ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Audio-visual interaction in sparse representation features for noise robust audio-visual speech recognition

Peng Shen[1], *Satoshi Tamura*[2], *and Satoru Hayamizu*[2]

[1]Graduate School of Engineering, Gifu University, 1-1 Yanagido, Gifu City, 501-1193, Japan
[2]Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu City, 501-1193, Japan
simon@asr.info.gifu-u.ac.jp, tamura@info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

Abstract

In this paper, we investigate audio-visual interaction in sparse representation to obtain robust features for audio-visual speech recognition. Firstly, we introduce our system which uses sparse representation method for noise robust audio-visual speech recognition. Then, we introduce the dictionary matrix used in this paper, and consider the construction of audio-visual dictionary. Finally, we reformulate audio and visual signals as a group sparse representation problem in a combined feature-space domain, and then we improve the joint sparsity feature fusion method with the group sparse representation features and audio sparse representation features. The proposed methods are evaluated using CENSREC-1-AV database with both audio noise and visual noise. From the experimental results, we showed the effectiveness of our proposed method comparing with traditional methods.
Index Terms: sparse representation, audio-visual speech recognition, feature fusion, noise reduction, joint sparsity model

1. Introduction

The audio-visual automatic speech recognition (ASR) system [1], [2], using both acoustic speech features and visual features has been investigated and found to increase the robustness and improve the accuracy of ASR. 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. To improve the performance of the system, noise reduction method was often employed on speech signal. Nevertheless, in real environment for example in a car, not only the speech signal but also the visual signal are often corrupted by audio and visual noise. Therefore, noise reduction method for both speech and visual signals is still categorized into challenging tasks for the audio-visual ASR system.

Recently, sparse representation (SR) [3] has gained considerable interests in signal processing. SR is known as a type of sampling theory, which 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, we can represent a certain signal with a small number of linear incoherent measurements. A robust audio-visual speech recognition system which has been motivated by the emerging theory of SR noise reduction [4],[5] was introduced. It shows effectiveness on both acoustic and visual speech respectively, and the feature fusion methods on audio-visual SR features were discussed.

In this paper, we investigate audio-visual interaction of the audio-visual dictionary matrix, and we reformulate audio and visual signals as a group sparse representation problem in a combined feature-space domain, and then we proposed a feature fusion method with the group sparse representation features and audio sparse representation features.

2. Sparse Representation Features

2.1. Sparse Representation Formulation

Consider an input vector $y \in R^d$, and a dictionary matrix $A \in R^{d \times n}(d < n)$ consisting of training vectors, and an unknown vector $x \in R^n$, such that $y = Ax$. If the dictionary A is overdetermined, the linear equations, $y = Ax$ can be uniquely determined by taking the pseudo-inverse $y = Ax$, which is a linear least squares problem. The problem can be solved by l_1 minimization:

$$(P_1): \quad argmin||x||_1 \quad subject \quad to \quad y = Ax. \quad (1)$$

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

There are several l_1-min solvers which can be used to solve the (P_1) problem, including orthogonal matching pursuit (OMP) [6], basis pursuit (BP), and LASSO. In this work, we use the OMP method to solve the (P_1) problem. The OMP solver works better when x is very sparse, and OMP is also a fast solver for the data of our work.

2.2. Noise Reduction via Sparse Representation

The speech signal is observed in an additive noise, then, a noisy signal m_t, can be written as

$$m_t = s_t + n_t, \tag{2}$$

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

$$\mathbf{y} = \mathbf{y}_s + \mathbf{y}_n = [A_s A_n][\mathbf{x}_s^T \mathbf{x}_n^T]^T = A\mathbf{x}, \tag{3}$$

where $\mathbf{x}_n$ indicates a vector of the noise exemplars and A_n indicates a dictionary matrix containing noise exemplars. $\mathbf{x}_s$ and A_s indicate the vector of speech sample and a matrix containing speech sample exemplars. The vector of $\mathbf{y}$ and dictionary A are feature parameters. For the case of visual speech, a signal m, s, and n are facial images for each time frame t and we use eigenlip expansions as feature parameters for $\mathbf{y}$ and A.

