Feature Reconstruction using Sparse Imputation for Noise Robust Audio-Visual Speech Recognition

Peng Shen*, Satoshi Tamura† and Satoru Hayamizu†
* Graduate School of Engineering, Gifu University, Gifu, Japan E-mail: simon@asr.info.gifu-u.ac.jp
† Department of Information Science, Gifu University, Gifu, Japan

***Abstract*—In this paper, we propose to use noise reduction technology on both speech signal and visual signal by using exemplar-based sparse representation features for audio-visual speech recognition. First, we introduce sparse representation classification technology and describe how to utilize the sparse imputation to reduce noise not only for audio signal but also for visual signal. We utilize a normalization method to improve the accuracy of the sparse representation classification, and propose a method to reduce the error rate of visual signal when using the normalization method. We show the effectiveness of our proposed noise reduction method and that the audio features achieved up to 88.63% accuracy at -5dB, a 6.24% absolute improvement is achieved over the additive noise reduction method, and the visual features achieved 27.24% absolute improvement at gamma noise.**

I. Introduction

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. In order to improve the performance of the system, noise reduction method was often employed on the 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, a 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 interest 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. The SR technology is similar to k-Nearest Neighbors (kNNs) and Support Vector Machines (SVMs) which are also known as exemplar-based techniques [4] used to characterize a signal from a few support training signals. SR is typically used for source separation and pattern classification. Blind source separation [5] and noise reduction [6] using SR have shown the effectiveness of SR in the source separation fields.

In this paper, we explore a robust audio-visual speech recognition, which has been motivated by emerging the theory of sparse imputation noise reduction. The speech noise can be often considered as additive in the speech signal in the time domain. The visual noise is a γ-transformation on the images to change the lighting conditions by simulating a car-driving condition, so feature reconstruction method by sparse imputation noise reduction is employed for not only audio but also visual signals. In this proposed method, given a test vector which is corrupted by noise and an over-complete training dictionary consisting of speech and visual examples with the noise, we can then represent the test vector as a linear combination of all training examples subject to a sparseness constraint on a coefficient vector. The nonzero coefficients reveal the true class of the speech sample. Then we can reconstruct the features by using the coefficients and the dictionary with only clean examples. A normalization method was used in this work to improve the accuracy of SR classification, and some changes also were done on the visual features to employ the sparse imputation noise reduction on the visual signal and improve the accuracy of the system.

II. Sparse Representation Features

A. 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 the l_1-minimization:

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

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 can be used to solve the (P_1) problem, including Orthogonal Matching Pursuit (OMP) [8], 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.

In order to create a set of SR features, first, consider a series of speech samples $Y = \{y_1, y_2, \ldots, y_n\}$, and a matrix A as the entire training set to include training samples from all k classes. For a speech sample y_n, we solve the problem $y_n = Ax_n$ subject to the sparseness constraint on the coefficient vector x_n. The dominant nonzero coefficients

in x_n reveal the true class of the speech sample. With the new x_n, a corresponding vector Ax_n is formed. Consequently, the given series of speech sample set can be represent as $Y' = \{Ax_1, Ax_2, , \ldots, Ax_n\}$.

B. Method 1: Addition Noise Reduction via SR

When a speech signal is corrupted by noise, we can consider the noise is an additive in the speech signal in the time domain, then, an observed signal g_t, can be written as

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

where s_t is a clean speech signal and n_t is a noise signal in time t. When the SR problem $y = Ax$ is applied to a noisy signal, then $y = Ax$ can be rewritten as

$$y = y_s + y_n = [A_s A_n][x_s^T x_n^T]^T = Ax, \tag{3}$$

where A_n indicates a dictionary matrix containing noise exemplars and x_n is the representation of noise via the noise exemplars in the dictionary. A_s and x_s indicate a matrix containing speech sample exemplars and the representation of the speech exemplars.

To reduce the noise in speech signal, we first construct a new dictionary matrix A as the entire training set including not only clean speech samples from all k classes but also the noisy samples. Then for a given speech sample corrupted by noise, we solve the problem (Eq.3), and will get a coefficient vector x, so that the dominant nonzero coefficients in x reveal the true class of the speech sample. Therefore, ideally, the speech sample y_s will be mapped into the clean speech sample partition and y_n will be mapped into the noise sample partition of the dictionary matrix A. Finally given the A_s and x_s, a corresponding vector $A_s x_s$ is formed, hence, the clean speech sample can be described as:

$$y_s = A_s x_s. \tag{4}$$

C. Method 2: Sparse Imputation

For a noise sample y_{noise}, and a matrix A_{noise} including training samples with the same noise, we solve the problem $y_{noise} = A_{noise}x$ subject to a sparseness constraint on the coefficient vector x. In order to reconstruct the input sample, we create a dictionary A_{clean} including only clean exemplars, the training samples of the clean dictionary correspond to the noise samples of the noise matrix. Then we reconstruct the SR features with the clean dictionary A_{clean}, in other words, the new SR clean feature can be described as $A_{clean}x$.

