trained a fully connected DNN, which has five hidden layers with 2,048 rectified linear units [44]. Its input is 11 consecutive frames of 40-dimensional log mel-scale triangle-shape filterbank features extracted using the Hidden Markov Model Toolkit (HTK) [45]. The features are normalized to zero mean and unit variance. Due to the fixed and undifferentiable shape of the triangular filters, the filter shapes are unchanged at all times.

Gaussian filterbank incorporated DNN (GFDNN)
The Gaussian filterbank layer (Eq. (3)) was inserted into the bottom of the baseline DNN (Fig. 3). Its input was 11 consecutive frames of 256-dimensional power spectra. The number of filters was set to 40, which is the same as the baseline system ($n = 1, 2, \ldots, 40$). The initial values of the gain parameter were set to 1.0. The center frequencies were spaced equally along the

Table 3 WERs [%] of baseline DNN and filterbank-incorporated DNNs.

System	WER [%]			
	SM	SF	Ave.	SM+SF
Baseline (Triangle)	12.4	20.4	16.4	13.4
GFDNN (fixed)	12.4	18.8	15.6	12.5
GFDNN (trained)	12.5	19.0	15.8	12.9
GtFDNN (fixed)	12.1	16.3	14.2	12.6
GtFDNN (trained)	12.1	15.9	14.0	12.6
ExpFDNN	12.3	17.0	14.7	12.9

mel-scale. The bandwidths were set so that the two-sigma range equals the corresponding bandwidth of the mel-scale filterbank. At the Gaussian filterbank layer, 120 parameters, which consist of φ (gains), γ (center frequencies), and β (bandwidths), were updated using backpropagation.

Gammatone filterbank incorporated DNN (GtFDNN)
The Gammatone filterbank layer (Eq. (7)) was inserted into the bottom of the baseline DNN. The initial values were set according to Eqs. (11), (12), and (13). The other setup is the same as GFDNN. At the Gammatone filterbank layer, 120 parameters were updated using backpropagation.

ExpFDNN
A DNN with an exponential filterbank layer [39] was trained for comparison. The initial values of the filterbank layer were set similar to the triangular filterbank. At the filterbank layer, 10,240 parameters (256 frequency bins of 40 filters) are updated using backpropagation. The other setup is identical as the GFDNN.

feature-space disctiminative linear regression (fDLR)
A linear layer was inserted into the bottom of the network for the baseline DNN, and it was inserted after the filterbank layer for the GtFDNN. The identity matrix was 40 by 40.

Learning hidden unit contribution (LHUC)
To adapt the models, we rescaled the hidden units using LHUC. In the experiment, the hidden units of the third layer were adapted which showed the best performance in the preliminary experiment. The number of parameters was 2,048.

Singular value decomposition (SVD)
We applied SVD on the 1st fully connected layer and kept top 420 singular values. These values were decided from the best performance.

Comparative adaptation methods, fDLR, LHUC, and SVD, are applicable to GtFDNN since the training of filterbank layer and adaptation of hidden layer is independent. Therefore, we applied comparative speaker adaptation methods on baseline (triangular shape filter) DNN and GtFDNN. Filterbank layer was un-adapted when the comparative methods are applied to GtFDNN.

We used Chainer [46] for training the DNNs. The models were trained using Adam [47] with batch normalization [48]. The 1% of the training data was used for the model selection. For training the GFDNN and GtFDNN, we first trained the DNN while fixing the filterbank parameters. Hereinafter, we refer to this model as a fixed model. Next, we trained the DNNs with a filterbank layer. Hereinafter, we refer to this model as a trained model. We followed the existing Kaldi recipe [49] for training GMM-HMM and decoding.

4.2 Results

4.2.1 Speaker Independent Model

The performance of the speaker-independent models are shown in Table 3. The baseline gender independent DNN, which takes triangular filterbanks, achieved an average WER of 13.4%. When we focus on the baseline models and the *fixed* (untrained) models, the latter outperformed the baseline models in all cases, even though the filter shape is the only difference between the two models. This difference changes the coverage of the frequency bin. The Gaussian and Gammatone filters focus on all the frequency bins while the baseline triangular filter zeroes out the frequency bins outside a certain bin distance. These results comparing fixed and baseline model show the importance of refined acoustic features. Mitra et al. also investigated the effectiveness of robust features for DNN including Gammatone filterbank [38]. The performance improvement of the fixed models correspond with their results. The Gammatone filter is widely used as an auditory filter. However, the difference between filter types did not show any performance gain.

