Audio-visual speech recognition using deep learning

Kuniaki Noda · Yuki Yamaguchi ·
Kazuhiro Nakadai · Hiroshi G. Okuno · Tetsuya Ogata

Abstract Audio-visual speech recognition (AVSR) system is thought to be one of the most promising solutions for reliable speech recognition, particularly when the audio is corrupted by noise. However, cautious selection of sensory features is crucial for attaining high recognition performance. In the machine-learning community, deep learning approaches have recently attracted increasing attention because deep neural networks can effectively extract robust latent features that enable various recognition algorithms to demonstrate revolutionary generalization capabilities under diverse application conditions. This study introduces a connectionist-hidden Markov model (HMM) system for noise-robust AVSR. First, a deep denoising autoencoder is utilized for acquiring noise-robust audio features. By preparing the training data for the network with pairs of consecutive multiple steps of deteriorated audio features and the corresponding clean features, the network is trained to output denoised audio features from the corresponding features deteriorated by noise. Second, a convolutional neural network (CNN) is utilized to extract visual features from raw mouth area images. By preparing the training data for the CNN as pairs of raw images and the corresponding phoneme label outputs, the network is trained to predict phoneme labels from the corresponding mouth area input images. Finally, a multi-stream HMM (MSHMM) is applied for integrating the acquired audio and visual HMMs independently trained with the respective features. By comparing the cases when normal and denoised mel-frequency cepstral coefficients (MFCCs) are utilized as audio features to the HMM, our unimodal isolated word recognition results demonstrate that approximately 65 % word recognition rate gain is attained with denoised MFCCs under 10 dB signal-to-noise-ratio (SNR) for the audio signal input. Moreover, our multimodal isolated word recognition results utilizing MSHMM with denoised MFCCs and acquired visual features demonstrate that an additional word recognition rate gain is attained for the SNR conditions below 10 dB.

K. Noda (✉) · T. Ogata
Graduate School of Fundamental Science and Engineering,
Waseda University, Tokyo 169-8555, Japan
e-mail: kuniaki.noda@akane.waseda.jp

T. Ogata
e-mail: ogata@waseda.jp

Y. Yamaguchi · H. G. Okuno
Graduate School of Informatics, Kyoto University,
Kyoto 606-8501, Japan
e-mail: yamaguch@kuis.kyoto-u.ac.jp

H. G. Okuno
e-mail: okuno@kuis.kyoto-u.ac.jp

K. Nakadai
Honda Research Institute Japan Co., Ltd., Saitama
351-0114, Japan
e-mail: nakadai@jp.honda-ri.com

1 Introduction

Human–machine interfaces for intelligent machines, such as smartphones, domestic robots, and auto-driving cars, are expected to become increasingly common in everyday life. Consequently, noise-robust speech recognition will become crucial for achieving effective human–machine interaction. Audio-visual speech recognition (AVSR) is thought to be one of the most promising solutions for reliable speech recognition, particularly when the audio is corrupted by

noise. The fundamental idea of AVSR is to use visual information derived from a speaker's lip motion to complement corrupted audio speech inputs. However, cautious selection of sensory features for the audio and visual inputs is crucial in AVSR because sensory features significantly influence the recognition performance.

Regarding sensory feature extraction mechanisms, deep learning approaches have recently attracted increasing attention among the machine-learning community [5]. For example, deep neural networks (DNNs) have successfully been applied to unsupervised feature learning for single modalities such as text [47], images [25], and audio [17]. The same approach has also been applied to the learning of fused representations over multiple modalities, resulting in significant improvements in speech recognition performance [37]. Expanding on ideas from recent successes in deep learning studies, we propose to apply two major DNN architectures, deep denoising autoencoder [48, 49] and convolutional neural network (CNN) [27], for audio feature extraction and visual feature extraction, respectively.

Audio feature extraction by a deep denoising autoencoder is achieved by training the network to predict original clean audio features, such as mel-frequency cepstral coefficients (MFCCs), from deteriorated audio features that are artificially generated by superimposing various strengths of Gaussian noises to original clean audio inputs. Acquired audio feature sequences are then processed with a conventional hidden Markov model (HMM) with a Gaussian mixture observation model (GMM-HMM) to conduct an isolated word recognition task. The main advantage of our audio feature extraction mechanism is that noise-robust audio features are easily acquired by a rather simple mechanism.

For the visual feature extraction mechanism, we propose application of a CNN, one of the most successfully utilized neural network architectures for image clustering problems. This is achieved by training the CNN with over a hundred thousand mouth area image frames in combination with corresponding phoneme labels. CNN parameters are learned in order to maximize the average across training cases for the log-probability of the correct label under the prediction distribution. Through supervised training, multiple layers of convolutional filters, which are responsible for extracting primitive visual features and predicting phonemes from raw image inputs, are self-organized. Our visual feature extraction mechanism has two main advantages: (1) the proposed model is easy to implement because dedicated lip-shape models or hand-labeled data are not required; (2) the CNN has superiority in shift- and rotation- resistant image recognition.

To perform an AVSR task by integrating both audio and visual features into a single model, we propose a multi-stream hidden Markov model (MSHMM) [6, 7, 20]. The main advantage of the MSHMM is that we can explicitly select the observation information source (i.e., from audio input to visual input) by controlling the stream weights of the MSHMM depending on the reliability of multimodal inputs. Our evaluation results demonstrate that the isolated word recognition performance can be improved by utilizing visual information, especially when audio information reliability is degraded. The results also demonstrate that the multimodal recognition attains an even better performance than when audio and visual features are separately utilized for isolated word recognition tasks.

The remainder of this study is organized as follows. In Section 2, we briefly review related work on audio and visual feature extraction mechanisms for automatic speech recognition (ASR). In Section 3, we describe several learning algorithms for various deep neural network architectures, including deep autoencoder and CNN. In Section 4, we describe the audiovisual dataset utilized for the evaluation experiments of our proposed speech recognition models. In Section 5, we introduce the general framework of the proposed AVSR system and describe implementations for audio feature extraction, visual feature extraction, and audio-visual integration. The proposed frameworks are evaluated in Section 6. We conduct isolated word recognition experiments to evaluate audio and visual features and integrated inputs. In Section 7, we discuss our current results and describe directions for future work. Conclusions are presented in Section 8.

2 Related work

2.1 Audio feature extraction mechanisms

The use of MFCCs has been a de facto standard for ASR for decades. However, advances in deep learning research have led to recent breakthroughs in unsupervised audio feature extraction methods and exceptional recognition performance improvements [13, 17, 32]. Advances in novel machine learning algorithms, improved availability of computational resources, and the development of large databases have led to self-organization of robust audio features by efficient training of large-scale DNNs with large-scale datasets.

One of the most successful applications of DNNs to ASR is the deep neural network hidden Markov model (DNN-HMM) [12, 36], which replaces the conventional Gaussian mixture model (GMM) with a DNN to represent the direct projection between HMM states and corresponding acoustic feature inputs. The idea of utilizing a neural network to replace a GMM and construct a hybrid model that combines a multilayer perceptron and HMMs was originally proposed decades ago [8, 41]. However, owing to limited computational resources, large and deep models were not

experimented with in the past, which led to hybrid systems that could not outperform GMM-HMM systems.

Other major approaches for application of DNNs to ASR involve using a deep autoencoder as a feature extraction mechanism. For example, Sainath et al. utilized a deep autoencoder as a dimensionality compression mechanism for self-organizing higher-level features from raw sensory inputs and utilized the acquired higher-level features as inputs to a conventional GMM-HMM system [43]. Another example is the deep denoising autoencoder proposed by Vincent et al. [48, 49]. This model differs from the former model in that the outputs of the deep autoencoder are utilized as the sensory feature rather than the compressed vectors acquired from the middle layer of the network. The key idea of the denoising model is to make the learned representations robust to partial destruction of the input by training a deep autoencoder to reconstruct clean repaired input from corrupted, partially destroyed input. We adopt the deep denoising autoencoder for acquiring noise-robust audio features by training the network to reconstruct clean audio features from deteriorated ones. In this study, the acquired denoised audio features are processed with a GMM-HMM system. The primary reason for utilizing a GMM-HMM rather than DNN-HMM is to apply the MSHMM seamlessly as a multimodal integration mechanism for the subsequent AVSR task.

2.2 Visual feature extraction mechanisms

Incorporation of speakers' lip movements as visual information for ASR systems is known to contribute to robustness and accuracy, especially in environments where audio information is corrupted by noise. In previous studies, several different approaches have been proposed for extracting visual features from input images [24, 34]. These approaches can be broadly classified into two representative categories.

The first is a top-down approach, where an a priori lip-shape representation framework is embedded in a model; for example, active shape models (ASMs) [31] and active appearance models (AAMs) [11]. ASMs and AAMs extract higher-level, model-based features derived from the shape and appearance of mouth area images. Model-based features are suitable for explicitly analyzing internal representations; however, some elaboration of lip-shape models and precise hand-labeled training data are required to construct a statistical model that represents valid lip shapes.

