

PAPER

Multistream sparse representation features for noise robust audio-visual speech recognition

Peng Shen^{1,*}, Satoshi Tamura^{2,†} and Satoru Hayamizu^{2,‡}

¹Graduate School of Engineering, Gifu University, 1-1 Yanagido, Gifu, 501-1193 Japan

²Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu, 501-1193 Japan

(Received 9 November 2012, Accepted for publication 23 April 2013)

Abstract: In this paper, we propose to use exemplar-based sparse representation features for noise robust audio-visual speech recognition. First, we introduce a sparse representation technology and describe how noise robustness can be realized by the sparse representation for noise reduction. Then, feature fusion methods are proposed to combine audio-visual features with the sparse representation. Our work provides new insight into two crucial issues in automatic speech recognition: noise reduction and robust audio-visual features. For noise reduction, we describe a noise reduction method in which speech and noise are mapped into different subspaces by the sparse representation to reduce the noise. Our proposed method can be deployed not only on audio noise reduction but also on visual noise reduction for several types of noise. For the second issue, we investigate two feature fusion methods—late feature fusion and the joint sparsity model method—to calculate audio-visual sparse representation features to improve the accuracy of the audio-visual speech recognition. Our proposed method can also contribute to feature fusion for the audio-visual speech recognition system. Finally, to evaluate the new sparse representation features, a database for audio-visual speech recognition is used in this research. We show the effectiveness of our proposed noise reduction on both audio and visual cases for several types of noise and the effectiveness of audio-visual feature determination by the joint sparsity model, in comparison with the late feature fusion method and traditional methods.

Keywords: Audio-visual speech recognition, Sparse representation, Noise reduction, Joint sparsity model

PACS number: 43.72.Ne [doi:10.1250/ast.35.17]

1. INTRODUCTION

The automatic speech recognition (ASR) system using only acoustic speech signals has achieved excellent results in clean environments, but the recognition performance becomes degraded when acoustic speech is corrupted by noise. To enhance the robustness of ASR in noisy or real environments, multimodal speech recognition, in which several features are used jointly, has been investigated and found to increase the robustness and improve the accuracy of ASR. The integration of multimodal features is more informative than the use of single-modal features individually. When one signal becomes corrupted, the missing information can be extracted from the others.

For the multimodal ASR, audio-visual speech recognition is often employed [1–3], which uses visual features,

typically lip information, in addition to the acoustic features to improve the ASR accuracy. The audio-visual ASR has achieved better performance than the audio-only ASR when the audio signal is corrupted by noise, and it can also achieve a slight improvement when the audio is clean.

Nevertheless, there are still some challenges [4] to create a robust audio-visual ASR. In a real environment, for example, in a car, not only the speech signal but also the visual signal is often corrupted by noise for various reasons, such as the presence of a camera, light, transformation and compression. Therefore, a noise reduction method for both speech and visual signals is still categorized into challenging tasks for the audio-visual ASR system. Another challenge is multimodal feature integration. There have been some approaches aimed at a robust integration. Concatenative feature fusion is a traditional method that is utilized widely. Hierarchical discriminant feature fusion [5] is also studied by seeking low-dimensional representations that cause inadequate

*e-mail: simon@asr.info.gifu-u.ac.jp

†e-mail: tamura@info.gifu-u.ac.jp

‡e-mail: hayamizu@gifu-u.ac.jp

modeling owing to the curse of dimensionality. Audio feature enhancement [6] on the basis of either visual features or concatenated audio-visual features also belongs in the study of the feature fusion field.

Recently, the theory of sparse representation (SR) [7,8] has been one of the hot topics in the signal processing and optimization communities. SR is known as a type of sampling theory and relies on the theory that many types of signals can be well approximated by a sparse expansion in terms of a suitable basis, that is, a certain signal can be represented by a small number of linear incoherent measurements. A similar denotation is compressed sensing (CS), which is used to investigate ways in which we can sample signals at roughly the “information rate” rather than the Nyquist rate [7]. CS is often used in sampling strategy and reconstruction algorithms. The SR technology in this paper is also known as an exemplar-based technique [9] similar to k -nearest neighbors (kNNs) and support vector machines (SVMs), which are often used to characterize a signal from a few support training signals. SR is typically used for source separation and pattern classification. Blind source separation [10] and noise reduction [11] using SR have shown the effectiveness of the SR in the source separation fields. For pattern classification tasks, SR has been shown to offer an improvement in accuracy over kNNs and SVMs and the Gaussian mixture model (GMM) methods [9]. Recently, it has also shown superior performance in face recognition [12] over linear SVM and 1-NN methods. A work on a large-vocabulary 50-hour broadcast news task demonstrated the benefit of using the audio SR classification features for large-scale tasks [13].

In this paper, we explore a robust audio-visual speech recognition [1,14], which has been motivated by the emerging theory of SR noise reduction. The method reformulates the noise reduction problem as an SR problem. For a test vector, given an over-complete training dictionary consisting of speech and noise samples, we can represent the test vector as a linear combination of all training samples subject to a sparseness constraint on the coefficient. The nonzero coefficients reveal the true class of the speech sample, and ideally, the speech samples will be mapped into the speech example category and the noise samples will be mapped into the noise example category of the dictionary. Then, we can achieve new clean speech features by choosing the coefficients of the speech example category.

The main work in this paper is twofold. First, we investigate the effectiveness of the proposed method on both acoustic and visual speech. Expressway noise and city road noise in a car, and white noise are used to evaluate the robustness of the proposed method for acoustic speech. For visual speech, we evaluate our system under the following noise conditions: Salt&Pepper noise, Gaussian

noise, and Poisson noise. Second, the proposed audio-visual ASR system uses new audio-visual features obtained from the audio and visual features before the recognition. Two proposed methods are investigated to create the audio-visual SR features. One is a traditional feature fusion method that combines time-synchronous audio SR features and visual SR features. The other method uses a joint sparsity model to measure the audio and visual samples together by solving the SR problem. To the best of our knowledge, such SR-based methods have not yet been studied for audio-visual signals.

This paper is organized as follows. In Sect. 2, we introduce a basic, general framework of SR technology, and then, the SR noise reduction method used in this research is discussed. Section 3 describes an audio-visual ASR system, a database, and the features used in this work. In Sect. 4, our proposed methods are introduced. We introduce our experiments and results in Sect. 5. In Sect. 6, we give a discussion on the results. Finally, we conclude our research and describe our future work in Sect. 7.