In comparison with the fixed models, the trained models did not show performance improvement. We considered that the difference of optimal center frequencies between male and female speakers made it difficult to learn universal center frequencies for both male and female speakers. In the following experiment, we only present the results of the trained models.

4.2.2 Gender Adaptation

In this section, we performed gender adaptation from SM to SF as shown in Table 4 to confirm the presence of the filters' shift to alleviate the vocal tract length. The first column is the duration of SF speech for adaptation. The row of 0 utterance is the WERs of the model without adaptation. The WERs of SM specific GFDNN and GtFDNN were worse at 26.8% and 24.4%, due to the gender mismatched condition. For evaluations of 10 and 20 utterances, 60 utterances in Table 4 were split into six or three folds, and averaged to alleviate any selection bias. In the scenario of limited adaptation data, the best performance was obtained when we adapted the filterbank layer of GFDNN. We considered that the focus on filter adaptation worked on the alleviation of gender mismatch effectively while discarding other mismatched conditions that are difficult to adapt under limited data. When the adaptation data increased to 0.5 hour and more, the best model was replaced by LHUC which has larger free parameters.

By adapting GFDNN from SM to SF, we considered

Table 4 WERs [%] of gender adaptation from adult male speakers to adult female speakers. Bold is the best performance among models.

Adaptation data	# speakers	GFDNN	GtFDNN				Baseline DNN (Triangle)			ExpFDNN
		filterbank	filterbank	fDLR	LHUC	SVD	fDLR	LHUC	SVD	filterbank
0.0 h	0	26.8	**24.4**	**24.4**	**24.4**	28.2	26.5	26.5	26.5	25.2
0.02 h (72 seconds)	20	**22.3**	22.7	23.3	22.9	28.2	26.0	32.8	28.6	23.6
0.03 h (108 seconds)	30	**18.9**	20.1	22.0	20.9	27.2	27.1	32.2	28.2	20.4
0.1 h (360 seconds)	51	**16.4**	16.5	17.8	20.3	22.3	19.1	31.8	23.7	17.2
0.5 h (1800 seconds)	166	15.7	14.9	17.0	**13.9**	18.2	17.4	31.2	18.2	15.7
1.0 h	166	15.4	15.4	16.5	**13.8**	14.6	19.4	31.6	16.2	14.1
10.0 h	166	14.9	15.6	16.1	**14.2**	14.3	16.6	31.2	14.3	14.6
30.0 h	166	15.0	15.2	16.1	**13.6**	14.0	16.6	31.4	14.1	14.7

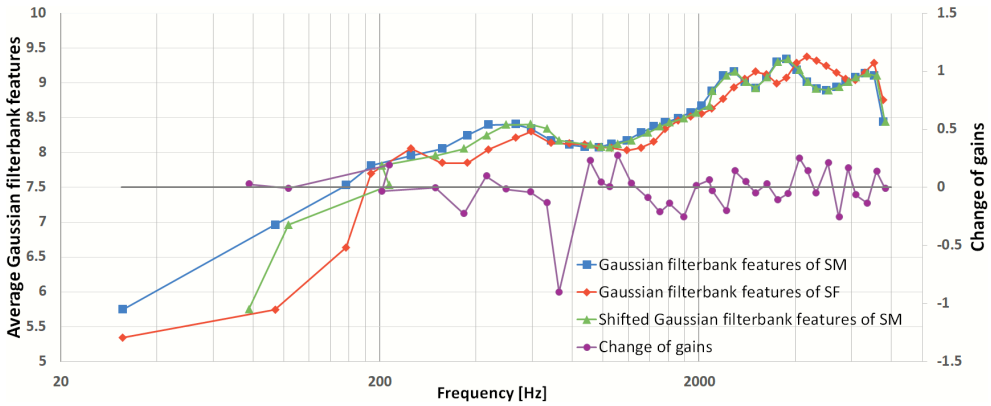


Fig. 5 Changes of gain parameters from SM-specific model (SM) to SF-adapted model and averaged Gaussian filterbank features of SM- and SF-speakers.