III. Database and Features

A. CENSREC-1-AV Database

An 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 digit utterances 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 (20dB, 15dB, 10dB, 5dB, 0dB and -5dB). Visual distortion was also conducted by simulating a driving-car condition by a gamma transformation. The gamma noise in this work is four times stronger than the baseline.

B. 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 at every 10ms. Different from the training data, the testing data includes not only the clean audio and visual data but also noisy data. In this paper, the audio features at several SNR levels (5dB, 0dB and -5dB) of the in-car noises recorded on expressway are also extracted. A 30-dimensional clean and gamma 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 to 100Hz, as same as the audio rate.

IV. Experiments

A. Dictionary Matrix A

In this work, a dictionary matrix A was constructed with the samples which are chosen based on phoneme classes. We use a time-aligned transcription [7] of the training data to locate the frame number of a phoneme class. We have two types of dictionary matrix, one for the additive noise reduction method, another for the sparse imputation method, so that we can compare these two methods.

The phoneme list used in CENSREC-1-AV database includes seventeen phoneme and sil. For a phoneme class $p_i (i = 1, 2, \ldots, 18)$, we randomly select a phone segment $p_{i,x} (x = 1, 2, \ldots n)$ corresponding to the phoneme class i from all the training data set, n is the selected phone segment number of train data in each class. Then the selected phone segment of the phoneme class p_i can be written as:

$$A_{p_i} = [p_{i,1}, p_{i,2}, \ldots, p_{i,n}], \tag{5}$$

where A_{p_i} is the selected phone segment of the phoneme class i, n 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 after cutting the starting and ending 10% frames of the phone segment.

In order to support additive noise reduction (Method 1), we also select the noisy examples $A_{sil,SNR}$ for sil with the SNR levels of 5dB, 0dB and -5dB. As a result, the dictionary A can be written as:

$$A = [A_{p_1}, \ldots, A_{p_{18}}, A_{sil,5dB}, \ldots, A_{sil,-5dB}]. \tag{6}$$

We create an audio dictionary A_a and a corresponding visual dictionary A_v for calculating audio and visual SR features.

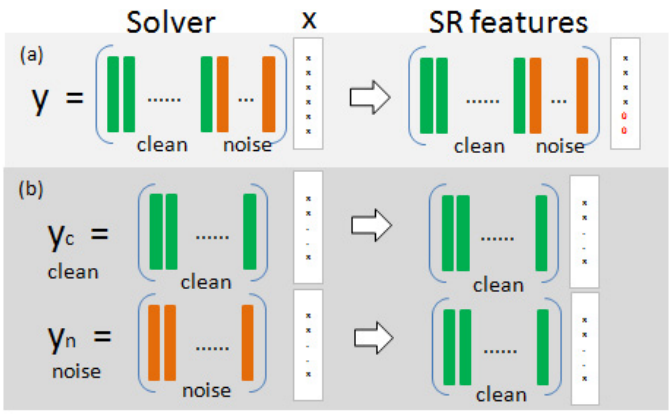

Fig. 1. Two methods of SR features reconstruction. Additive noise reduction method (a); Sparse Imputation method (b).

For the sparse imputation noise reduction method (Method 2), we construct dictionaries for clean and noise data respectively. The exemplars are selected from the clean training data set for clean dictionary and noise training data set for noise dictionaries. We finally got four dictionaries for audio data and two dictionaries for visual data: A_{clean}, A_{5dB}, A_{0dB}, A_{-5dB}, V_{clean} and V_{gamma}.

B. Experiment 1 using Additive Noise Reduction

In this experiment, the noise samples are considered as an additive noise in the speech signal, therefore the SR features are created using the method which describes in the previous section. Figure 1 (a) shows the method in this experiment, where, 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 additive noise reduction method (Eq. 3, 4). Using the same method, we can get the visual SR features y_v^{sr}. And then the two SR features are integrated into audio-visual SR features. The SR features are created for both training data and testing data including all the audio noise conditions: expressway 5dB, 0dB, -5dB.

$$\begin{aligned} y_a^{sr} &= A_a x_a \\ y_v^{sr} &= A_v x_v \\ y_{av}^{sr} &= ((y_a^{sr})^T, (y_v^{sr})^T)^T \end{aligned} \tag{7}$$

In order to improve the performance of the SR classification, a normalization method was used on the input samples y and the dictionary matrix A. An m-th column in the dictionary is normalized with mean u_m and standard deviation σ_m of the column. The normalized column can be described as

$$A'_m = [\frac{a_{m,1} - \mu_m}{\sigma_m}, \frac{a_{m,2} - \mu_m}{\sigma_m}, \ldots, \frac{a_{m,D} - \mu_m}{\sigma_m}]^T \tag{8}$$

Then, with the normalized y' and A', we can get the the new SR features $y_a'^{sr}$, then a reverse normalization is used to reconstruct the features.