The second is a bottom-up approach. Various methods can be used to directly estimate visual features from the image; for example, dimensionality compression algorithms, such as discrete cosine transform [35, 44], principal component analysis (PCA) [3, 35], and discrete wavelet transform [35]. These algorithms are commonly utilized to extract lower-level image-based features, which are advantageous because they do not require dedicated lip-shape models or hand-labeled data for training; however, they are vulnerable to changes in lighting conditions, translation, or rotation of input images. In this study, we adopt the bottom-up approach by introducing a CNN as a visual feature extraction mechanism, because it is possible that CNNs can overcome the weaknesses of conventional image-based feature extraction mechanisms. The acquired visual features are also processed with a GMM-HMM system.

Several approaches for application of CNNs to speech recognition studies have been proposed. Abdel-Hamid et al. [1, 2] applied their original functionally extended CNNs for sound spectrogram inputs and demonstrated that their CNN architecture outperformed earlier basic forms of fully connected DNNs on phone recognition and large vocabulary speech recognition tasks. Palaz et al. [39] applied a CNN for phoneme sequence recognition by estimating phoneme class conditional probabilities from raw speech signal inputs. This approach yielded comparable or better phoneme recognition performance relative to conventional approaches. Lee et al. [29] applied a convolutional deep belief network (DBN) for various audio classification tasks, such as speaker identification, gender classification, and phone classification, that showed better performance compared with conventional hand-crafted audio features. Thus, CNNs are attracting increasing attention in speech recognition studies. However, applications of CNNs has been limited to audio signal processing; applications to lipreading remains unaddressed.

2.3 Audio-visual integration mechanisms

Multimodal recognition can improve performance compared with unimodal recognition by utilizing complementary sources of information [9, 15, 42]. Multimodal integration is commonly achieved by two different approaches. First, in the feature fusion approach, feature vectors from multiple modalities are concatenated and transformed to acquire a multimodal feature vector. For example, Ngiam et al. [37] utilized a DNN to extract fused representations directly from multimodal signal inputs by compressing the input dimensionality. Huang et al. [19] utilized a DBN for audio-visual speech recognition tasks by combining mid-level features learned by single modality DBNs. However, these approaches have difficulty explicitly and adaptively selecting the respective information gains depending on the dynamic changes in the reliability of multimodal information sources. Alternatively, in the decision fusion approach, outputs of unimodal classifiers are merged to determine a final classification. Unlike the previous method, decision fusion techniques can improve robustness by incorporating stream reliabilities associated with multiple information

sources as a criterion of information gain for a recognition model. For example, Gurban et al. [14] succeeded in dynamic stream weight adaptation based on modality confidence estimators in the MSHMM for their AVSR problem. Placing emphasis on the simplicity and explicitness of the decision fusion approach, we adopted the MSHMM as our multimodal integration mechanism.

3 Deep neural networks

Efforts to apply neural network architectures as feature extraction mechanisms have been attempted for decades [30]. However, the following three factors have made a major breakthrough in the application of DNNs to the problems of image classification and speech recognition. First, popularization of low-cost, high-performance computational environments, i.e., high-end consumer personal computers equipped with general-purpose graphics processing units (GPGPUs), has allowed a wider range of users to conduct brute force numerical computation with large datasets [10]. Second, the availability of larger datasets has enabled unsupervised learning mechanisms to self-organize robust features that can outperform conventional handcrafted features. Third, the development of powerful machine learning techniques, e.g., improved optimization algorithms, has enabled large-scale neural network models to be efficiently trained with large datasets, which has made deep neural networks generating robust features possible [17]. In this section, we introduce representative deep learning algorithms that have contributed to the recent development of deep learning studies.

3.1 Deep autoencoder

The deep autoencoder is a variant of a DNN commonly utilized for dimensionality compression and feature extraction [18, 37]. DNNs are artificial neural network models with multiple layers of hidden units between inputs and outputs. An autoencoder is a variant of a multilayered artificial neural network with a bottleneck-shaped structure (the number of nodes for the central hidden layer becomes smaller than that for the input (encoder) and output (decoder) layers), and the network is trained to model the identity mappings between inputs and outputs. Regarding dimensionality compression mechanisms, a simple and commonly utilized approach is PCA. However, Hinton et al. demonstrated that the deep autoencoder outperformed PCA in image reconstruction and compressed feature acquisition [18].

To train DNNs, Hinton et al. first proposed an unsupervised learning algorithm to use greedy layer-wise unsupervised pretraining followed by fine-tuning methods to overcome the high prevalence of unsatisfactory local optima in learning objectives of deep models [18]. Subsequently, Martens proposed a novel approach by introducing a second-order optimization method, Hessian-free optimization, to train deep networks [33]. The proposed method efficiently trained the models by a general optimizer without pretraining. Placing emphasis on the theoretical clarity of their algorithm, we adopted the learning method proposed by Martens for optimizing our deep autoencoder.

3.2 Hessian-free optimization

The Hessian-free algorithm originates from Newton's method, a well-known numerical optimization technique. A canonical second-order optimization scheme, such as Newton's method, iteratively updates parameter $\theta \in \mathbb{R}^N$ of an objective function f by computing gradient vector p and updates θ as $\theta_{n+1} = \theta_n + \alpha p_n$ with learning parameter α. The core idea of Newton's method is to locally approximate f around each θ, up to the second order, by the following quadratic equation:

$$M_{\theta_n}(\theta) \equiv f(\theta_n) + \nabla f(\theta_n)^T p_n + \frac{1}{2} p_n^T B_{\theta_n} p_n, \qquad (1)$$

where B_{θ_n} is a damped Hessian matrix of f at θ_n. As H can become indefinite, the Hessian matrix is re-conditioned to be $B_{\theta_n} = H(\theta_n) + \lambda I$, where $\lambda \geq 0$ is a damping parameter and I is the unit matrix.

Using the standard Newton's method, $M_{\theta_n}(\theta)$ is optimized by computing $N \times N$ matrix B_{θ_n} and solving the system $B_{\theta_n} p_n = -\nabla f(\theta_n)^T$. However, this computation is very expensive for a large N, which is a common case even with modestly sized neural networks. To overcome this, the variant of Hessian-free optimization developed by Martens utilizes the linear conjugate gradient (CG) algorithm to optimize quadratic objectives in combination with a positive semidefinite Gauss-Newton curvature matrix, instead of the possibly indefinite Hessian matrix. The name "Hessian-free" indicates that the CG does not necessarily require the costly, explicit Hessian matrix; instead, the matrix-vector product between the Hessian matrix H or Gauss-Newton matrix G and gradient vector p is sufficient (for implementation details, see [33], [40], and [45]).

3.3 Convolutional neural network

A CNN is a variant of a DNN commonly utilized for image classification problems [26–28]. CNNs integrate three architectural ideas to ensure spatial invariance: local receptive fields, shared weights, and spatial subsampling. Accordingly, CNNs are advantageous compared with ordinary fully connected feed-forward networks in the following three ways.

First, the local receptive fields in the convolutional layers extract local visual features by connecting each unit only to small local regions of an input image. Local receptive fields can extract visual features such as oriented-edges, endpoints, and corners. Typically, pixels in close proximity are highly correlated and distant pixels are weakly correlated. Thus, the stack of convolutional layers is structurally advantageous for recognizing images by effectively extracting and combining the acquired features. Second, CNNs have an advantage to some degree relative to spatial invariance with respect to shift, scale, or local distortion of inputs by forcing sharing of same weight configurations across the input space. Units in a plane are forced to perform the same operation on different parts of the image. As CNNs are equipped with several local receptive fields, multiple features are extracted at each location. In principle, fully connected networks are also able to perform similar invariances. However, learning such weight configurations requires a very large number of training datasets to cover all possible variations. Third, subsampling layers, which perform local averaging and subsampling, are utilized to reduce the resolution of the feature map and sensitivity of the output to input shifts and distortions (for implementation details, see [27]).

In terms of computational scalability, shared weights allow CNNs to possess fewer connections and parameters compared with standard feed-forward neural networks with similar-sized layers. Moreover, current improvements in computational resource availability, especially with highly-optimized implementations of two-dimensional convolution algorithms processed with GPGPUs, has facilitated efficient training of remarkably large CNNs with datasets containing millions of images [22, 25].