2. SPARSE REPRESENTATION FEATURES

The SR features utilized in this research are calculated by solving an SR noise reduction problem. In this section, we will first illustrate the basic, general SR technique and how to represent speech samples in the SR framework. Then, a noise reduction method motivated by the SR technology is introduced.

2.1. Sparse Representation Formulation

Let us denote an input vector $\mathbf{y} \in R^d$, a dictionary matrix $\mathbf{A} \in R^{d \times n}$ ($d < n$) consisting of training vectors, and an unknown vector $\mathbf{x} \in R^n$, such that $\mathbf{y} = \mathbf{A}\mathbf{x}$. In a system of linear equations, $\mathbf{y} = \mathbf{A}\mathbf{x}$, if the dictionary \mathbf{A} is over-determined, the solution can be uniquely determined by taking the pseudo-inverse, $\mathbf{y} = \mathbf{A}\mathbf{x}$, which is a linear least-squares problem. The problem can be solved by l_1 minimization [15]:

$$(P_1) : \min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (1)$$

Since $d < n$, and if \mathbf{x} is sufficiently sparse and \mathbf{A} is incoherent to the basis in which \mathbf{x} is sparse, the solution can be uniquely recovered by solving (P_1) .

The (P_1) problem can be solved using several l_1 -min solvers, including orthogonal matching pursuit (OMP) [16], basis pursuit (BP) [17], and LASSO [18]. The OMP solver works better when \mathbf{x} is very sparse; therefore, in this work, the OMP is used to solve the (P_1) problem. OMP is also a fast solver for the data of our work.

To create a set of SR features, first, matrix \mathbf{A}_i was created with d -dimensional training samples v_i taken from class i as columns; in other words, $\mathbf{A}_i = [v_{i,1}, v_{i,2}, \dots, v_{i,n_i}] \in R^{d \times n_i}$. Given sufficient training sam-

ples from the i -th class, any new (test) sample $\mathbf{y} \in R^d$ from the same class with dimension d will approximately lie in the linear span of the training samples associated with class i , that is,

$$\mathbf{y} = x_{i,1}\mathbf{v}_{i,1} + x_{i,2}\mathbf{v}_{i,2} + \dots + x_{i,n_i}\mathbf{v}_{i,n_i}. \quad (2)$$

Since the membership i of the test sample is initially unknown, we define a new matrix \mathbf{A} as the entire training set to include training samples from k classes:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k] = [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \dots, \mathbf{v}_{k,n_k}], \quad (3)$$

where $\mathbf{A} \in R^{d \times N}$, d is the dimension of each feature vector \mathbf{x} , and N is the total number of all training samples from all classes. Then, the linear representation of \mathbf{y} can be rewritten in terms of all training samples as

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (4)$$

where $\mathbf{x} = [0, \dots, 0, x_{i,1}, x_{i,2}, \dots, x_{i,n_i}, 0, \dots, 0]^T \in R^N$ is a coefficient vector, which ideally should be sparse and should be zero for the elements in \mathbf{A} , except those associated with the same class as \mathbf{y} .

Given a series of speech samples \mathbf{y} , in this paper we use the mel-scale frequency cepstral coefficients (MFCCs) for audio samples and eigenlip components for visual samples. Then, we can get the speech sample set Y , $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_\tau\}$, where τ is the total number of the sample set. For a speech sample, and given an appropriate \mathbf{A} , we solve the problem $\mathbf{y} = \mathbf{A}\mathbf{x}$ subject to a sparseness constraint on the coefficient vector \mathbf{x} . As Sainath *et al.* discuss [13], the dominant nonzero coefficients in \mathbf{x} reveal the true class of the speech sample. With the new \mathbf{x} , a corresponding vector $\mathbf{A}\mathbf{x}$ is formed. Furthermore, the given series of speech sample sets can be represented as $\{\mathbf{A}_1\mathbf{x}_1, \mathbf{A}_2\mathbf{x}_2, \dots, \mathbf{A}_\tau\mathbf{x}_\tau\}$.

2.2. Noise Reduction via Sparse Representation

When a speech signal is corrupted by noise, we can consider the noise to be an additive in the speech signal in the time domain; then, a noisy signal, m_t , can be written as

$$m_t = s_t + n_t, \quad (5)$$

where s_t is a clean speech signal and n_t is a noise signal in time t . When the SR problem $\mathbf{y} = \mathbf{A}\mathbf{x}$ is applied to a noisy signal, Eq. (4) can be rewritten as

$$\mathbf{y} = \mathbf{y}_s + \mathbf{y}_n = [\mathbf{A}_s\mathbf{A}_n][\mathbf{x}_s^T\mathbf{x}_n^T]^T = \mathbf{A}\mathbf{x}, \quad (6)$$

where \mathbf{x}_n indicates a vector of the noise exemplars and \mathbf{A}_n indicates a dictionary matrix containing noise exemplars. \mathbf{x}_s and \mathbf{A}_s indicate the vector of the speech sample and a matrix containing speech sample exemplars. The vectors of \mathbf{y} and \mathbf{x} are feature parameters. For the case of visual speech, signals m , s , and n are facial images for each time

frame t and we use eigenlip expansions as feature parameters for \mathbf{y} and \mathbf{x} .

Equation (6) shows a linear noise reduction method, whereas an MFCCs domain has the nonlinear relation;

$$F(m_t)^2 = F(s_t)^2 + F(n_t)^2 \quad (7)$$

where $F(m_t)$ denotes the Fourier transform of time signal m_t , and $F(m_t)^2$ denotes the power spectrum of m_t . Mel-frequency cepstrum coefficients (MFCCs) are obtained by taking logarithm of power spectrum in mel-scaled frequency and their cosine transform. When the vectors, \mathbf{y} , \mathbf{x}_s and \mathbf{x}_n are MFCCs, the following equation stands as approximation;

$$\exp(\mathbf{y}) = [\mathbf{A}_{\exp(s)}\mathbf{A}_{\exp(n)}][\mathbf{x}_s^T\mathbf{x}_n^T]^T = \mathbf{A}\mathbf{x}, \quad (8)$$

where \mathbf{y} , s , and n are MFCCs features of noisy speech, clean speech, and noise, respectively. Note that parameters in MFCCs are linear for channel distortion, such as difference in microphones or transmission lines.