Equation 3 shows a linear noise reduction method, whereas an MFCC domain has the non-linear relation.

$$F(m_t)^2 = F(s_t)^2 + F(n_t)^2 \tag{4}$$

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

$$exp(\mathbf{y}) = exp(\mathbf{s}) + exp(\mathbf{n}), \tag{5}$$

where $\mathbf{y}$, $\mathbf{s}$, and $\mathbf{n}$ are MFCC features of noisy speech, clean speech and noise, respectively. Note that parameters in MFCC are linear for channel distortion, such as the difference in microphones or transmission lines. The equation 3 can be written as

$$exp(\mathbf{y}) = [A_{exp(\mathbf{s})} A_{exp(\mathbf{n})}][\mathbf{x}_\mathbf{s}^T \mathbf{x}_\mathbf{n}^T]^T = A\mathbf{x}. \tag{6}$$

To reduce the noise in speech signal, a dictionary matrix A is constructed from 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, equation 6 is solved, and a coefficient vector $\mathbf{x}$ is obtained, so that the dominant nonzero coefficients in $\mathbf{x}$ reveal the true class of the speech sample. Therefore, ideally, the 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 A. Finally, a corresponding vector $A_s\mathbf{x}_s$ is formed with A_s and $\mathbf{x}_s$, hence, we can describe the clean speech sample as

$$\mathbf{y}_s = A_s\mathbf{x}_s. \tag{7}$$

3. Database and Features

3.1. CENSREC-1-AV Database

The evaluation framework CENSREC-1-AV [7] for audio-visual ASR system is utilized in this work. The data in CENSREC-1-AV is constructed by concatenating eleven Japanese connected utterance of digits from zero to nine, silence (sil), and short pause (sp). It includes a training data set and a testing data set. The training data consists of 3,234 utterances. 1,963 utterances were collected in the testing data. The testing data set includes not only clean audio and visual data but also noisy data. The audio noisy data were created by adding in-car noises recorded on city road and expressway to clean speech data at several SNR levels. Visual distortion was also conducted by simulating a driving-car condition by a gamma transformation.

3.2. 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 10ms. Different from the training data, the testing data include not only the clean audio and visual data but also noisy data. To ensure the recognition accuracy of baseline covers a wide range, the audio features at several SNR levels (5dB, 0dB, and -5dB) of in-car noises recorded on an expressway, and SNR levels (15dB, 10dB, and 5dB) of classical noise is also extracted.

A 30-dimensional visual feature is also computed, that consists of 10-dimensional eigenlip components [2] and their Δ and $\Delta\Delta$ coefficients. Feature interpolation is subsequently conducted using a 3-degree spline function in order to make the feature rate 100Hz, as same as the audio rate. Salt&Pepper noise with noise density of 0.05, Gaussian white noise with zero mean and 0.01 variance are employed in our work.

4. Methods

4.1. Audio-visual Dictionary Matrix A

In this work, we represent a given sample by a few training samples of the dictionary in the MFCC or PCA domain by solving the SR noise reduction problem. Therefore, our dictionary matrix A is constructed with the clean speech samples and noise samples that are chosen on the base of phone classes. A time-aligned transcription [7] 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 $p_i (i = 1, 2, \ldots, 18)$, we randomly select a phone segment $p_{i,x} (x = 1, 2, \ldots\ q)$ 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

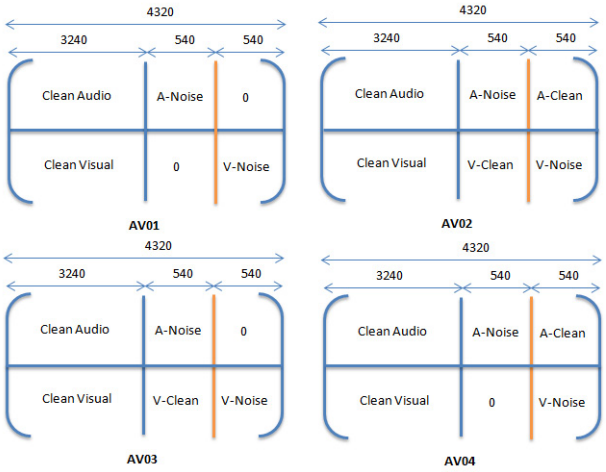

Figure 1: The construction of audio-visual dictionary matrix.

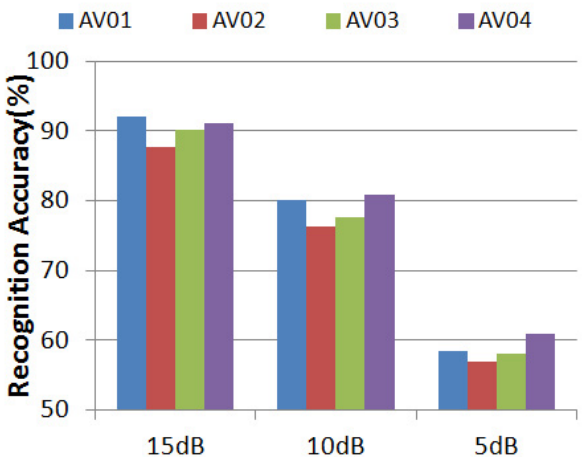

Figure 2: The influence of dictionary construction.