4 The dataset

A Japanese audiovisual dataset [23, 51] was used for the evaluation of the proposed models. In the dataset, speech data from six males (400 words: 216 phonetically-balanced words and 184 important words from the ATR speech database [23]) were used. In total, 24000 word recordings were prepared (one set of words per speaker; approximately 1 h of speech in total). The audio-visual synchronous recording environment is shown in Fig. 1. Audio data was recorded with a 16 kHz sampling rate, 16-bit depth, and a single channel. To train the acoustic model utilized for the assignment of phoneme labels to image sequences, we extracted 39 dimensions of audio features, composed of 13 MFCCs and their first and second temporal derivatives. To synchronize the acquired features between audio and video, MFCCs were sampled at 100 Hz. Visual data was a full-frontal 640×480 pixel 8-bit monochrome facial view recorded at 100 Hz. For visual model training and
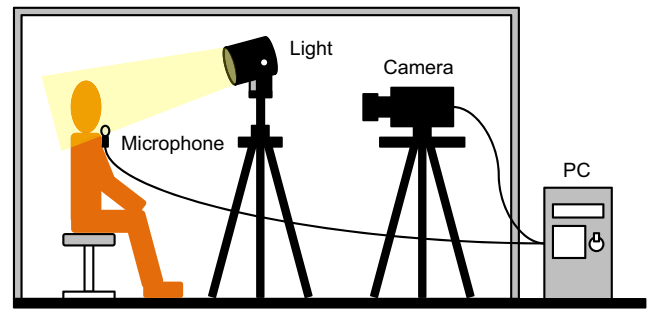
Fig. 1 Audio-visual synchronous data recording environment

evaluation, we prepared a trimmed dataset composed of multiple image resolutions by manually cropping 128×128 pixels of the mouth area from the original data and resizing the cropped data to 64×64, 32×32, and 16×16 pixels.

5 Model

A schematic diagram of the proposed AVSR system is shown in Fig. 2. The proposed architecture consists of two feature extractors to process audio signals synchronized with lip region image sequences. For audio feature extraction, a deep denoising autoencoder [48, 49] is utilized to filter out the effect of background noise from deteriorated audio features. For visual feature extraction, a CNN is utilized to recognize phoneme labels from lip image inputs. Finally, a multi-stream HMM recognizes isolated words by binding acquired multimodal feature sequences.

5.1 Audio feature extraction by deep denoising autoencoder

For the audio feature extraction, we utilized a deep denoising autoencoder [48, 49]. Eleven consecutive frames of audio features are used as the short-time spectral representation of speech signal inputs. To generate audio input feature sequences, partially deteriorated sound data are artificially generated by superimposing several strengths of Gaussian noises to original sound signals. In addition to the original clean sound data, we prepared six different deteriorated sound data; the signal-to-noise-ratio (SNR) was from 30 to -20 dB at 10 dB intervals. Utilizing sound feature extraction tools, the following types of sound features are generated from eight variations of original clean and deteriorated sound signals. HCopy command of the hidden Markov model toolkit (HTK) [52] is utilized to extract 39 dimensions of MFCCs. Auditory Toolbox [46] is utilized to extract 40 dimensions of log mel-scale fiterbank (LMFB). Finally, the deep denoising autoencoder is trained to reconstruct clean audio features from deteriorated features by preparing the deteriorated dataset as input and the corresponding clean dataset as the target of the network.
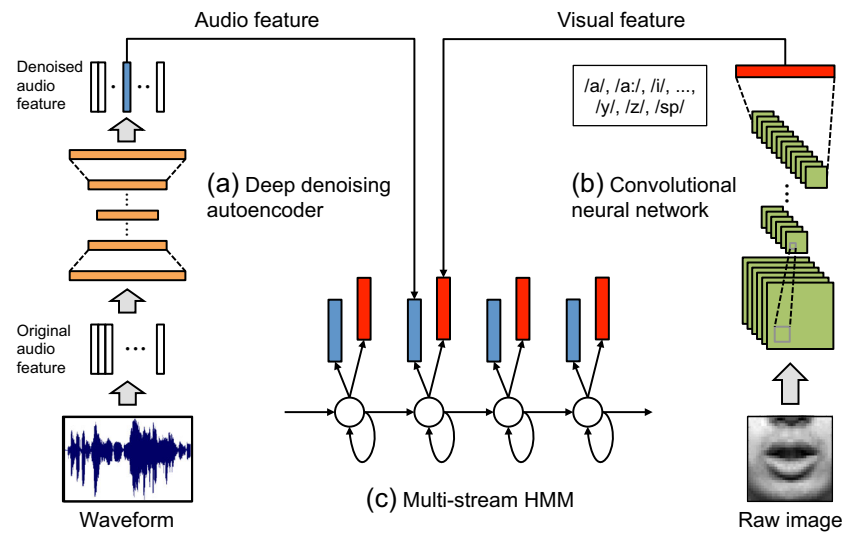
Fig. 2 Architecture of the proposed AVSR system. The proposed system is composed of two deep learning architectures, a deep denoising autoencoder and CNN for the audio and visual feature extraction, respectively. The deep denoising autoencoder is trained to predict clean audio features from deteriorated ones to filter out the effect of noise from the input. The CNN is trained to predict phoneme labels from the mouth area image inputs to generate the visual feature sequences from the lip motion of speakers. Finally, an MSHMM is utilized for the isolated word recognition by integrating the acquired audio and visual features

Among a 400-word dataset, sound signals from 360 training words (2.76×10^5 samples) and the remaining 40 test words (2.91×10^4 samples) from six speakers are used to train and evaluate the network, respectively.

The denoised audio features are generated by recording the neuronal outputs of the deep autoencoder when 11 frames of audio features are provided as input. To compare the denoising performance relative to the construction of the network, several different network architectures are compared. Table 1 summarizes the number of input and output dimensions, as well as layer-wise dimensions of the deep autoencoder.

In the initial experiment, we compared three different methods to acquire denoised features with respect to MFCCs and LMFB audio features. The first generated 11 frames of output audio features and utilized the middle frame (SequenceOut). The second acquired audio feature from the activation pattern of the central middle layer of

the network (BottleNeck). For these two experiments, a bottleneck-shaped network was utilized (Table 1 (a)). The last generated a single frame of an output audio feature that corresponds to the middle frame of the inputs (SingleFrame-Out). For this experiment, a triangle-shaped network was utilized (Table 1 (b)).

In the second experiment, we compared the performance relative to the number of hidden layers of the network utilizing an MFCCs audio feature. In this experiment, we prepared four straight-shaped networks with different numbers of layers (i.e., one to seven layers) at intervals of two (Table 1 (c)–(f)). Outputs were acquired by generating 11 frames of output audio features and utilizing the middle frame. Regarding the activation functions of the neurons, a linear function and logistic nonlinearity are utilized for the central middle layer of the bottleneck-shaped network and the remaining network layers, respectively. Parameters for the network structures are empirically determined with reference to previous studies [18, 21].

The deep autoencoder is optimized to minimize the objective function $\mathcal{E}$ defined by the sum of L2-norm between the output of the network and target vector across training dataset $\mathcal{D}$ under the model parameterized by θ, represented as

$$\mathcal{E}(\mathcal{D}, \theta) = \sqrt{\sum_{i=1}^{|\mathcal{D}|} (\hat{x}^{(i)} - x^{(i)})^2}, \qquad (2)$$

where $\hat{x}^{(i)}$ and $x^{(i)}$ are the output of the network and corresponding target vector from the i-th data sample,

Table 1 Settings for audio feature extraction

IN*	OUT*	LAYERS*	
429	429	300-150-80-40-80-150-300	(a)
429	39	300-150-80	(b)
429	429	300-300-300-300-300-300-300	(c)
429	429	300-300-300-300-300	(d)
429	429	300-300-300	(e)
429	429	300	(f)

*IN, OUT, and LAYERS give the number of input and output dimensions, and layer-wise dimensions of the network, respectively

respectively. To optimize the deep autoencoder, we adopted the Hessian-free optimization algorithm proposed by Martens [33]. In our experiment, the entire dataset was divided into 12 chunks with approximately 85000 samples per batch. We utilized 2.0×10^{-5} for the L2 regularization factor on the connection weights. For the connection weight parameter initialization, we adopted the sparse random initialization scheme to limit the number of non-zero incoming connection weights to each unit to 15. Bias parameters were initialized at 0. To process the substantial amount of linear algebra computation involved in this optimization algorithm, we developed a software library using the NVIDIA CUDA Basic Linear Algebra Subprograms [38]. The optimization computation was conducted on a consumer-class personal computer with an Intel Core i7-3930K processor (3.2 GHz, 6 cores), 32 GB RAM, and a single NVIDIA GeForce GTX Titan graphics processing unit with 6 GB on-board graphics memory.

5.2 Visual feature extraction by CNN

For visual feature extraction, a CNN is trained to predict phoneme label posterior probabilities corresponding to the mouth area input images. Mouth area images of 360 training words from six speakers were used to train and evaluate the network. To assign phoneme labels to every frame of the mouth area image sequences, we trained a monophone HMM with MFCCs utilizing the HTK and assigned 40 phoneme labels, including Japanese 39 phonemes (Table 2) and short pause /sp/, to the visual feature sequence by conducting a forced alignment by using the HVite command in the HTK.