To reduce the noise in the speech signal, we first construct a new dictionary matrix \mathbf{A} as the entire training set including not only the clean speech samples from all k classes but also the noise samples. Then, for a given speech sample corrupted by noise, we solve Eq. (8) and get coefficient vector \mathbf{x} , so that the dominant nonzero coefficients in \mathbf{x} reveal the true class of the speech sample. Therefore, ideally, speech sample \mathbf{y}_s will be mapped into the clean speech sample category and \mathbf{y}_n will be mapped into the noise sample category of the dictionary matrix \mathbf{A} . Finally, given \mathbf{A}_s and \mathbf{x}_s , a corresponding vector $\mathbf{A}_s\mathbf{x}_s$ is formed; hence, the clean speech sample can be described as

$$\mathbf{y}_s = \mathbf{A}_s\mathbf{x}_s. \quad (9)$$

3. AUDIO-VISUAL SPEECH RECOGNITION

In this section, we introduce a corpus for our audio-visual ASR, then, a basic audio-visual ASR system based on HMMs is presented. Finally, we describe the audio and visual features used in our system.

3.1. CENSREC-1-AV Database

The corpus CENSREC-1-AV (CENSREC: Corpora and Environments for Noisy Speech REcognition) was described in [14]. CENSREC-1-AV is an evaluation framework for an audio-visual ASR system in noisy environments. The data in CENSREC-1-AV are constructed by concatenating eleven Japanese connected utterances of digits from zero to nine, silence (sil), and short pause (sp). Each utterance consists of 1–7 digits, each digit being pronounced as ichi (1), ni (2), san (3), yon (4), go (5), roku (6), nana (7), hachi (8), kyu (9), and zero or maru (0). The corpus was recorded in an office environment with one

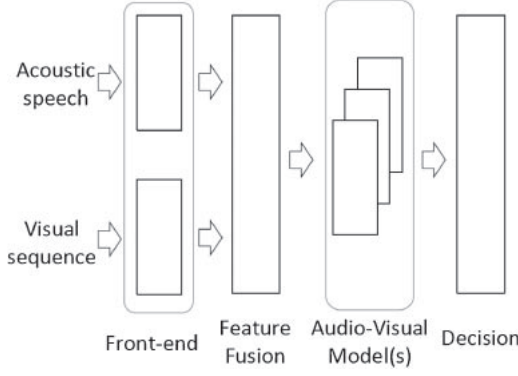


Fig. 1 Simplified general architecture for early-fusion bimodal recognition.

microphone and two cameras, one for color movies and the other for infrared pictures obtained using a lens filter. The sampling frequency of speech is 48 kHz and the frame rate of visual data is 29.97 Hz (NTSC standard). The original image size of the two movies is 720×480 pixels. The visual data were recorded from the front, showing the upper part of the body, with little movement.

CENSREC-1-AV includes a training data set and a testing data set. The training data set is used to build audio and visual models as well as to compute eigenvectors for visual parameterization. It consists only of clean audio and clean visual data comprising 3,234 utterances in total, spoken by 20 female and 22 male speakers. 1,963 utterances were collected in the testing data set spoken by 26 female and 25 male subjects, none of whom participated in the preparation of the training set.

The above database was chosen because it includes not only speech data and mouth images but also a baseline system. We can easily evaluate our own audio-visual SR features using this database.

3.2. Audio-Visual ASR System

Our speech recognition system is based on hidden Markov models (HMMs) to train the acoustic, visual, and audio-visual models. The HMMs are left to right; each digit HMM has sixteen states, and an HMM for silence has three states. The central state in the silence HMM and the state in the short-pause HMM are shared (tied state). The model training method in CENSREC-1-AV is a late fusion bimodal [19], and the training of the HMMs is first performed for each of the stream, acoustic and visual signals separately. Then, a multistream HMM is constructed with the acoustic HMM and the visual HMM. In this research, an early fusion bimodal training method (Fig. 1) is used for training the multistream HMM to evaluate our audio-visual features calculated using the joint sparsity model. The acoustic and visual models are trained separately by the same method as the visual training

method in CENSREC-1-AV. In the multistream HMM, an output log likelihood is computed as

$$b_{av}(\mathbf{o}_{av}) = \lambda_a b_a(\mathbf{o}_a) + \lambda_v b_v(\mathbf{o}_v), \quad (10)$$

where \mathbf{o}_a is an acoustic feature and \mathbf{o}_v is a visual feature, and $\mathbf{o}_{av} = (\mathbf{o}_a^T, \mathbf{o}_v^T)^T$, T being a transposition. Audio and visual log likelihoods are denoted by $b_a(\mathbf{o}_a)$ and $b_v(\mathbf{o}_v)$, respectively. λ_a and λ_v are stream weighting factors. Testing is performed with the stream weights λ_a and λ_v that are optimized under the constraint

$$\lambda_a + \lambda_v = 1. \quad (11)$$

The audio-only testing is carried out with the acoustic model prepared in the training procedure, and visual-only testing is performed with the visual model. The audio-visual testing is performed with the multistream model, and the audio stream weight λ_a is tested from 0.5 to 1.0 with 0.1 step, to obtain the best accuracy.

3.3. Audio and Visual Features

To create the audio features, 12-dimensional MFCCs and a static log power, and their first and second derivatives are extracted from an audio frame. As a result, a 39-dimensional audio feature is obtained every 10 ms. Unlike the training data, the testing data include not only the clean audio but also noisy data. To ensure that the recognition accuracy of the baseline covers a wider range, the audio features at several SNR levels (5 dB, 0 dB and -5 dB) of in-car noises recorded on an expressway and a city road, and SNR levels (15 dB, 10 dB and 5 dB) of white noise, bubble noise, classical noise, and piano noise are also extracted.

Our visual features are extracted from the clean visual data. To extract visual features, the rectangle around the mouth of the speaker is located first, then, downsizing and gray-scale conversion are applied on the mouth image, yielding a 1,040-dimensional signal vector. Principle component analysis (PCA) is conducted using all the training vectors. We extract a 30-dimensional visual feature, which includes 10-dimensional eigenlip components [3] and their Δ and $\Delta\Delta$ coefficients. Since the visual frame rate is 29.97 Hz, which is different from that of audio parameters, feature interpolation is subsequently conducted using a 3-degree spline function to make the feature rate 100 Hz, same as the audio rate. The visual features with noise are also prepared for our experiment. The visual noises are Salt&Pepper noise with noise densities of 0.03 and 0.05, Gaussian white noise with zero mean and 0.01 and 0.02 variances, and Poisson noise.