Table 5 Shift of center frequencies [Hz] caused by gender adaptation from SM to SF speakers using 10 hours of data. SM → SF shows center frequencies of unadapted and adapted models. SF column shows center frequencies of SF-specific model trained using SF speech data.

	SM → SF			SF
n	Before Adaptation	After Adaptation	Difference	-
6	315.8	369.7	54.1 (17.2%)	310.5
7	382.6	438.1	55.5 (14.5%)	374.3
8	449.4	508.1	58.7 (13.1%)	439.3
9	530.6	592.0	61.4 (11.6%)	524.3
10	597.2	669.0	71.8 (12.0%)	589.5
11	701.1	741.1	39.9 (5.7%)	687.1
12	783.9	912.9	129.1 (16.5%)	783.7
13	866.1	976.1	110.0 (12.7%)	874.9
14	953.6	1033.9	80.4 (8.4%)	965.4
15	1055.1	1103.9	48.7 (4.6%)	1061.5

that the frequency shift of the filters are caused by the differences of the vocal tract lengths. Table 5 shows the relation among the center frequencies of SM-dependent GFDNN, adapted GFDNN from SM to SF using 10 hours of data, and SF-dependent GFDNN. Theoretically, an ideal frequency shift is approximately 11.8%, as described in Sect. 3.3. The column of difference shows the actual shift was approximately 4.6% to 17.2%, which resembles the theoretical value at low- and middle-frequency region (300Hz ∼ 1000Hz). These results show that the optimization of the filterbank layer causes a shift of the center frequencies to discriminatively perform frequency warping. This characteristic corresponds to the VTLN function.

The last column of Table 5, SF, shows the center frequencies of the SF-dependent GFDNN. When we focus on the SM- and SF-dependent GFDNNs, the relation between the two models cannot be observed in the experiment. Instead, the learned center frequencies based on SF speakers showed lower frequencies than those of the SM speakers at $n = 6 \sim 12$ because the optimal position of the filters in the training stage depends of the condition of the following DNN. However, in the adaptation stage, the filterbank layer was updated, and the parameters of the following DNN were fixed. In this situation, the filterbank layer can be handled independently of the following DNN to perform frequency warping.

Figure 5 shows the gender-dependent Gaussian filterbank features and the change of gain parameters. The square markers and diamond markers show the Gaussian filterbank features of SM and SF speakers, respectively. The triangle markers show the SM-dependent features with horizontal shifting according to the change of center frequencies caused by gender adaptation to SF-speakers from SM-speakers. We can see that the features of SM at low-frequency region shifted toward the ones of SF speakers. Next, the change of gain parameters are depicted at the right vertical axis. To emphasize the conspicuous change of gains, their relative changes are plotted with circle markers by computing $\log(gains\_of\_SF / gains\_of\_SM)$. In the low-frequency region, the change of gain was relatively small while the shift of center frequency was remarkable. Con-

Table 6 WERs [%] of the triphone level supervised speaker adaptation. Bold is the best performance among models.

#utt	GFDNN	GtFDNN				Baseline DNN (Triangle)			ExpFDNN
	filterbank	filterbank	fDLR	LHUC	SVD	fDLR	LHUC	SVD	filterbank
0	12.5	**12.1**	**12.1**	**12.1**	13.2	12.4	12.4	13.4	12.3
1	12.3	**12.2**	13.6	13.3	13.1	36.0	13.2	13.3	12.4
2	**12.2**	12.9	13.6	14.1	13.1	15.4	12.8	13.4	12.7
3	12.5	12.8	13.6	14.4	13.2	12.7	**12.4**	13.4	13.1
4	**12.4**	12.6	13.2	14.1	13.1	13.1	12.5	13.3	13.0
5	**12.0**	12.3	13.4	13.9	13.0	13.0	12.3	13.2	12.8
10	**11.4**	**11.4**	13.4	12.9	12.8	13.0	12.4	12.6	11.7
15	**11.2**	**11.2**	13.4	12.5	12.6	12.2	12.0	12.7	11.4
20	11.4	11.3	13.3	12.1	12.5	12.7	11.9	12.5	**11.2**

versely, the shift of center frequency was relatively small at a high-frequency region (∼ 2000Hz). The variances of the filterbank features are relatively large enough to overlap the SM- and SF-speakers. Therefore, the optimization of the gain parameters is a secondarily important factor in gender adaptation. In contrast to the above two parameters, no discrimination of the change of the bandwidth parameters was observed in gender adaptation.