To enhance shift- and rotation-invariance, artificially modulated images created by randomly shifting and rotating the original images are added to the original dataset. In addition, images labeled as short pause /sp/ are eliminated, with the exception of the five adjacent frames before and after the speech segments. The image dataset (3.05×10^5 samples) were shuffled and 5/6 of the data were used for training; the remainder was used evaluation of a phoneme

Table 2 39 types of Japanese phonemes

Category	Phoneme labels
Vowels	/a/ /i/ /u/ /e/ /o/ /a:/ /i:/ /u:/ /e:/ /o:/
Consonants	/b/ /d/ /g/ /h/ /k/ /m/ /n/ /p/ /r/ /s/ /t/ /w/ /y/ /z/ /ts/ /sh/ /by/ /ch/ /f/ /gy/ /hy/ /j/ /ky/ /my/ /ny/ /py/ /ry/
Others	/N/ /q/

recognition experiment. From our preliminary experiment, we confirmed that phoneme recognition precision degrades if images from all six speakers are modeled with a single CNN. Therefore, we prepared an independent CNN for each speaker.[1] The visual features for the isolated word recognition experiment are generated by recording the neuronal outputs (phoneme label posterior probability distribution) from the last layer of the CNN when mouth area image sequences corresponding to 216 training words were provided as inputs to the CNN.

A seven-layered CNN is used in reference to the work by Krizhevsky et al. [22]. Table 3 summarizes construction of the network containing four weighted layers: three convolutional (C1, C3, and C5) and one fully connected (F7). The first convolutional layer (C1) filters the input image with 32 kernels of 5×5 pixels with a stride of one pixel. The second and third convolutional layers (C3 and C5) take the response-normalized and pooled output of the previous convolutional layers (P2 and P4) as inputs and filter them with 32 and 64 filters of 5×5 pixels, respectively. The fully connected layer (F7) takes the pooled output of the previous convolutional layer (P6) as input and outputs a 40-way soft-max, regarded as a posterior probability distribution over the 40 classes of phoneme labels. A max-pooling layer follows the first convolution layer. Average-pooling layers follow the second and third convolutional layers. Response-normalization layers follow the first and second pooling layers. Rectified linear unit nonlinearity is applied to the outputs of the max-pooling layer as well as the second and third convolutional layers. Parameters for the network structures are empirically determined in reference to previous studies [22, 27].

The CNN is optimized to maximize the multinomial logistic regression objective of the correct label. This is equivalent to maximizing the likelihood $\mathcal{L}$ defined by the sum of log-probability of the correct label across training dataset $\mathcal{D}$ under the model parameterized by θ, represented as

$$\mathcal{L}(\mathcal{D}, \theta) = \sum_{i=1}^{|\mathcal{D}|} \log(P(Y = y^{(i)} | x^{(i)}, \theta)), \qquad (3)$$

where $y^{(i)}$ and $x^{(i)}$ are the class label and input pattern corresponding to the i-th data sample, respectively. The prediction distribution is defined with the softmax function as

$$P(Y = i | x, \theta) = \frac{\exp(h_i)}{\Sigma_{j=1}^{C} \exp(h_j)}, \qquad (4)$$

---

[1] We think this degradation is mainly due to the limited variations of lip region images that we prepared to train the CNN. To generalize the higher-level visual features that enable a CNN to attain speaker invariant phoneme recognition, we believe that more image samples from different speakers are needed.

Table 3 Construction of convolutional neural network

IN*	OUT*	LAYERS*
256/1024/4096	40	C1-P2-C3-P4-C5-P6-F7**

*IN, OUT, and LAYERS give the input dimensions, output dimensions, and network construction, respectively

**C, P, and F denote the convolutional, local-pooling, and fully connected layer, respectively. The numbers after the layer types represent layer indices

where h_i and C are the total input to output unit i and number of classes, respectively. The CNN is trained using a stochastic gradient descent method [22]. The update rule for the connection weight w is defined as

$$v_{i+1} = \alpha v_i - \gamma \epsilon w_i - \epsilon \left\langle \frac{\partial \mathcal{L}}{\partial w} |w_i \right\rangle_{\mathcal{D}_i} \quad (5)$$

$$w_{i+1} = w_i + v_{i+1} \quad (6)$$

where i is the learning iteration index, v_i is the update variable, α is the factor of momentum, ϵ is the learning rate, γ is the factor of weight decay, and $\left\langle \frac{\partial \mathcal{L}}{\partial w} |w_i \right\rangle_{\mathcal{D}_i}$ is the average over the i-th batch data $\mathcal{D}_i$ of the derivative of the objective with respective to w, evaluated at w_i. In our experiment, the mini batches are one-sixth of the entire dataset for each speaker (approximately 8500 samples per batch). We utilized $\alpha = 0.9$, $\epsilon = 0.001$, and $\gamma = 0.004$ in our leaning experiment. The weight parameters were initialized with a zero-mean Gaussian distribution with standard deviation 0.01. The neuron biases in all layers were initialized at 0. We used open source software (cuda-convnet) [22] for practical implementation of the CNN. The software was processed on the same computational hardware as the audio feature extraction experiment.

5.3 Audio-visual integration by MSHMM

In our study, we adopt a simple MSHMM with manually selected stream weights for the multimodal integration mechanism. We utilize the HTK for the practical MSHMM implementation. The HTK can model output probability distributions composed of multiple streams of GMMs [52]. Each observation vector at time t is modeled by splitting it into S independent data streams o_{st}. The output probability distributions of state j is represented with multiple data streams as

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s}, \quad (7)$$

where o_t is a speech vector generated from the probability density $b_j(o_t)$, M_s is the number of mixture components in stream s, c_{jsm} is the weight of the m'th component, $\mathcal{N}(\cdot; \mu, \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, and the exponent γ_s is a stream weight for stream s.

Definitions of MSHMM are generated by combining multiple HMMs independently trained with corresponding audio and visual inputs. In our experiment, we utilize 16 mixture components for both audio and visual output probability distribution models. When combining two HMMs, GMM parameters from audio and visual HMMs are utilized to represent stream-wise output probability distributions. Model parameters from only the audio HMM are utilized to represent the common state transition probability distribution. Audio stream weights γ_a are manually prepared from 0 to 1.0 at 0.1 intervals. Accordingly, visual stream weights γ_v are prepared to satisfy $\gamma_v = 1.0 - \gamma_a$. In evaluating the acquired MSHMM, the best recognition rate is selected from the multiple evaluation results corresponding to all stream weight pairs.
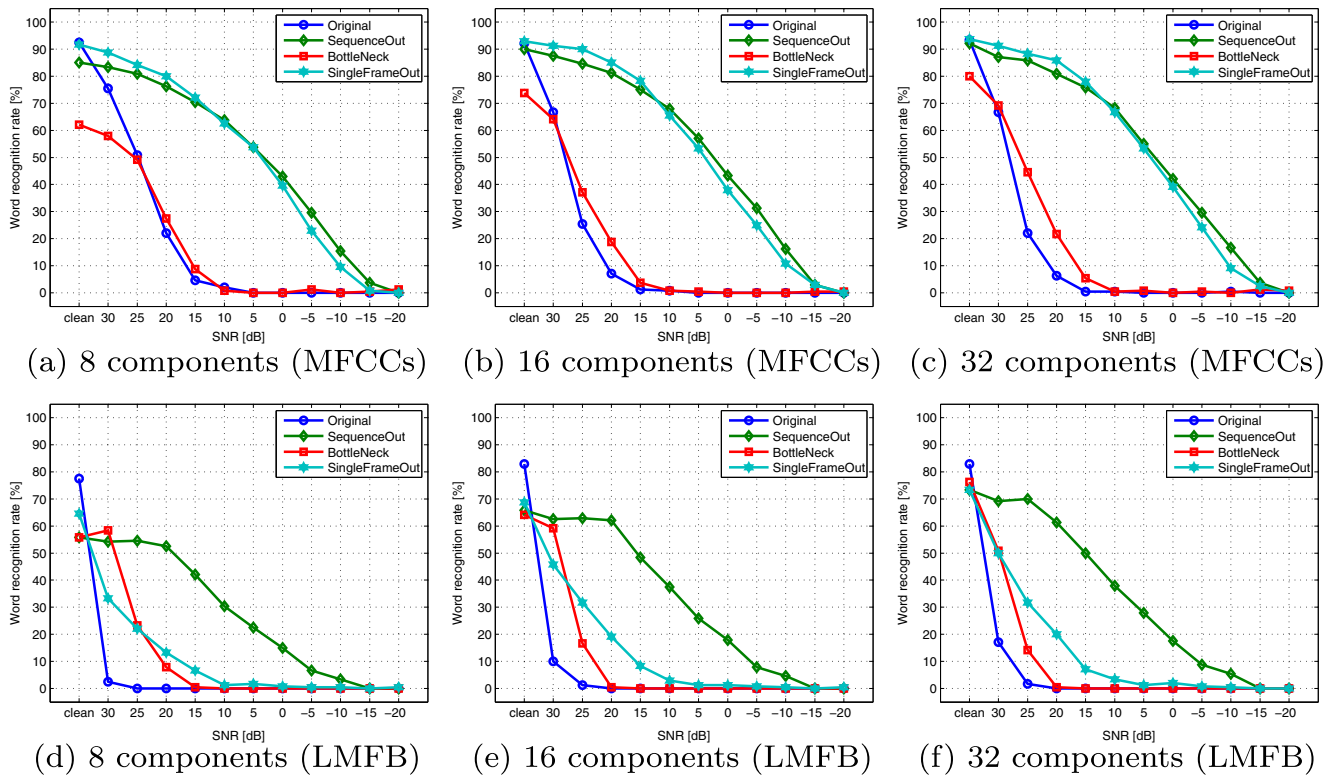
6 Results