4. PROPOSED METHOD

In this section, we present two proposed methods to validate the effectiveness of the SR features on our audio-

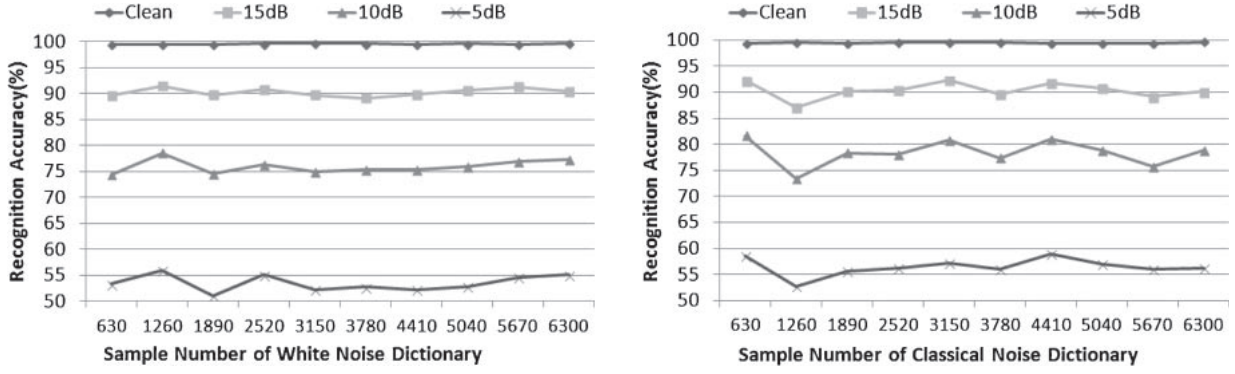


Fig. 2 Influence of dictionary size.

visual ASR system. We will first describe a method of constructing the dictionary of \mathbf{A} . Then, we describe a proposed noise reduction method for acoustic and visual signals, and two proposed methods to improve the robustness of audio-visual ASR.

4.1. Dictionary Matrix \mathbf{A}

In most speech applications of SR, the dictionary either consists of fundamental basis functions, such as Fourier coefficients, wavelets, and speech spectrographic representations, or is *learned* [13]. In this work, we represent a given sample by a few training samples of the dictionary in the MFCCs or PCA domain by solving the SR noise reduction problem. Therefore, we construct our dictionary matrix \mathbf{A} with the clean speech samples and noise samples that are chosen on the basis of phone classes. A time-aligned transcription [14] of the training data is used to locate the frame number of a phone class. The phone list used in the CENSREC-1-AV database includes seventeen phones and *sil*. For phone class p_i ($i = 1, 2, \dots, 18$), we randomly select a phone segment $p_{i,x}$ ($x = 1, 2, \dots, q$) corresponding to the phone class i from the training data set; q is the selected phone segment number of training data in each class. Then, the selected phone segment of the phone class p_i can be written as

$$\mathbf{A}_{p_i} = [p_{i,1}, p_{i,2}, \dots, p_{i,q}], \quad (12)$$

where \mathbf{A}_{p_i} is the selected phone segment of the phone class i and q is sixty in this work. The frame length of $p_{i,x}$ is about five to thirty. For every phone segment $p_{i,x}$, we randomly select three frames f_j after cutting the first and last 10% of the phone segment, that is,

$$p_{i,x} = [f_{i,x,1}, f_{i,x,2}, f_{i,x,3}]. \quad (13)$$

To support noise reduction, we also select noisy samples $\mathbf{A}_{\text{sil,SNR}}$ from the nonspeech segment with the SNR levels 5 dB, 0 dB, and -5 dB for acoustic samples. The audio dictionary \mathbf{A}_a can be written as

$$\mathbf{A}_a = [\mathbf{A}_{p_1}, \dots, \mathbf{A}_{p_{18}}, \mathbf{A}_{\text{sil,5dB}}, \dots, \mathbf{A}_{\text{sil,-5dB}}]. \quad (14)$$

We create an audio dictionary \mathbf{A}_a and a corresponding visual dictionary \mathbf{A}_v for calculating audio and visual SR features. Because we have only one noise level for visual samples, we select the visual noise samples $\mathbf{A}_{\text{sil,noise}}^v$ three times to keep the length of the visual dictionary the same as that of the audio dictionary.

$$\mathbf{A}_v = [\mathbf{A}_{p_1}^v, \dots, \mathbf{A}_{p_{18}}^v, \mathbf{A}_{\text{sil,noise}}^v, \dots, \mathbf{A}_{\text{sil,noise}}^v] \quad (15)$$

Finally, a multistream dictionary \mathbf{A}_{av} consisting of audio and visual samples is obtained by integrating the audio dictionary \mathbf{A}_a and the visual dictionary \mathbf{A}_v .

An experiment was carried out to evaluate the performance changes with dictionary size. Two data sets are chosen; one is white noise as stationary noise and classical music as nonstationary noise. We change the value of q as 10, 20, \dots , 100 so that the dictionary size becomes 630, 1,260, \dots , 6,300.

Figure 2 shows the influence of dictionary size. The left graph indicates the results for stationary noise (white noise). From the results, we can see that the performance fluctuates when the sample number is smaller than 3,150. The accuracy improves as the number of samples in the dictionary becomes larger than 3,150, especially at the SNR levels of 10 dB and 5 dB. For nonstationary noise (Fig. 2 right graph), we can see that the performance also fluctuates with a small-size dictionary. However, it is less stable than the stationary noise when the sample number is greater than 3,150.

In our experiments, the frame length of one phone segment is five to thirty, so we require a certain number of samples to better represent a phone. However, for the small-size dictionary, for example, 630, only ten phone segments for each phone were used to select the samples. We believe this amount to be insufficient for a stable performance. When the dictionary size is small, the performance will depend heavily on the quality of the selected samples. Because we chose the samples from our