4.2.3 Speaker Adaptation

Table 6 shows the supervised adaptation result. The models trained by SM in Table 3 were used as source models. The 0 utterance row shows the WERs, which were recognized using the model without adaptation. By adapting the filterbank layer of GtFDNN using 15 utterances, WER was improved from 12.1% to 11.2%, and a word error reduction rate (WERR) of 7.4% was obtained. This WERR is better than the unadapted GtFDNN at a significance level of 0.005 under a statistical sign test. These results show that the adjustment of the filter shapes can handle the diversity of speakers.

Performance gains were observed when adaptation was applied for GFDNN with 5 utterances ($p < 0.03$) and GtFDNN with 10 utterances ($p < 0.005$). These results are better than the other adaptation methods, although baseline DNN with LHUC and ExpFDNN also showed a performance improvement when more than 15 or 10 utterances are available for each method. Table 6 also shows the adaptation result using fDLR, LHUC, and SVD under the same GtFDNN. The adaptation of filterbank layer obtained the best performance on all conditions of adaptation utterances among other adaptation methods.

Finally, we depicted the relation between adaptation utterances and WERs per speaker in GtFDNN in Fig. 6. The WERs of almost all the speakers decreased linearly over the adaptation utterances. However, Speaker 1 showed unexpected behavior when the value of the horizontal axis was 2 to 5.

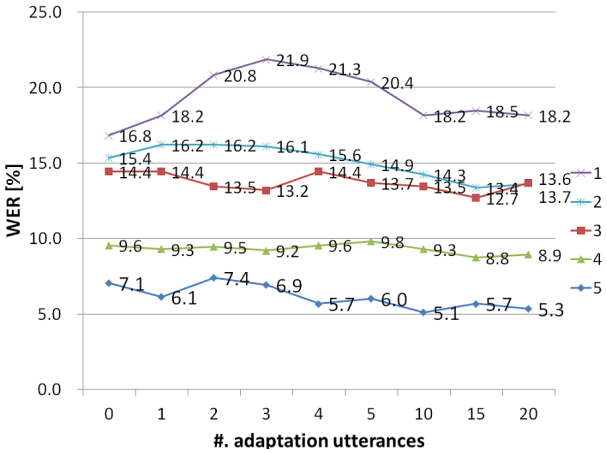


Fig. 6 Relation between number of adaptation utterances and WERs per speaker (Speaker 1 to Speaker 5). GtFDNN in Table 6 is used for error analysis.

5. Conclusions

In this paper, we evaluated a filterbank-incorporated DNN which has a filterbank layer at the bottom of the DNN. Compared with the baseline DNN, which uses log mel-scale triangular shape filterbank features as its input, the proposed method can discriminatively learn optimal filter shapes. When we carried out speaker adaptation, we found that filterbank-incorporated DNNs showed effectiveness in speaker adaptation under limited adaptation data. We also carried out gender adaptation from male to female speakers and discussed the relation between the physical characteristics of the vocal tract length and an optimal filterbank from an engineering viewpoint.

The filterbank layer is a simple module of a neural network and can be combined with other modules, e.g. CNNs, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) [50]. In future work, we will conduct an evaluation under noisy conditions.

Acknowledgments

The present work was supported (in part) through the Leading Graduate School Program, "Innovative program for training brain-science-information-architects by analysis of massive quantities of highly technical information about the brain," by the Ministry of Education, Culture, Sports, Science and Technology. This work was also supported by JSPS (Japan Society for the Promotion of Science) KAKE-

NHI Grand Number 25280062 and 15K00233.

References

[1] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-R. Mohamed, N. Jitaly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol.29, no.6, pp.82–97, 2012.

[2] T.N. Sainath, R.J. Weiss, A. Senior, K.W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," Proc. Interspeech, pp.1–5, 2015.

[3] H.B. Sailor and H.A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," IEEE Trans. Audio, Speech, Language Process., vol.24, no.12, pp.2341–2353, 2016.

[4] Z. Zhu, J.H. Engel, and A. Hannun, "Learning multiscale features directly from waveforms," Proc. Interspeech, pp.1305–1309, 2016.