selected phone segment of the phone class p_i can be written as

$$A_{p_i} = [p_{i,1}, p_{i,2}, \ldots, p_{i,q}], \quad (8)$$

where 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. In every phone segment $p_{i,x}$, three frames f_j are randomly selected after cutting the start and last 10% of frames of the phone segment, that is

$$p_{i,x} = [\mathbf{f}_{i,x,1}, \mathbf{f}_{i,x,2}, \mathbf{f}_{i,x,3}]. \quad (9)$$

To support noise reduction, we also select the noisy samples $A_{sil,SNR}$ from the nonspeech segment with the SNR levels of 5dB/15dB, 0dB/10dB and -5dB/5dB for acoustic samples. The audio dictionary A_a can be written as

$$A_a = [A_{p_1}, \ldots, A_{p_{18}}, A_{sil,5dB}, \ldots, A_{sil,-5dB}]. \quad (10)$$

We create an audio dictionary A_a and a corresponding visual dictionary 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 $A^v_{sil,noise}$ three times to keep the length of visual dictionary as same as the audio dictionary.

$$A_v = [A^v_{p_1}, \ldots, A^v_{p_{18}}, A^v_{sil,noise}, \ldots, A^v_{sil,noise}]. \quad (11)$$

Finally, a multi-stream dictionary A_{av} consisting of audio and visual samples is obtained by integrating the audio dictionary A_a and the visual dictionary A_v.

Figure 1 shows four types of multi-stream dictionary. These dictionaries (AV01,..., AV04) are used to evaluate the interaction of audio and visual modalities when we solve the audio-visual sparse representation problem. An experiment was performed to evaluate the interaction of audio and visual when the audio-visual SR problem is solved.

Figure 2 shows the results of this experiment to investigate the influence of dictionary construction with the joint sparsity model (proposed method 3). The classical and gaussian data set was utilized in this experiment. From the results, we can know that AV04 gets the best recognition accuracy, especially in the SNR 10dB and 5dB. Therefore, the dictionary of AV04 was chosen to perform our experiments.

In this paper, we investigated four methods to obtain audio-visual SR features. Method 1 and 2 have already been proposed in [5], to compare with the proposed methods, we introduce them briefly. Method 3 and 4 are the proposed methods in this paper.

4.2. Method 1: Late Feature Fusion

In this method, firstly, audio SR features and visual SR features are created separately. To create the audio SR features $\mathbf{y}^{sr}_a$, we use the audio dictionary A_a and solve the equation $\mathbf{y}_a = A_a\mathbf{x}_a$ with the noise reduction method equations 3 and 7. In the same way, we can obtain the visual SR features $\mathbf{y}^{sr}_v$. Then, the two SR features are integrated into audio-visual SR features. Figure 3 depicts the feature extraction process. In this figure, the most left-hand graph shows the extraction method used in this subsection.

$$\mathbf{y}^{sr}_a = A_a\mathbf{x}_a \quad (12)$$

$$\mathbf{y}^{sr}_v = A_v\mathbf{x}_v \quad (13)$$

$$\mathbf{y}^{sr}_{av} = ((\mathbf{y}^{sr}_a)^T, (\mathbf{y}^{sr}_v)^T)^T \quad (14)$$

4.3. Method 2: Late Feature Fusion(w/t weight)

The second left graph in figure 3 illustrates how the audio-visual SR features of this method can be obtained. Firstly, we solve the audio and visual SR problems separately as same as method 1. Then, equation 15 is utilized to obtain the audio-visual coefficient $\mathbf{x}_{av}$ with $\mathbf{x}_a$