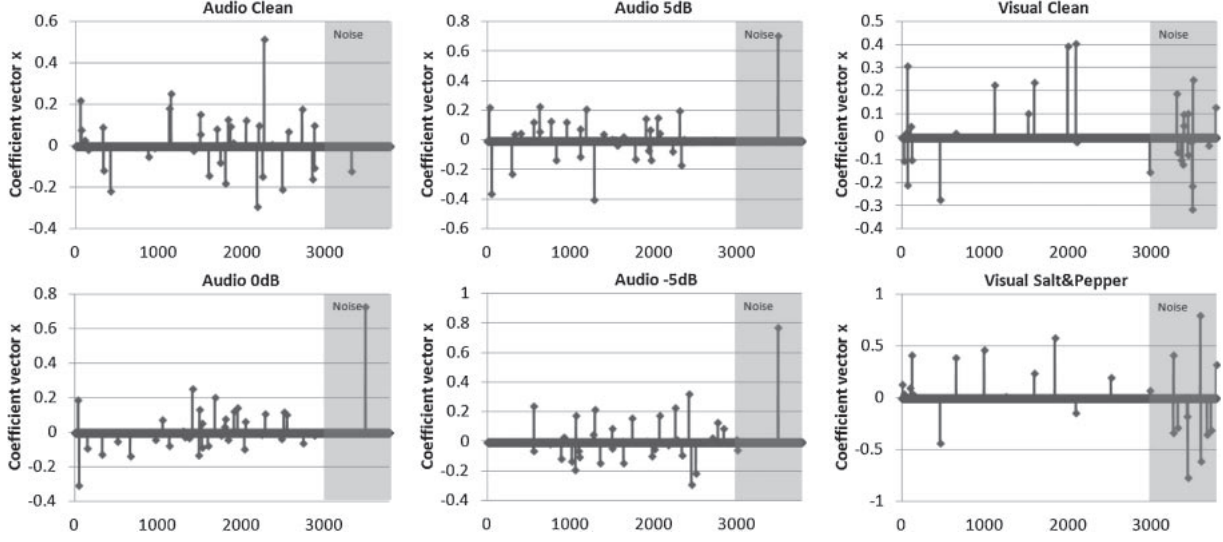


Fig. 3 Coefficient vectors x of clean, 5 dB, 0 dB, and -5 dB of audio, and clean and Salt&Pepper of visual for the speech file FBJ_135A (the input vector is one of a frame of phoneme i with frame number 66). The horizontal axis of each image represents the frame number of the dictionary. The vertical axis is the value of x .

training data set randomly, the performance might drop when the selected samples are insufficient and inappropriate. Compared with stationary noise, nonstationary noise might require more samples to achieve a stable performance. Considering the computation complexity and the performance, we chose the dictionary size of 3,780 (q is 60).

4.2. Acoustic SR Feature

To reduce the noise, for example, expressway noise, in speech signals by our proposed method, we first construct a new dictionary matrix A consisting of clean samples and noisy samples with SNR levels of 5 dB, 0 dB, and -5 dB. For a given speech sample y , we solve Eq. (8), and get coefficient vector x . Then, with Eq. (9), we can get clean speech. The left four graphs in Fig. 3 show the coefficient value x for a given speech sample. The samples of dictionary from 3,240 to 3,780 are noise samples. We know that the coefficient value of clean speech is seldom mapped into the noise sample category, but for 5 dB, 0 dB, and -5 dB samples, the noise sample category has an important influence.

Six types of noise are prepared to evaluate the effectiveness of the acoustic SR features: expressway noise and city road noise in a car, classical and piano music noise, bubbles, and white noise. We create different dictionary matrices A for all the noises.

4.3. Visual SR Feature

When a visual signal is corrupted by additive noise, we expect our proposed method to still be effective. The right two graphs in Fig. 3 show the coefficient values x for a

given visual sample. From these two figures, we can see that not only the noisy sample but also the clean sample is mapped into the noise example category.

Three types of noise for visual features are also prepared for our experiment. They are Salt&Pepper noise with a noise density of 0.05, Gaussian white noise with zero mean and 0.01 variance, and Poisson noise. We create different dictionary matrices for each noise type.

4.4. Audio-Visual: Late Feature Fusion

In this experiment, the SR features are created using the method described in the previous section. In the traditional audio-visual ASR system, audio features and visual features are obtained, and audio-visual features are then integrated with the audio features and visual features. Finally, the audio-visual features are tested with the audio-visual model. In this experiment, firstly, the audio SR features and visual SR features are created separately. To create the audio SR features y_a^{sr} , we use the audio dictionary A_a and solve the problem $y_a = A_a x_a$ with the noise reduction method Eqs. (6), (9). By the same method, we can obtain the visual SR features y_v^{sr} . Then, the two SR features are integrated into audio-visual SR features. Figure 4 depicts the feature extraction process. In this figure, the left-hand side shows the extraction method used in this subsection. The SR features are created for both training data and testing data including all the audio noise conditions.

$$y_a^{sr} = A_a x_a \quad (16)$$

$$y_v^{sr} = A_v x_v \quad (17)$$

$$y_{av}^{sr} = ((y_a^{sr})^T, (y_v^{sr})^T)^T \quad (18)$$

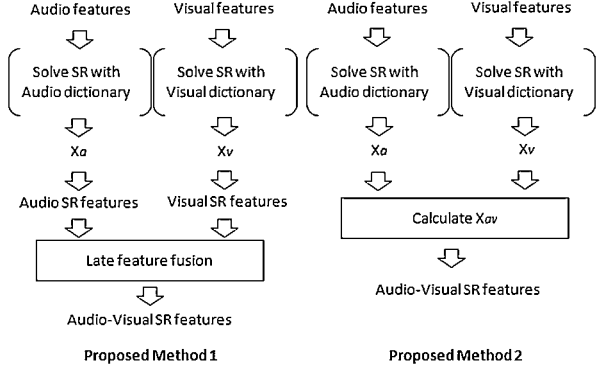


Fig. 4 Proposed methods for creating audio-visual SR features. Proposed Method 1 is the late feature fusion method and Proposed Method 2 is the joint sparsity model method.

4.5. Audio-Visual: Joint Sparsity Model

In the previous method, the audio features and visual features were created by solving the SR problem. This means that the audio and visual signals can be treated as separate problems so there are no influences between the two SR features when creating the SR features. The narrow-band array processing and localization using the sparsity model is already known from a previous study [20], in which a joint sparsity model was suggested and the localization robustness was explored. In this experiment, because the audio signal and visual signal give large and different contributions to the recognition accuracy, we extend the joint sparsity model to match our needs.

We solve the audio and visual SR problems separately and then we create the audio-visual coefficient \mathbf{x}_{av} . The audio-visual SR features are created with this new audio-visual coefficient \mathbf{x}_{av} . The right-hand side in Fig. 4 illustrates how the audio-visual SR feature can be obtained. Firstly, Eq. (19) is used to obtain the audio-visual coefficient \mathbf{x}_{av} with \mathbf{x}_a and \mathbf{x}_v . w is the audio stream weighting factor. Then, we will construct our dictionary matrix \mathbf{A}_{av} for the joint sparsity model in the same manner as just described in the previous section.

$$\mathbf{x}_{av} = \mathbf{x}_a \times w + \mathbf{x}_v \times (1 - w) \quad (19)$$

$$\mathbf{A}_{av} = (\mathbf{A}_a^T, \mathbf{A}_v^T)^T \quad (20)$$

$$\mathbf{y}_{av}^{sr} = \mathbf{A}_{av} \mathbf{x}_{av} \quad (21)$$

This method is used to create audio-visual SR features for both training data and test data, and the audio stream weight w is changed from 1.0 to 0.4 with 0.1 steps.