6.1 ASR performance evaluation

The acquired audio features are evaluated by conducting an isolated word recognition experiment utilizing a single-stream HMM. To recognize words from the audio features acquired by the deep denoising autoencoder, monophone HMMs with 8, 16, and 32 GMM components are utilized. While training is conducted with 360 train words, evaluation is conducted with 40 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To enable comparison with the baseline performance, word recognition rates utilizing the original audio features are also prepared. To evaluate the robustness of our proposed mechanism against the degradation of audio input, partially deteriorated sound data were artificially generated by superimposing several strengths of Gaussian noises to original sound signals. In addition to the original clean sound data, we prepared 11 different deteriorated sound data such that the SNR was 30 dB to −20 dB at 5 dB intervals.

Figure 3 shows word recognition rates from the different word recognition models for MFCCs and LMFB audio features evaluated with 12 different SNRs for sound inputs. These results demonstrate that MFCCs generally outperforms LMFB. The sound feature acquired by integrating consecutive multiple frames with a deep denoising autoencoder has an effect on higher noise robustness compared with the original input. By comparing the audio features acquired from the different network architectures, it was

(a) 8 components (MFCCs)   (b) 16 components (MFCCs)   (c) 32 components (MFCCs)

(d) 8 components (LMFB)   (e) 16 components (LMFB)   (f) 32 components (LMFB)

Fig. 3 Word recognition rate evaluation results utilizing audio features depending on the number of Gaussian mixture components for the output probability distribution models of HMM. Changes of word recognition rates depending on the types of audio features (MFCCs for (**a**) to (**c**) and LMFB for (**d**) to (**f**)), the types of feature extraction mechanism, and changes of the SNR of audio inputs are shown

observed that "SingleFrameOut" obtains the highest recognition rates for the higher SNR range, whereas "Sequence-Out" outperforms for the lower SNR range. While "Bottle-Neck" performs slightly better than the original input for the middle SNR range, the advantage is scarce. Overall, approximately a 65 % word recognition gain was attained with denoised MFCCs under 10 dB SNR. Although there is a slight recognition performance difference depending on the increase of the number of Gaussian mixture components, the effect is not significant.

Figure 4 shows word recognition rates for the different number of hidden layers of the deep denoising autoencoder utilizing MFCCs audio features evaluated with 12 different SNRs for sound inputs. The deep denoising autoencoder with five hidden layers obtained the best noise robust word recognition performance among all SNR ranges.

6.2 Visual-based phoneme recognition performance evaluation

After training the CNN, phoneme recognition performance is evaluated by recording neuronal outputs from the last layer of the CNN when the mouth area image sequences corresponding to the test image data are provided to the CNN. Table 4 shows that the average phoneme recognition performance for the 40 phonemes, normalized with the number of samples for each phoneme over six speakers, attained approximately 48 % when 64×64 pixels of mouth area images are utilized as input.

Figure 5 shows the mean and standard deviation of the phoneme-wise recognition rate from six different speakers for four different input image resolutions.

This result generally demonstrates that visual phoneme recognition works better for recognizing vowels than consonants. The result derives from the fact that the mean recognition rate for all vowels is 30–90 %, whereas for all other phonemes it is 0–60 %. This may be attributed to the fact that generation of vowels is strongly correlated

Table 4 Speaker-wise visual-based phoneme recognition rates and averaged values [%] depending on the input image sizes

Img. size	p1	p2	p3	p4	p5	p6	Avr.
16×16	42.13	43.40	39.92	39.03	47.67	46.73	43.15
32×32	43.77	47.07	42.77	41.05	49.74	50.83	45.87
64×64	45.93	50.06	46.51	43.57	49.95	51.44	47.91

*p1–p6 correspond to the six speakers

(a) 8 components (b) 16 components (c) 32 components
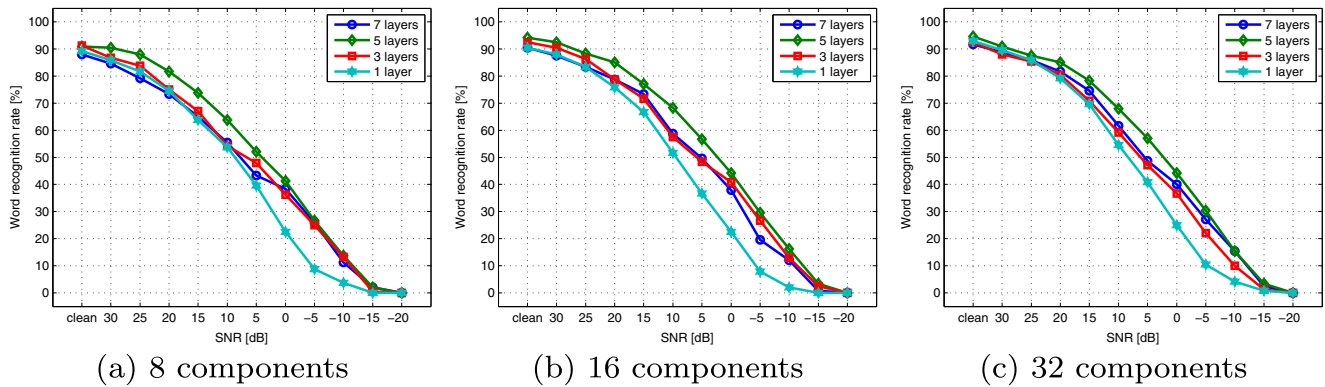
Fig. 4 Word recognition rate evaluation results utilizing MFCCs depending on the number of Gaussian mixture components for the output probability distribution models of HMM. Changes of word recognition rates depending on the number of hidden layers of DNN and changes of the SNR of audio inputs are shown

with visual cues represented by lips or jaw movements [4, 50].

Figure 6 shows the confusion matrix of the phoneme recognition evaluation results. It should be noted that, in most cases, wrongly recognized consonants are classified as vowels. This indicates that articulation of consonants is attributed to not only the motion of the lips but also the dynamic interaction of interior oral structures such as tongue, teeth, oral cavity, which are not evident in frontal facial images.

Visually explicit phonemes, such as bilabial consonants (/m/, /p/, or /b/), are expected to be relatively well discriminated by a VSR system. However, the recognition performance was not as high as expected. To improve the recognition rate, the procedure to obtain phoneme target labels for the CNN training should be improved. In general pronunciation, consonant sounds are shorter than vowel sounds; therefore, the labeling for consonants is more time critical than vowels. In addition, the accuracy of consonant labels directly affects recognition performance because the number of training samples for consonants is much smaller than for vowels.

6.3 Visual feature space analysis

To analyze how the acquired visual feature space is self-organized, the trained CNN is used to generate phoneme posterior probability sequences from test image sequences. Forty dimensions of the resulting sequences are processed by PCA, and the first three principal components are extracted to visualize the acquired feature space. Figure 7 shows the visual feature space corresponding to the five representative Japanese vowel phonemes, /a/, /i/, /u/, /e/, and /o/, generated from 64×64 pixels image inputs. The cumulative contribution ratio with 40 selected components was 31.1 %.