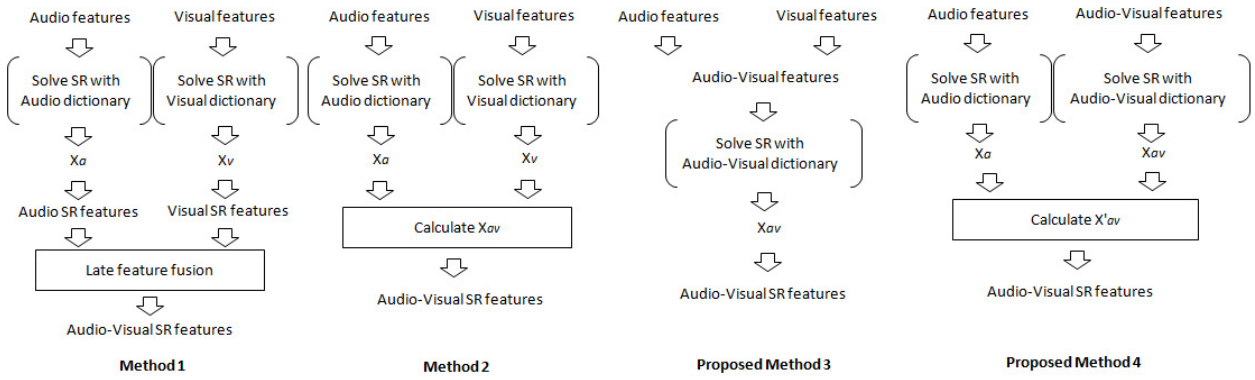

Figure 3: Feature fusion methods

and $\mathbf{x}_v$. w is an audio stream weighting factor. And the dictionary matrix A_{av} is constructed with equation 16. Finally, equation 17 is utilized to create the audio-visual SR features with the audio-visual coefficient $\mathbf{x}_{av}$ and the dictionary matrix A_{av}.

$$\mathbf{x}_{av} = \mathbf{x}_a \times w + \mathbf{x}_v \times (1 - w) \tag{15}$$

$$A_{av} = (A_a^T, A_v^T)^T \tag{16}$$

$$\mathbf{y}_{av}^{sr} = A_{av}\mathbf{x}_{av} \tag{17}$$

4.4. Method 3: Joint Sparsity Model(JSM)

The previous two methods, audio features and visual features were obtained by solving the SR problem respectively. It means that the audio and visual SR can be treated as separated problems, therefore, there are no influences between the two SR features when solving the SR problem. The narrow-band array processing and localization using sparsity model is already known in the literature [8], in which a joint sparsity model was suggested and localization robustness was explored. In this experiment, the integrated audio-visual features $\mathbf{y}_{av}$ and audio-visual dictionary A_{av04} was used to solve the SR problem $\mathbf{y}_{av} = A_{av04}\mathbf{x}_{av}$.

4.5. Method 4: Feature Fusion(w/t weight & JSM)

Method 3 treats audio and visual with no discrimination when solve the SR problem. Because the audio signal and visual signal have a huge different contribution to the recognition accuracy, the results of method 3 will be affected because of the difference. An audio-only adaptation / multimodal visual adaptation method [9] investigated the multimodal visual adaptation, it showed the effectiveness of using multi-modal visual adaptation instead of visual adaptation. In this experiment, we extend the joint sparsity model by replacing the visual coefficient $\mathbf{x}_v$ with audio-visual coefficient $\mathbf{x}_{av}$ to match our needs.

The most right-hand graph in figure 3 illustrates how the audio-visual SR features can be obtained. Firstly, we solve the audio-visual SR problem with the audio-visual dictionary as same as method 3 to obtain an audio-visual coefficient $\mathbf{x}_{av04}$, and solve the audio SR problem to obtain an audio coefficient $\mathbf{x}_a$. Then, equation 18 is utilized to obtain the audio-visual coefficient $\mathbf{x}_{av}^{'}$ with $\mathbf{x}_a$ and $\mathbf{x}_{av}$. w is the audio stream weighting factor. Finally, with the equation 19, the new audio-visual SR features can be obtained.

$$\mathbf{x}_{av}^{'} = \mathbf{x}_a \times w + \mathbf{x}_{av04} \times (1 - w) \tag{18}$$

$$\mathbf{y}_{av}^{sr} = A_{av04}\mathbf{x}_{av}^{'} \tag{19}$$

5. Experiments

In this work, we created audio-visual SR features (y_{av}^{sr}) for both training data and test data. Then, models were learned using the training data. The test data was used to evaluate the proposed methods. For methods 2 and 4, we changed the weight w from 0.1 to 1.0 with 0.1 steps to obtain the best recognition accuracy.

Figure 4 shows the recognition accuracies for audio-visual SR features of the proposed methods. The baseline results of CENSREC-1-AV, which is a standard audio-visual ASR system, and method 1 and 2 results are also included for comparison. In these experiments two audio and two visual noises were prepared. For the classical and gaussian data set, we can know that method 3 obtained better accuracy than method 1, but for the expressway and Salt&Pepper data set, method 1 obtained better accuracy. The audio-only recognition results [5] showed that with the SR method, the accuracy of expressway in -5dB can be improved from 49.11% to 85.23%, however, the classical data set in 10dB, the accuracy can be improved from 49.84% to 79.34% due to non-stationary noise. And the visual-only of Salt&Pepper and gaussian can only be improved 8.5% and 6.96% respectively. We can see that,