5. EXPERIMENTS

In this section, the experiments are conducted using the methods described in the previous section.

5.1. Experiment Using Acoustic SR Feature

In this experiment, we created the audio SR features by solving Eq. (8) with dictionary matrix \mathbf{A} , which consists of audio samples. This method was used to create audio SR features for both training data and test data. Then, the acoustic model was learned using the training data. The test data were divided into two data sets: one is a development data set that is used to optimize the recognition parameters, i.e., an insertion penalty, and the other subdata set is used to evaluate our proposed method. The development data set was created by selecting five data for every speaker, so we have 255 data in total in the development data. The other 1,708 data were used as the testing data set.

Figure 5 shows the recognition accuracy for audio only for expressway, city road, white, bubble, classical music, and piano music noises by both spectrum subtraction (SS) [21] and the proposed method. Figure 6 shows the average recognition accuracy of Fig. 5 for nonstationary noise and stationary noise. Results for the baseline system are also included for comparison. In the experiments for expressway, city road, white, bubble, classical music, and piano music noises, the dictionary matrices with the corresponding noise samples were used. To evaluate unmatched conditions, the unmatched experiments were also performed. We processed the expressway noise speech data with the dictionary consisting of city road noise samples. For the city road, white, bubble, classical, and piano data, we used the dictionary consisting of expressway, white 5 dB samples, bubble 15 dB samples, different classical music, and different piano music, respectively. The results show that the SS method has better performance than the proposed method for stationary noise: expressway, city road, and white noise. For nonstationary noise, the proposed method has better performance than SS. For clean SR features, the performance was almost the same as that with the baseline features. The result of unmatched conditions shows that the proposed method is applicable for unmatched conditions.

5.2. Experiment Using Visual SR Feature

The procedure of the visual SR feature experiment is the same as the audio SR feature experiment. For the visual data, three types of noise were prepared to evaluate our proposed method. We created a dictionary matrix for each type of noise. Training was also carried out for each of them with the clean training data set.

Figure 7 shows the recognition accuracy only for visual Salt&Pepper, Gaussian, and Poisson noises by both median filtering and our proposed methods. Unmatched-condition experiments were performed for Salt&Pepper, Gaussian, and Poisson noises with a dictionary consisting of Salt&Pepper noise with a density of 0.03 and Gaussian noise with zero mean and 0.02 variance and zero mean

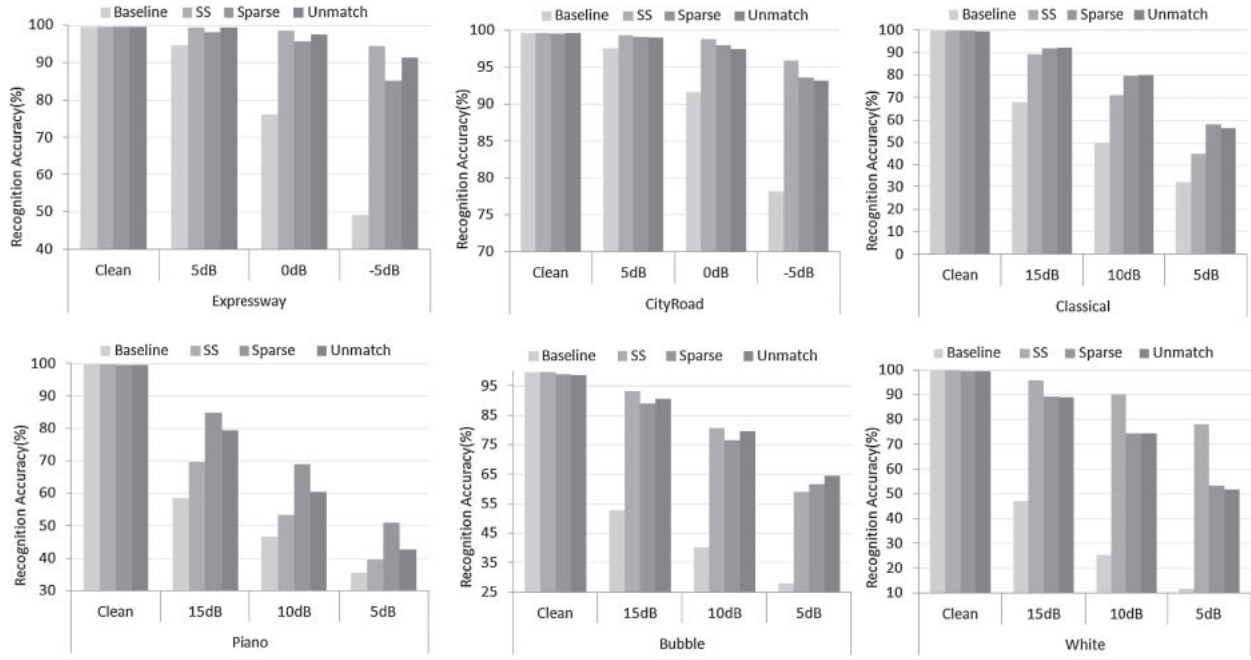


Fig. 5 Recognition accuracy for audio only for expressway, city road, classical, piano, bubble, and white noises by the baseline, spectrum subtraction (SS), and the proposed method (matched/unmatched).

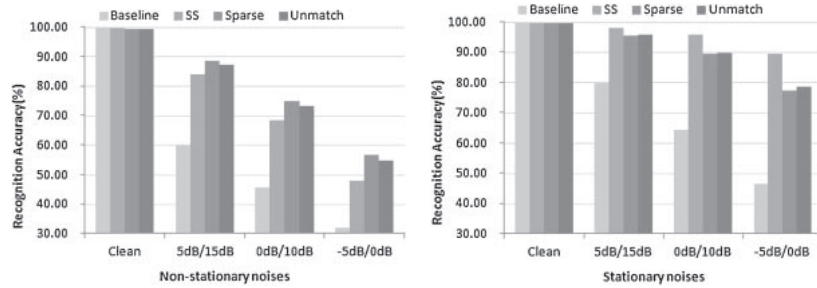


Fig. 6 Average recognition accuracy for audio only for nonstationary noise (classical, piano, and bubble) and stationary noise (expressway, city road, and white) by the baseline, spectrum subtraction (SS) and the proposed method (matched/unmatched).