As demonstrated in the graph, raw mouth area images corresponding to the five vowel phonemes are discriminated by the CNN and clusters corresponding to the phonemes are self-organized in the visual feature space. This result
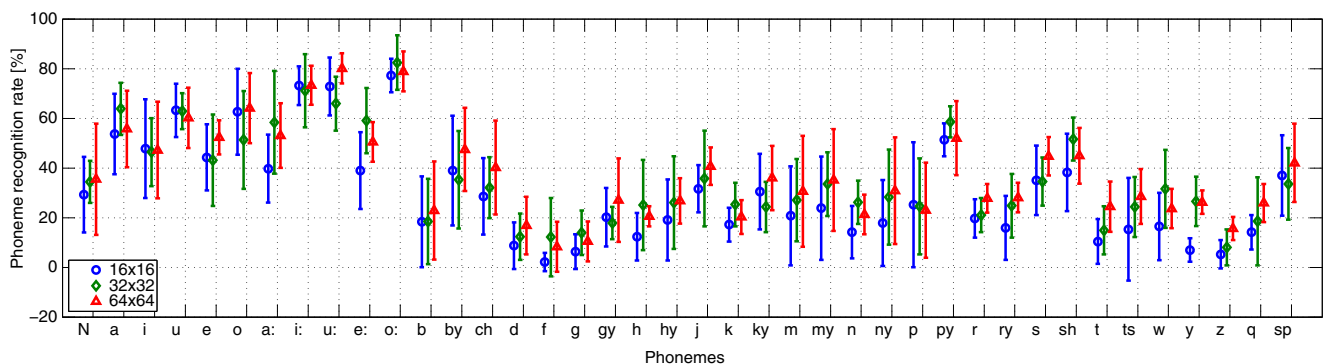
Fig. 5 Phoneme-wise visual-based phoneme recognition rates, the mean and the standard deviations from six speakers' results are shown. Four different shapes of the plots correspond to the recognition results when four different visual features, acquired by the CNN from four different image resolutions for the mouth area image inputs, are utilized

Fig. 6 Visual-based phoneme-recognition confusion matrix; mean values from six speakers' results (64 × 64 pixels image input)
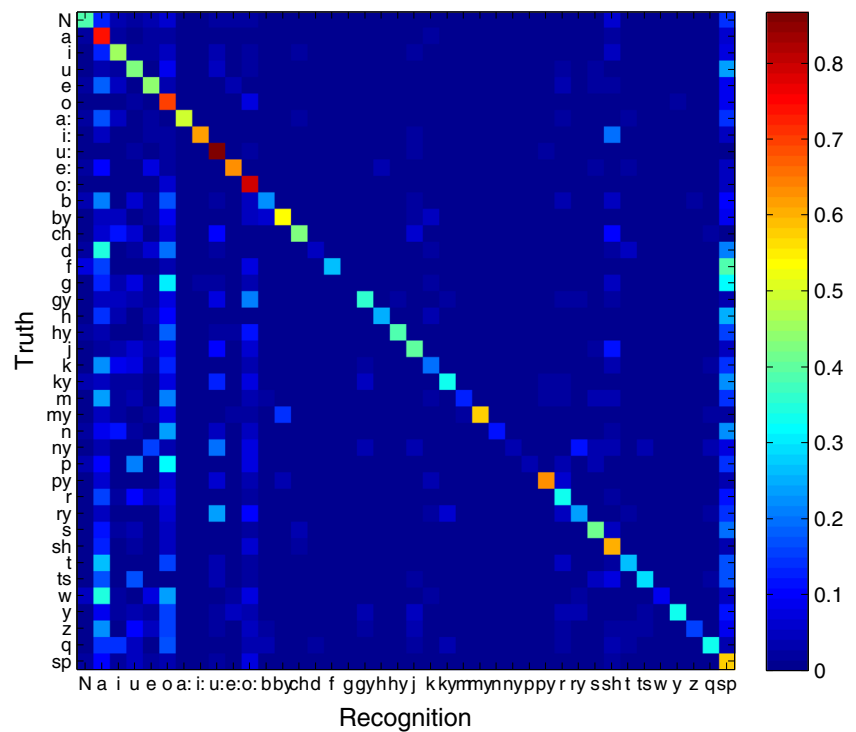
indicates that the acquired phoneme posterior probability sequences can be utilized as visual feature sequences for isolated word recognition tasks.

6.4 VSR performance evaluation

The acquired visual features are evaluated by conducting an isolated word recognition experiment utilizing a single-stream HMM. To recognize words from the phoneme label sequences generated by the CNN trained with 360 training words, monophone HMMs with 1, 2, 4, 8, 16, 32, and 64 Gaussian components are utilized. While training is conducted with 360 train words, evaluation is conducted with 40 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To compare with the baseline performance, word recognition rates utilizing two other visual features are also prepared. One feature has 36 dimensions, generated by simply rescaling the images to 6 × 6 pixels, and the other feature has 40 dimensions, generated by compressing the raw images by PCA.

Figure 8 shows the word recognition rates acquired from 35 different models with a combination of five types of visual features and seven different numbers of Gaussian mixture components for GMMs. Comparison of word recognition rates from different visual features within the same number of Gaussian components shows that visual features acquired by the CNN attain higher recognition rates than the other two visual features. However, the effect of the different input image resolutions is not prominent. Among

all word recognition rates, visual features acquired by the CNN with 16 × 16 and 64 × 64 input image resolutions attain a rate of approximately 22.5 %, the highest word recognition rate, when a mixture of 32 Gaussian components is used.
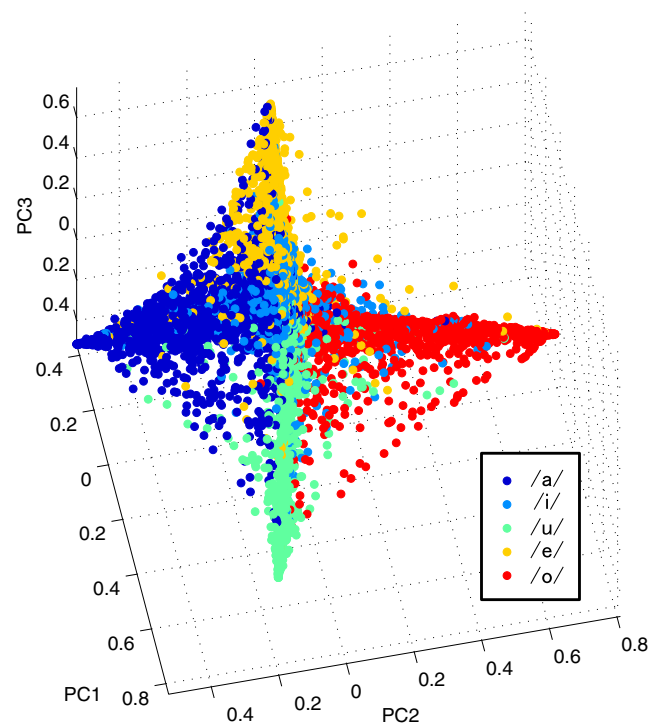
Fig. 7 Visual feature distribution for the five representative Japanese vowel phonemes (64 × 64 pixels image input)
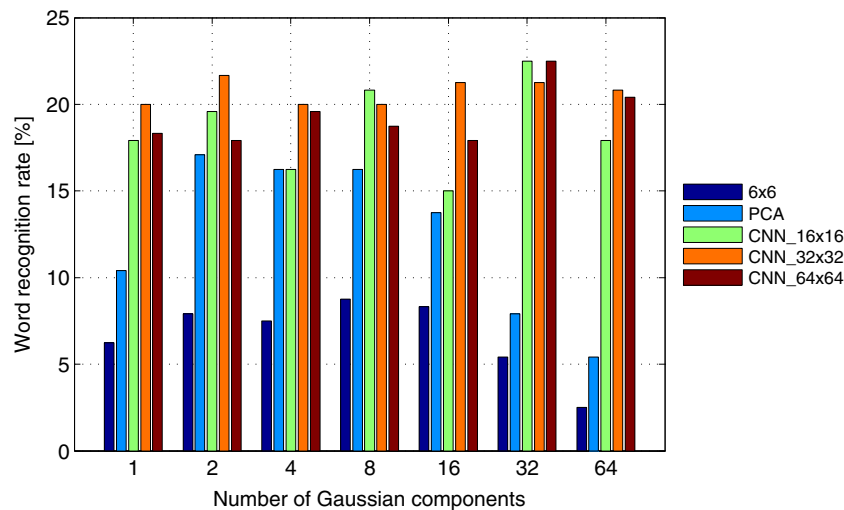
Fig. 8 Word recognition rates using image features. Evaluation results from 40 test words over six speakers depending on the different mixture of Gaussian components for the HMM are shown. Six visual features, including two image-based features, one generated by simply resampling the mouth area image into 6×6 pixels image and the other generated by compressing the dimensionality of the image into 40 dimensions by PCA, and four visual features acquired by predicting the phoneme label sequences from four different resolutions of the mouth area images utilizing the CNN, are employed in this evaluation experiment

6.5 AVSR performance evaluation

We evaluated the advantages of sensory features acquired by the DNNs and noise robustness of the AVSR by conducting an isolated word recognition task. Training data for the MSHMM are composed of image and sound features generated from 360 training words of six speakers. For sound features, we utilized the neuronal outputs of the straight-shaped deep denoising autoencoder with five hidden layers (Table 1 (d)) when clean MFCCs are provided as inputs. For visual features, we utilized the output phoneme label sequences generated from 32×32 pixels mouth area image inputs by the CNN. Evaluation data for the MSHMM are composed of image and sound features generated from the 40 test words. Thus, closed-speaker and open-vocabulary evaluation was conducted. To evaluate the robustness of our proposed mechanism against the degradation of audio input, partially deteriorated sound data were artificially generated by superimposing several strengths of Gaussian noises to original sound signals. In addition to the original clean sound data, we prepared 11 different deteriorated sound data such that the SNR was 30 dB to −20 dB at 5 dB intervals. In our evaluation experiment, we compared the performance under four different conditions. The initial two models were the unimodal models that utilize single-frame MFCCs and the denoised MFCCs acquired by the straight-shaped deep denoising autoencoder with five hidden layers. These are identical to the models "Original" and "5 layers" presented in Figs. 3 and 4, respectively. The third model was the unimodal model that utilized visual features acquired by the CNN. The fourth model was the multimodal model that binds the acquired audio and visual features by the MSHMM.