[5] Z. Chen, S. Watanabe, H. Erdogan, and J.R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," Proc. Interspeech, pp.3274–3278, 2015.

[6] Z.Q. Wang and D. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," Proc. Interspeech, pp.2839–2843, 2015.

[7] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," Proc. ICASSP, pp.5480–5484, 2017.

[8] G. Saon, H. Saltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," Proc. ASRU, pp.55–59, 2013.

[9] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," IEEE/ACM Trans. Audio, Speech, Language Process., vol.22, no.12, pp.1713–1725, 2014.

[10] H. Seki, D. Enami, F. Zhu, K. Yamamoto, and S. Nakagawa, "Speech recognition of short time utterance based on speaker clustering," IEICE Trans. Inf. & Syst., vol.J100-D, no.1, pp.81–92, Jan. 2017 (in Japanese).

[11] S.H.K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fMLLR based feature-space speaker adaptation of DNN acoustic models," Proc. Interspeech, 2015.

[12] T.N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," Proc. ICASSP, pp.8614–8618, 2013.

[13] S.P. Rath, D. Povey, K. Veselý, and J.H. Černocký, "Improved feature processing for deep neural networks," Proc. Interspeech, pp.109–113, 2013.

[14] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," Proc. ASRU, pp.24–29, 2011.

[15] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," Proc. EUROSPEECH, pp.2171–2174, 1995.

[16] B. Li and K.C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," Proc. Interspeech, pp.526–529, 2010.

[17] S. Xue, H. Jiang, L. Dai, and Q. Liu, "Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition," Journal of Signal Processing Systems, vol.82, no.2, pp.175–185, 2016.

[18] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," IEEE Trans. Audio, Speech, Language Process., vol.24, no.8, pp.1450–1463, 2016.

[19] Y. Zhao, J. Li, J. Xue, and Y. Gong, "Investigating online low-footprint speaker adaptation using generalized linear regression and click-through data," Proc. ICASSP, pp.4310–4314, 2015.

[20] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," Proc. ICASSP, pp.7893–7897, 2013.

[21] D. Saito, N. Minematsu, and K. Hirose, "Rotational properties of vocal tract length difference in cepstral space," Journal of Research Institute of Signal Processing, vol.15, no.5, pp.363–374, 2011.

[22] D.H.H. Nguyen, X. Xiao, E.S. Chng, and H. Li, "Feature adaptation using linear spectro-temporal transform for robust speech recognition," IEEE Trans. Audio, Speech, Language Process., vol.24, no.6, pp.1006–1019, 2016.

[23] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE Trans. Audio, Speech, Language Process., vol.22, no.10, pp.1533–1545, 2014.

[24] K. Kameyama, K. Mori, and Y. Kosugi, "A neural network incorporating adaptive Gabor filters for image texture classification," International Conference on Neural Networks, pp.1523–1528, 1997.

[25] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," IEEE/ACM Trans. Audio, Speech, Language Process., vol.24, no.3, pp.459–468, 2016.

[26] L. Samarakoon and K.C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," IEEE/ACM Trans. Audio, Speech, Language Process., vol.24, no.12, pp.2241–2250, 2016.

[27] H.N.B. Pinheiro, F.M.P. Neto, A.L.I. Oliveira, T.I. Ren, G.D.C. Cavalcanti, and A.G. Adami, "Optimizing speaker-specific filter banks for speaker verification," Proc. ICASSP, pp.5350–5354, 2017.

[28] C. Charbuillet, B. Gas, M. Chetouani, and J.L. Zarader, "Filter bank design for speaker diarization based on genetic algorithms," Proc. ICASSP, pp.673–676, 2006.

[29] T. Kobayashi and J. Ye, "Discriminatively learned filter bank for acoustic features," Proc. ICASSP, pp.649–653, 2016.

[30] L. Burget and H. Heřmanský, "Data driven design of filter bank for speech recognition," Text, Speech and Dialogue, Lecture Notes in Computer Science, vol.2166, pp.299–304, Springer, 2001.

[31] Y. Suh and H.R. Kim, "Data-driven filter-bank-based feature extraction for speech recognition," Proc. SPECOM, pp.154–157, 2004.

[32] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," Proc. ICASSP, pp.2721–2725, 2017.

[33] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks," Proc. ICASSP, pp.791–795, 2017.