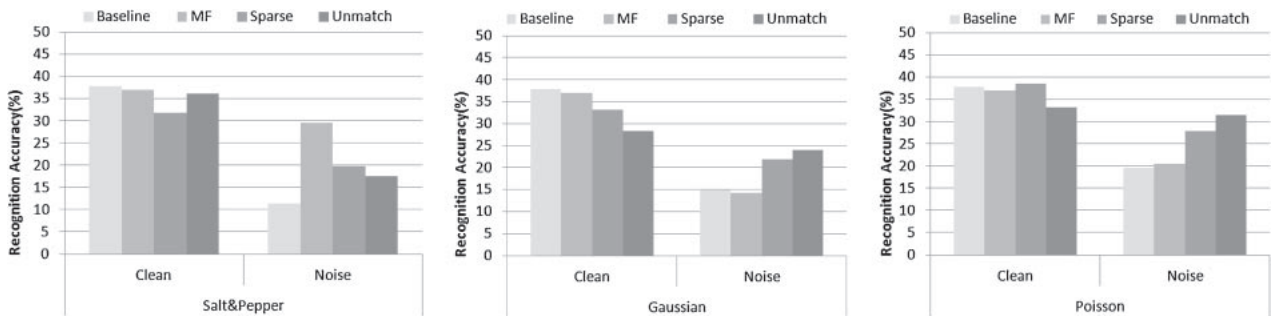


Fig. 7 Recognition accuracy for visual only for Salt&Pepper noise, Gaussian noise, and Poisson noise by the baseline, median filtering (MF) and the proposed method (matched/unmatched).

and 0.01 variance, respectively. The results show that the median filtering method has better performance for Salt&Pepper noise because it is an effective method for

reducing Salt&Pepper noise. Nevertheless, it cannot improve the performance for Gaussian noise and achieved only 0.89% better than the baseline. The performance of

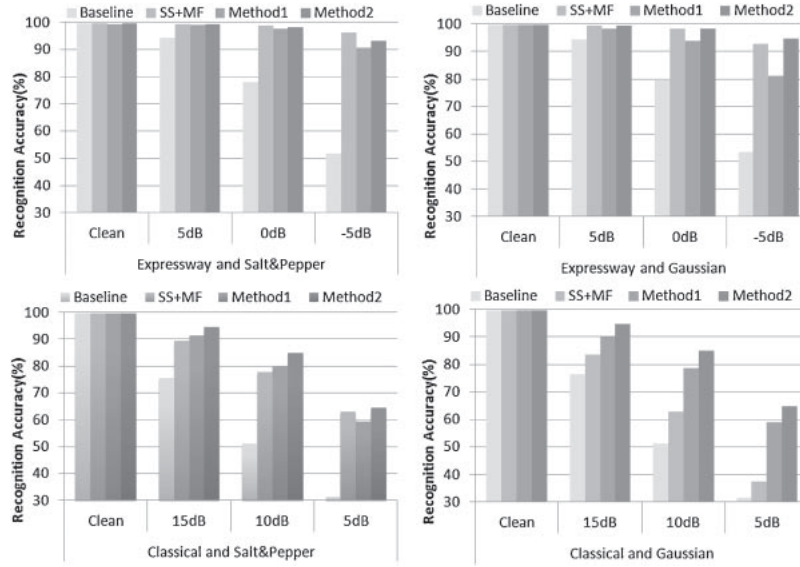


Fig. 8 Recognition accuracy of method 1 (Late feature fusion) and method 2 (joint sparsity model).

our proposed method achieved slightly lower performance under clean conditions of Salt&Pepper and Gaussian noises. In contrast, it achieved 8.5%, 6.96%, and 8.17% better than the baseline features of Salt&Pepper, Gaussian, and Poisson noises, respectively. Compared with the baseline noise results, the unmatched-conditions experiment results had better accuracies and were better than those of matched-condition experiments when using a heavily noisy dictionary.

5.3. Experiment Using Late Feature Fusion and Joint Sparsity Model

We created four data sets to evaluate our proposed method; for audio noise we used nonstationary classical music noise and stationary expressway noise, for visual noise, Salt&Pepper noise with a density of 0.05, which is random white and black noise, and Gaussian noise, which is a distributed noise, were used. The audio-visual SR features were created by two methods: late feature fusion and joint sparsity model. The audio-visual ASR system described in the previous chapter is used to learn the audio-visual models with the audio-visual training SR features and to test the audio-visual ASR system with the audio-visual testing SR features. We optimize the recognition parameters with the development data set to achieve the best performance for each condition.

Figure 8 shows the recognition accuracy for the audio-visual results by both method 1 (late feature fusion) and method 2 (joint sparsity model). The baseline and SS+MF results (audio with SS and visual with MF) are also included for comparison. From the result, we can know that proposed method 1 and method 2 yielded results better than the baseline results. Compared with proposed methods

1 and 2, for the classical and Gaussian data set, method 2 achieved 93.17% when the SNR was 5 dB, which is more than 6% better than the performance of method 1 when the stream weight of audio signals was 0.8. The best accuracy was achieved when the stream weight of audio signals was 1.0 for SNRs of 15 dB and 10 dB. This means that our audio-visual SR features are created using only the audio classification coefficient. For the expressway and Salt&Pepper data set, the best accuracy was achieved when the audio stream weight was 0.9 for 15 dB, 10 dB, and 5 dB. In general, the recognition accuracy of method 2 is better than that of method 1 for all four data sets, and only for the expressway and Salt&Pepper data set did the proposed method 2 yield better results compared with the SS+MF method.

6. DISCUSSION

As discussed regarding the results of audio-only SR features in Figs. 5 and 6 in the previous section, the SS method has better performance than our proposed method in the cases of stationary noise, such as expressway and white noises. This result means that the use of MFCCs, compared with SS, limited the noise reduction. Nevertheless, our proposed method achieved better performance than the SS method in the case of nonstationary noise. In the SS method, the noise spectrum must be estimated and subtracted from the noisy speech spectrum. To estimate the noise spectrum, nonspeech segments are often used. This is based on the assumption that the noise is stationary. Thus, in the case where the noise is nonstationary, the SS technique has a limitation in noise spectrum estimation. However, our proposed methods are aimed at mapping the noise speech samples into clean samples and noise

samples, so ideally, if the noise sample is sufficient, our dictionary can cover the features of nonstationary noise.