Figure 9 shows word recognition rates from the four different word recognition models under 12 different SNRs for sound inputs. These results demonstrate that when two modalities are combined to represent the acoustic model, the word recognition rates are improved, particularly for lower SNRs. At minimum, the same or a better performance was attained compared with cases when both features are independently utilized. For example, the MSHMM attained an additional 10 % word recognition rate gain under 0 dB SNR for the audio signal input compared with the case when single-stream HMM and denoised MFCCs are utilized as the recognition mechanism and input features, respectively. Although there is a slight recognition performance difference depending on the increase of the number of Gaussian mixture components, the effect is not significant.

7 Discussion and future work

7.1 Current need for the speaker dependent visual feature extraction model

In our study, we demonstrated an isolated word recognition performance from visual sequence inputs by the integration of CNN and HMM. We showed that the CNN works as a phoneme recognition mechanism with mouth region image inputs. However, our current results are attained by preparing an independent CNN corresponding to each

(a) 8 components   (b) 16 components   (c) 32 components
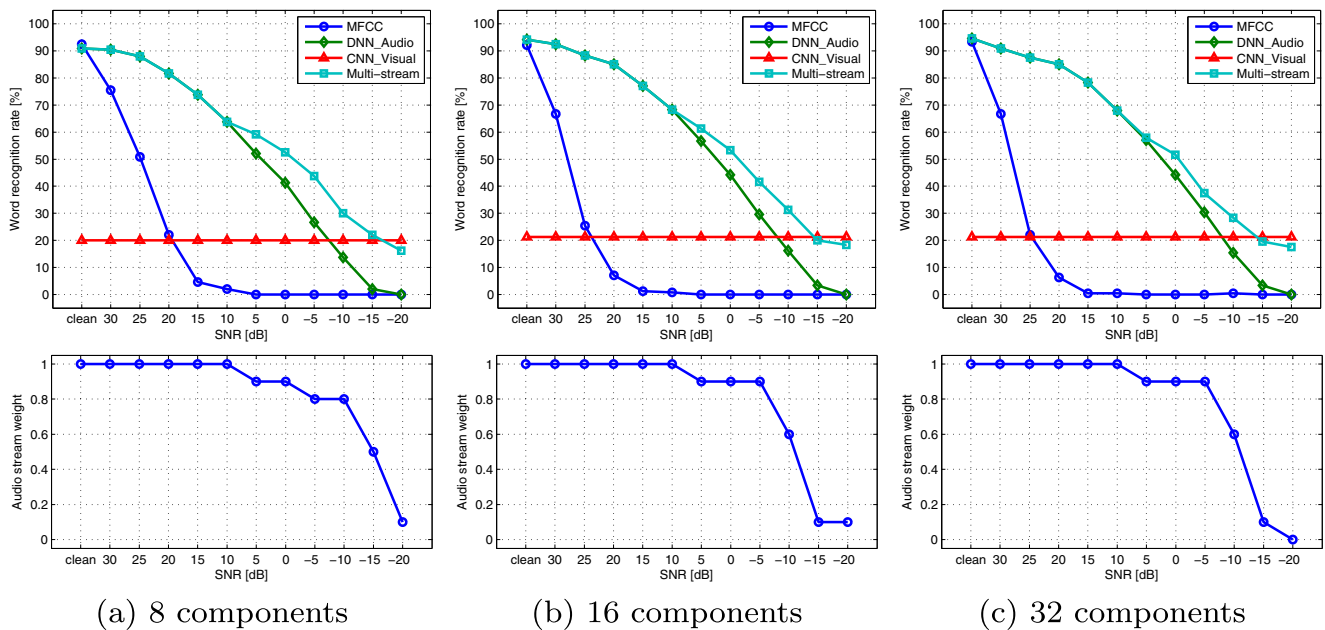
Fig. 9 Word recognition rate evaluation results utilizing dedicated features and multimodal features depending on the number of Gaussian mixture components for the output probability distribution models of HMM. Top: changes of word recognition rates depending on the types of utilized features and changes of the SNR of audio inputs. "MFCC," "DNN_Audio," "CNN_Visual," and "Multi-stream" denote the original MFCCs feature, audio feature extracted by the deep denoising autoencoder, visual feature extracted by the CNN, and MSHMM composed of "DNN_Audio" and "CNN_Visual" features, respectively. Bottom: audio stream weights that give the best word recognition rates for the MSHMM depending on changes of the audio inputs SNR

speaker. As generally discussed in previous deep learning studies [22, 25], the number and variation of training samples are critical for maximizing the generalization ability of a DNN. A DNN (CNN) framework is scalable; however, it requires a sufficient training dataset to reduce overfitting [22]. Therefore, in future work, we need to investigate the possibility of realizing a VSR system applicable to multiple speakers with a single CNN model by training and evaluating our current mechanism with a more diverse audio-visual speech dataset that has large variations, particularly for mouth region images.

7.2 Adaptive stream weight selection

Our AVSR system utilizing MSHMM achieved satisfactory speech recognition performance, despite its quite simple mechanism, especially for audio signal inputs with lower reliability. The transition of the stream weight in accordance with changes of the SNR for the audio input (Fig. 9) clearly demonstrates that the MSHMM can prevent the degradation of recognition precision by shifting the observation information source from audio input to visual input, even if the quality of the audio input degrades. However, to apply our AVSR approach to real-world applications, automatic and adaptive selection of the stream weight in relation to changes in audio input reliability becomes an important issue to be addressed.

7.3 Relations of our AVSR approach with DNN-HMM models

As an experimental study for an AVSR task, we adopted a rather simple tandem approach, a connectionist-HMM [16]. Specifically, we applied heterogeneous deep learning architectures to extract the dedicated sensory features from audio and visual inputs and combined the results with an MSHMM. We acknowledge that a DNN-HMM is known to be advantageous for directly estimating the state posterior probabilities of an HMM from raw sensory feature inputs over conventional GMM-HMM owing to the powerful nonlinear projection capability of DNN models [17]. In future, it might be interesting to formulate an AVSR model based on the integration of DNN-HMM and MSHMM. This novel approach may succeed because of the recognition capability of DNNs and simplicity and explicitness of the proposed decision fusion approach.

8 Conclusion

In this study, we proposed an AVSR system based on deep learning architectures for audio and visual feature extraction and an MSHMM for multimodal feature integration and isolated word recognition. Our experimental results demonstrated that, compared with the original MFCCs,

the deep denoising autoencoder can effectively filter out the effect of noise superimposed on original clean audio inputs and that acquired denoised audio features attain significant noise robustness in an isolated word recognition task. Furthermore, our visual feature extraction mechanism based on the CNN effectively predicted the phoneme label sequence from the mouth area image sequence, and the acquired visual features attained significant performance improvement in the isolated word recognition task relative to conventional image-based visual features, such as PCA. Finally, an MSHMM was utilized for an AVSR task by integrating the acquired audio and visual features. Our experimental results demonstrated that, even with the simple but intuitive multimodal integration mechanism, it is possible to attain reliable AVSR performance by adaptively switching the information source from audio feature inputs to visual feature inputs depending on the changes in the reliability of the different signal inputs. Although automatic selection of stream weight was not attained, our experimental results demonstrated the advantage of utilizing an MSHMM as an AVSR mechanism. The next major target of our work is to examine the possibility of applying our current approach to develop practical, real-world applications. Specifically, future work will include a study to evaluate how the VSR approach utilizing translation, rotation, or scaling invariant visual features acquired by the CNN contributes to robust speech recognition performance in a real-world environment, where dynamic changes such as reverberation, illumination, and facial orientation, occur.