[34] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An application of discriminative feature extraction to filter-bank-based speech recognition," IEEE Trans. Audio, Speech, Language Process., vol.9, no.2, pp.96–110, 2001.

[35] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," A Meeting of the IOC Speech Group on Auditory Modelling at RSRE, vol.2, no.7, pp.1–18, 1987.

[36] A. Adiga, M. Magimai, and C.S. Seelamantula, "Gammatone wavelet cepstral coefficients for robust speech recognition," TENCON IEEE Region 10 Conference (31194), pp.1–4, 2013.

[37] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," IEEE International Conference on Circuits and Systems (ISCAS), pp.305–308, 2013.

[38] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions," Proc. Interspeech, pp.895–899, 2014.

[39] T.N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," Proc. ASRU, pp.297–302, 2013.

[40] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech & Language,

vol.10, no.4, pp.249–264, 1996.
[41] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, no.CMU-CS-97-148, 1997.
[42] A. Behrman, Speech and Voice Science, Plural Publishing, 2007.
[43] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.
[44] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. International Conference on Artificial Intelligence and Statistics, pp.315–323, 2011.
[45] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book, University of Cambridge.
[46] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: A next-generation open source framework for deep learning," Proc. NIPS, 2015.
[47] D.P. Kingma and J.L. Ba, "ADAM: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. ICML, pp.448–456, 2015.
[49] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, no.EPFL-CONF-192584, 2011.
[50] D. Yu, and J. Li, "Recent progresses in deep learning based acoustic models," IEEE/CAA Journal of Automatica Sinica, vol.4, no.3, pp.396–409, 2017.

Hiroshi Seki graduated from Toyohashi University of Technology with his Bachelor's and Master's degrees in 2014 and 2016. Since 2016, he has been studying at Toyohashi University of Technology as a doctoral student.

Kazumasa Yamamoto received his B.E., M.E. and Dr. Eng. degrees in Information and Computer Sciences from Toyohashi University of Technology, Toyohashi, Japan, in 1995, 1997 and 2000. From 2000 to 2007, he was a research associate in the Department of Electrical and Electronic Engineering, Shinshu University, Nagano, Japan. In 2007, he joined Toyohashi University of Technology. From 2007 to 2012, he was an assistant professor, and from 2013 to 2017, he was an associate professor in the Department of Computer Science and Engineering. In 2013, he was transferred temporary as an associate professor to the Department of Information and Computer Engineering, National Institute of Technology, Toyota College, Toyota, Japan. Since 2017, he has been an associate professor in the Department of Computer Science, Chubu University, Kasugai, Japan. In 2012, he was a visiting researcher in the Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, USA. His major research interests include speech recognition, spoken dialogue system and sound signal processing. He is a member of IEEE, ISCA, IEICE, IPSJ and ASJ.

Tomoyosi Akiba received the B.E. degree in 1990 in information science, the M.E. degree in 1992, and the Ph.D. degree in 1995 in system science from Tokyo Institute of Technology, Tokyo, Japan. In 1995, he became a Researcher in the Electrotechnical Laboratory, MITI, Japan. In 2004, he became an Associate Professor in Toyohashi University of Technology. His research interests include natural language processing and spoken language processing. He is a member of ISCA, Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), Acoustical Society of Japan (ASJ), and the Association for Natural Language Processing (NLP).

Seiichi Nakagawa received Dr. Eng. degree from Kyoto University, Kyoto, Japan, in 1977. He joined the Faculty of Kyoto University, in 1976, as a Research Associate in the Department of Information Sciences. In 1980, he joined Toyohashi University of Technology, which was founded in 1977. From 1980 to 1983, he was an Assistant Professor, and from 1983 to 1990 he was an Associate Professor. From 1990 to 2014, he was a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan. He retired from the university in 2014, and he is now Special Appointment Professor of "Organization for Leading-Graduate-School Program." Since 2017, he is also professor of Department of Computer Science, Cubu Univercity. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. His major research interests include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He received the 1997/2001/2013 Best Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronics and Telecommunication Engineers. He received the Prize for Distinguished Achievements in Acoustics (ASJ, 2015) and the Achievement Award from IEICE (2015). He was the founder of the Special Interest Group of Spoken Language Processing (SLP) of the Information Processing Society of Japan in 1994. He is a Fellow of IEICE and IPSJ.