Similarly, Fig. 7 shows that the median filtering method has better performance than our proposed method for Salt&Pepper noise. This result also means that the use of PCA features, compared with image processing of a median filter, induced some limitation of visual noise reduction. However, our proposed method had better performance than the baseline for not only the Salt&Pepper but also Gaussian and Poisson noises.

Compared with the conventional techniques, SS can achieve better performance for stationary noise, and the median filtering method achieves better performance for Salt&Pepper noise. The SR method has improved performance for various noises including stationary audio noise and Salt&Pepper visual noise, and it is better than SS for nonstationary noises and median filtering for Gaussian and Poisson noises. In particular, the SR method is an effective method that can be employed not only for audio noise reduction but also for visual noise reduction. Thus we confirmed that our proposed method can simplify the complexity of the audio-visual ASR system.

Figure 8 shows the audio-visual SR features of proposed methods 1, 2, baseline, and SS+MF. The audio and visual SRs were separately performed in proposed method 1. This means that SR for audio features was performed using only audio data. In contrast, the audio-visual SR was conducted so that the SR for the audio signals utilized not only audio but also visual information in proposed method 2. Our results show that computing the audio and visual features together using the same dictionary matrix is a useful robust technique of audio-visual ASR.

7. CONCLUSION AND FUTURE STUDY

In this paper, we proposed an audio-visual ASR system with an SR framework to create a robust ASR system. The SR noise reduction and the joint sparsity model were also utilized in our proposed system. We set up experiments to examine the effectiveness of the SR features. The results showed that the SR features offer a significant improvement for both audio-only and visual-only features. The joint sparsity model can improve the performance for audio-visual SR features compared with the SR features obtained by the late feature fusion method.

For future work, there are two topics to be studied. One is the extension of our SR technique for audio-visual ASR in a manner similar to the recent work [9] on an audio-only framework. Another is the study of the relationship between the SR technique and model adaptation [22], because it is related to the audio-visual interaction for model adaptation.

REFERENCES

- [1] S. Tamura, K. Iwano and S. Furui, "A stream-weight optimization method for multistream HMMs based on likelihood value normalization," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 469–472 (2005).
- [2] G. Potamianos and C. Neti, "Automatic speechreading of impaired speech," *Proc. Int. Conf. Auditory-Visual Speech Processing*, pp. 177–182 (2001).
- [3] C. Miyajima, K. Tokuda and T. Kitamura, "Audiovisual speech recognition using MCE-based HMMs and model-dependent stream weights," *Proc. Int. Conf. Spoken Language Processing*, pp. 1023–1026 (2000).
- [4] G. Potmianos, C. Neti, J. Luetttin and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson and P. Perrier, Eds. (MIT Press, Cambridge, MA, 2004).
- [5] G. Potamianos, J. Luetttin and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 165–168 (2001).
- [6] L. Girin, J.-L. Schwartz and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Am.*, **109**, 3007–3020 (2001).
- [7] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Trans. Signal Process. Mag.*, **25**, 21–30 (2008).
- [8] D. Donoho, M. Elad and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, **52**, 6–18 (2006).
- [9] T. N. Sainath, A. Carmi, D. Kanevsky and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4370–4373 (2010).
- [10] Z. H. Wu, Y. Shen, Q. Wang, J. Liu and B. Li, "Blind source separation based on compressed sensing," *Proc. 6th Int. ICST Conf. Communications and Networking in China (CHINACOM)*, pp. 794–798 (2011).
- [11] M. N. Schmidt, J. Larsen and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," *IEEE Workshop, Machine Learning for Signal Processing*, pp. 431–436 (2007).
- [12] J. Wright, A. Y. Yang and A. Ganesh, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**, 210–227 (2009).
- [13] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky and A. Sethy, "Sparse representation features for speech recognition," *Proc. INTERSPEECH*, pp. 2254–2257 (2010).
- [14] S. Tamura, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," *Proc. Int. Conf. Auditory-Visual Speech Processing*, pp. 85–88 (2010).
- [15] A. Yang, A. Ganesh, Z. Zhou, S. Sastry and Y. Ma, "Fast l_1 -minimization algorithms and an application in robust face recognition: A review," *Tech. Rep. UCB/EECS-2010-13*, UC Berkeley (2010).
- [16] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, **41**, 3397–3415 (1993).
- [17] S. Chen, D. Donoho and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, **43**, 129–159 (2001).
- [18] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc., Ser. B*, **58**, 267–288 (1996).
- [19] C. C. Chibelushi, F. Deravi and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, **4**, 23–37 (2002).

- [20] D. Malioutov, M. Cetin and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, **53**, 3010–3022 (2005).
- [21] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 208–211 (1979).
- [22] S. Tamura, M. Oonishi and S. Hayamizu, "Audio-visual interaction in model adaptation for multi-modal speech recognition," *Proc. Int. Conf. APSIPA ASC*, Xi'an, China, PID. 15, 4 pp. (2011).



Peng Shen received his M.S. degree in computer science and technology from Shandong Agriculture University, China in 2003. Then he works as an information management engineer in Lenovo from 2004 to 2007. He received his M.S. degree in information science from Gifu University, Japan in 2010, where he is currently a doctoral course student. His research interests

are multimodal speech recognition, robust speech recognition and pattern recognition. He is a student member of ASJ and IEEE.



Satoshi Tamura received M.S. and Ph.D. degrees in information science and engineering from Tokyo Institute of Technology, Japan in 2002 and 2005 respectively. He became a research associate at Faculty of Engineering, Gifu University, Japan in 2005. Currently he has been an assistant professor in Gifu University from 2007. His research interests are speech information processing, such as multimodal speech recognition, robust speech recognition and application of speech recognition to real system. He is also interested in audio-visual information processing. He is a member of ASJ, IEICE, IPSJ, JSAI, JAMI, and ISCA.



Satoru Hayamizu received his B.E., M.E., and Ph.D. degrees from The University of Tokyo in 1978, 1981, and 1993, respectively. From 1981 to 2001, he worked at Electro-technical Laboratory, AIST, Ministry of International Trade and Industry. From 2001 to 2002, he worked at National Institute of Advanced Industrial Science and Technology. Since 2002, he has been a professor at Faculty of Engineering, Gifu University. His research interests are speech processing, pattern recognition, machine learning and media informatics. He is a member of ASJ, IEICE, IPSJ, JSAI, IEEE, ACM, and ISCA.