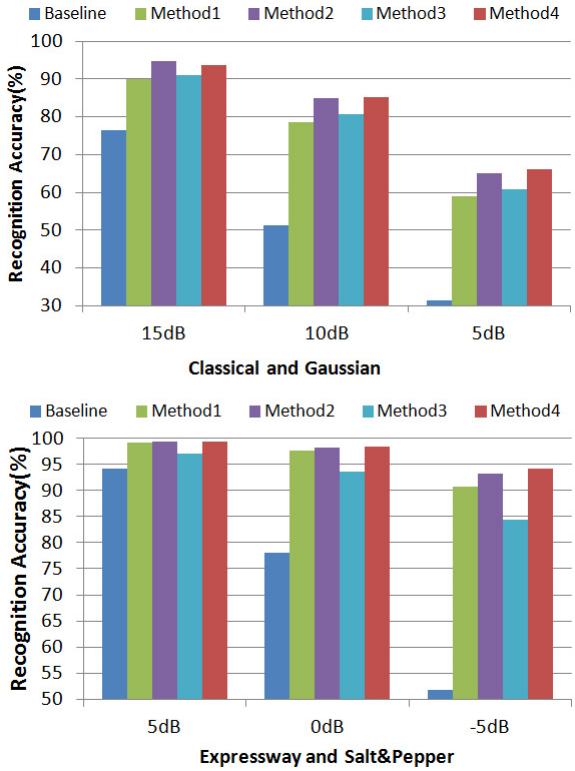

Figure 4: Recognition accuracy of proposed methods.

when the difference of the two streams recognition accuracy is large, method 3 is not a robust method. In contract, when the difference is smaller, method 3 is better than method 1.

From the results of method 4, the best recognition accuracy was obtained when the weight is 1.0 in 15dB, 0.7 in 10dB and 0.7 in 5dB for the classical and gaussian data set, and they are 0.7 (5dB, 0dB and -5dB) for the expressway and Salt&Pepper data set. Compared with method 2, method 4 improved the performance on both data sets, achieved 1.28% for classical and gaussian data set in 5dB, and 1.01% for expressway and Salt&Pepper data set in -5dB. Method 4 uses audio-visual coefficients instead of visual coefficients to calculate the SR features. Therefore, the classification accuracy using the joint sparsity model and audio-only is better than using audio-only and visual-only.

6. Conclusions

In this paper, we focus on audio-visual ASR system with an SR noise reduction framework to create a robust ASR system. Then, we proposed a joint sparsity feature fusion method which uses audio-visual interaction to improve the traditional feature fusion method. The experiment results showed that the joint sparsity model is an effectiveness method to create audio-visual SR features. For future work, there are still some work need to do to improve the current method and simplify the complexity of the system to create real-time system.

7. Acknowledgment

The part of this work was supported by JSPS KAKENHI Grant (Grant-in-Aid for Young Scientists (B)) No.25730109.

8. References

[1] S. Tamura, K. Iwano, and S. Furui, “A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization,” Proc. Int. Conf. ICASSP2005, vol.1, pp.469-472 (2005).

[2] C. Miyajima, K. Tokuda, T. Kitamura, “Audiovisual speech recognition using MCE-based HMMs and model-dependent stream weights,” Proc. Int. Conf. ICSLP2000, vol.2, pp.1023-1026 (2000).

[3] E.J. Candes and M.B. Wakin, “An Introduction To Compressive Sampling,” IEEE Trans. Signal Processing Magazine, vol.25, no.2, pp.21-30 (2008).

[4] P. Shen and S. Tamura, and S. Hayamizu, “Feature reconstruction using sparse imputation for noise robust audio-visual speech recognition,” Proc. Int. Conf. Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, no.125 (2012).

[5] P. Shen and S. Tamura, and S. Hayamizu, “Multi-Stream Sparse Representation Features for Noise Robust Audio-visual Speech Recognition,” Trans. Acoustical Science and Technology (In press).

[6] S.G. Mallat and Zhifeng Zhang, “Matching pursuits with time-frequency dictionaries,” IEEE Trans. Signal Processing, vol.41, no.12, pp.3397-3415 (1993).

[7] S.Tamura et al., “CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition,” Proc. Int. Conf. AVSP2010, pp.85-88 (2010).

[8] D. Malioutov, M. Cetin, and A. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” IEEE Trans. Signal Processing, vol.53, no.8, pp.3010-3022 (2005).

[9] S.Tamura, M. Oonishi, and S. Hayamizu, “Audio-visual Interaction in Model Adaptation for Multimodal Speech Recognition,” Proc. Int. Conf. APSIPA ASC2011, Xi’an, China, PID:15 (2011).