References

1. Abdel-Hamid O, Jiang H. (2013) Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France

2. Abdel-Hamid O, rahman Mohamed A, Jiang H, Penn G (2012) Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Kyoto, pp 4277–4280

3. Aleksic PS, Katsaggelos AK (2004) Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 5, Montreal, pp 917–920

4. Barker J, Berthommier F (1999) Evidence of correlation between acoustic and visual features of speech. In: Proceedings of the

5. Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127

6. Bourlard H, Dupont S (1996) A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proceedings of the 4th International Conference on Spoken Language Processing, vol 1, Philadelphia, pp 426–429

7. Bourlard H, Dupont S, Ris C (1996) Multi-stream speech recognition.IDIAP research report

8. Bourlard Ha, Morgan N (1994) Connectionist speech recognition: a hybrid approach. Springer US, Boston

9. Brooke N, Petajan ED (1986) Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. In: Proceedings of the International Conference on Speech Input and Output, Techniques and Applications, London, pp 104–109

10. Coates A, Huval B, Wang T, Wu DJ, Ng AY, Catanzaro B (2013) Deep learning with COTS HPC. In: Proceedings of the 30th international conference on machine learning, Atlanta, pp 1337–1345

11. Cootes T, Edwards G, Taylor C (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685

12. Dahl GE, Acero A (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 20(1):30–42

13. Feng X, Zhang Y, Glass J (2014) Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Florence, pp 1759–1763

14. Gurban M, Thiran JP, Drugman T, Dutoit T (2008) Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition. In: Proceedings of the 10th International Conference on Multimodal Interfaces, Chania, pp 237–240

15. Heckmann M, Kroschel K, Savariaux C (2002) DCT-based video features for audio-visual speech recognition. In: Proceedings of the 7th International Conference on Spoken Language Processing, vol 3, Denver, pp 1925–1928

16. Hermansky H, Ellis D, Sharma S (2000) Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 3, Istanbul, pp 1635–1638

17. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Proc Mag 29:82–97

18. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–7

19. Huang J, Kingsbury B (2013) Audio-visual deep learning for noise robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, pp 7596–7599

20. Janin A, Ellis D, Morgan N (1999) Multi-stream speech recognition: Ready for prime time? In: Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest, Hungary

21. Krizhevsky A, Hinton GE (2011) Using very deep autoencoders for content-based image retrieval. In: Proceedings of the 19th European Symposium on Artificial Neural Networks. Bruges, Belgium

22. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems

23. Kuwabara H, Takeda K, Sagisaka Y, Katagiri S, Morikawa S, Watanabe T (1989) Construction of a large-scale Japanese speech database and its management system. In: Proceedings of the

14th International Congress of Phonetic Sciences, San Francisco, pp 5–9

IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, pp 560–563

24. Lan Y, Theobald BJ, Harvey R, Ong EJ, Bowden R (2010) Improving visual features for lip-reading. In: Proceedings of the International Conference on Auditory-Visual Speech Processing. Hakone,Japan

25. Le QV, Ranzato M, Monga R, Devin M, Chen K, Corrado GS, Dean J, Ng AY (2012) Building high-level features using large scale unsupervised learning. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, pp 81–88

26. LeCun Y, Bottou L (2004) Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2, Washington, pp 97–104

27. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

28. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, pp 609–616

29. Lee H, Pham P, Largman Y, Ng AY (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Proceedings of the Advances in Neural Information Processing Systems 22, Vancouver, pp 1096–1104

30. Lerner B, Guterman H, Aladjem M, Dinstein I (1999) A comparative study of neural network based feature extraction paradigms. Pattern Recogn Lett 20(1):7–14

31. Luettin J, Thacker N, Beet S (1996) Visual speech recognition using active shape models and hidden Markov models. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 2, Atlanta, pp 817–820

32. Maas AL, O'Neil TM, Hannun AY, Ng AY (2013) Recurrent neural network feature enhancement: The 2nd chime challenge. In: Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments.Vancouver, Canada

33. Martens J (2010) Deep learning via Hessian-free optimization. In: Proceedings of the 27th International Conference on Learning, Machine, Haifa, pp 735–742

34. Matthews I, Cootes T, Bangham J, Cox S, Harvey R (2002) Extraction of visual features for lipreading. IEEE Trans Pattern Anal Mach Intell 24(2):198–213

35. Matthews I, Potamianos G, Neti C, Luettin J (2001) A comparison of model and transform-based visual features for audio-visual LVCSR. In: Proceedings of the IEEE International Conference on Multimedia and Expo. Tokyo, Japan

36. Mohamed A, Dahl GE, Hinton GE (2012) Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process 20(1):14–22

37. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning

38. NVIDIA Corporation (2014) CUBLAS library version 6.0 user guide. CUDA Toolkit Documentation

39. Palaz D, Collobert R, Magimai.-Doss M (2013) Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France

40. Pearlmutter B (1994) Fast exact multiplication by the Hessian. Neural Comput 6(1):147–160

41. Renals S, Morgan N, Member S, Bourlard H, Cohen M, Franco H (1994) Connectionist probability estimators in HMM speech recognition 2(1):161–174

42. Robert-Ribes J, Piquemal M, Schwartz JL, Escudier P (1996) Exploiting sensor fusion architectures and stimuli complementarity in av speech recognition. In: Stork D, Hennecke M (eds) Speechreading by Humans and Machines. Springer, Berlin Heidelberg, pp 193–210

43. Sainath TN, Kingsbury B, Ramabhadran B (2012) Auto-encoder bottleneck features using deep belief networks. In:Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Kyoto, pp 4153–4156

44. Scanlon P, Reilly R (2001) Feature analysis for automatic speechreading. In: Proceedings of the IEEE 4th Workshop on Processing, Multimedia Signal, Cannes, pp 625–630

45. Schraudolph NN (2002) Fast curvature matrix-vector products for second-order gradient descent. Neural Comput 14(7):1723–38

46. Slaney M (1998) Auditory toolbox: A MATLAB toolbox for auditory modeling work version 2. Interval research corproation

47. Sutskever I, Martens J, Hinton G (2011) Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, pp 1017–1024

48. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning, New York, pp 1096–1103

49. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408

50. Yehia H, Rubin P, Vatikiotis-Bateson E (1998) Quantitative association of vocal-tract and facial behavior. Speech Comm 26:23–43

51. Yoshida T, Nakadai K, Okuno HG (2009) Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In: Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots, Paris, pp 604–609

52. Young S, Evermann G, Gales M, Hain T, Liu XA, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2009) The HTK Book (for HTK Version 3.4),.Cambridge University Engineering Department

Kuniaki Noda received the BS and MS in Mechanical Engineering in 2000 and 2002, respectively, from Waseda University, Japan. From 2002 to 2012, he worked for Sony Corporation. From 2009 to 2010, he was a visiting researcher at EPFL, Switzerland. Currently, he is a Ph.D. candidate at Waseda University. His research interests include autonomous robot, multimodal integration, deep learning, and high performance computing on GPU. He received various awards including the Hatakeyama Award from the Japan Society of Mechanical Engineers in 1999, the Best Paper Award of ICDL-EPIROB 2011, and the Best Paper Award of RSJ in 2012.

Yuki Yamaguchi received the BS and MS in Informatics in 2012 and 2014, respectively, from Kyoto University, Japan. His research interests include human-robot interaction, autonomous robot, deep learning and automatic audio-visual speech recognition. He received some awards including Best Paper Award for Young Researcher of IPSJ National Convention in 2013.

Kazuhiro Nakadai received B.E. in electrical engineering in 1993, M.E. in information engineering in 1995, and Ph.D. in electrical engineering in 2003 from The University of Tokyo. He was working with Nippon Telegraph and Telephone and NTT Comware Corporation for four years as a system engineer from 1995 to 1999. He was working with Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Agency (JST) as a researcher from 1999 to 2003. He is currently Principal Researcher for Honda Research Institute Japan, Co., Ltd. From 2006 to 2010, he was concurrently Visiting Associate Professor at Tokyo Institute of Technology, and he is Visiting Professor at Tokyo Institute of Technology since 2011. He also has another position of Visiting Professor at Waseda University from 2011. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multi-modal integration and robot audition. He is a member of RSJ, JSAI, ASJ, and IEEE.

Hiroshi G. Okuno received the BA and Ph.D from the University of Tokyo in 1972 and 1996, respectively. He worked for NTT, JST, Tokyo University of Science, and Kyoto University. He is currently a professor of Graduate Program for Embodiment Informatics, Graduate School of Creative Science and Engineering, Waseda University, and a professor emeritus, Kyoto University. He was visiting scholar at Stanford University from 1986 to 1988. He is currently engaged in computational auditory scene analysis, music information processing and robot audition. He received various awards including the 2013 Award of the Minister of Education, Culture, Sports, Science Technology of Japan for Science and Technology (Research Division), the 2nd Advanced Robotics Best Paper Award in 2014, the 1990 Best Paper Award of the Japanese Society for Artificial Intelligence (JSAI), and IROS-2010 NTF Award for Entertainment Robots and Systems in 2010. He coedited Computational Auditory Scene Analysis (CRC Press, 1998), Advanced Lisp Technology (Taylor and Francis, 2002), and New Trends in Applied Artificial Intelligence (IEA/AIE) (Springer, 2007). He is a fellow of IEEE and, JSAI and a member of AAAI, ACM, ASA, RSJ, IPSJ, JSSST and JCSST.

Tetsuya Ogata received the BS, MS and DE degrees in Mechanical Engineering, in 1993, 1995 and 2000, respectively, from Waseda University. He was a Research Fellow of JSPS, a Research Associate of Waseda University, a Research Scientist of RIKEN Brain Science Institute, and an Associate Professor of Kyoto University. He is currently a Professor of Faculty of Science and Engineering, Waseda University. Since 2009, he has been a JST PRESTO researcher. His research interests include human-robot interaction, dynamics of human-robot mutual adaptation and inter-sensory translation in robot systems. He is a member of IEEE, RSJ, JSAI, IPSJ, JSME, SICE, etc.