Our speech recognition system is based on multi-stream Hidden Markov Models (HMMs), which we chose for their ability to vary the importance of each stream to the recognition. In this research, a early fusion bimodal training method is used for training the multi-stream HMM to evaluate our audio-visual features. The audio-visual testing is done with the multi-stream model, and the audio stream weight is tested with 0.0 and 1.0, so that we can evaluate the accuracy of audio and visual respectively. We optimize the recognition parameters, i.e. an insertion penalty and stream weights, manually to achieve the best performance of audio and visual features.

TABLE I
RECOGNITION ACCURACY OF ADDITIVE NOISE REDUCTION METHOD (METHOD 1.1) AND ADDITIVE NOISE REDUCTION WITH THE NORMALIZATION ON TEST SAMPLE AND DICTIONARY (METHOD 1.2).

	SNR	Baseline	Method 1.1	Method 1.2
Audio only	Clean	99.67	99.49	99.12
	5dB	86.84	91.33	97.55
	0dB	65.05	74.76	93.08
	-5dB	47.55	55.77	82.39
Visual only	Clean	42.29	42.65	42.77
	Gamma	7.20	11.22	11.45

Table I shows the recognition accuracy for audio only, visual only results of the proposed additive noise reduction method (Method 1.1) and additive noise reduction with normalization on test sample and dictionary (Method 1.2). Results for the baseline system are also included for comparison. Looking the result of audio only, SR features of Method 1.2 achieved a recognition rate of 82.39% when the SNR was -5dB, more than 34% better than the baseline features. And the accuracy rate is more than 93% at 0dB. For the clean data, we can see the performance was almost the same as the baseline features. This confirms that the SR features with noise reduction with normalization can significantly improve the performance when the signal is corrupted by noise. Although, the visual dictionary consists of no noise visual exemplars, the result of the gamma noise is still improved with the new SR visual features.

C. Experiment 2 using Sparse Imputation

The sparse imputation noise reduction method described in the previous section can be applied not only to additive noise but also to the distortion of image. In this experiment, we create four dictionaries for audio data and two dictionaries for visual data. Figure 1 (b) shows the process of this method. To create the SR features, for a test input signal y_{0dB}^{sr}, we use the audio dictionary A_{0dB} and solve the problem $y_{0dB} = A_{0dB} x_{0dB}$. and then the clean audio dictionary A_{clean} is used to reconstruct the new SR features, that is $y_{0dB}^{sr} = A_{clean} x_{0dB}$. Similarly, we can reconstruct all the audio SR features $y_{asnr}^{sr}(asnr = \{clean, 5dB, 0dB, -5dB\})$ and all the visual SR features $y_{vsnr}^{sr}(vsnr = \{clean, gamma\})$. And then the two SR features are integrated into audio-visual SR features. The SR features are created for both training data and testing data.

Figure 2 shows visual features of two speaking continuous digits. From this figure we can know that the value of the first dimensional in (a) is positive number, but in (b) is negative number. And the value of the first dimension is much

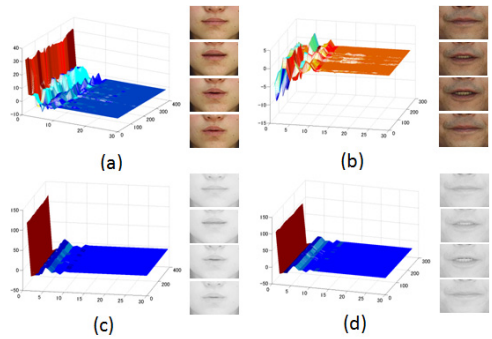

Fig. 2. Visual features of the baseline database. (a), (c) is the clean, gamma data of FBJ-15O3456A, (b), (d) is the clean, gamma of MED-37A. The x axis of each image represents the dimension of the features, the y axis is the value of feature, z is frame number.

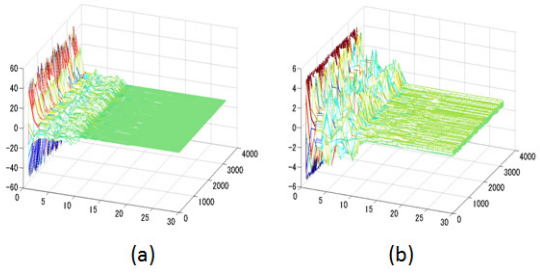

Fig. 3. The features of clean visual dictionary. (a) without normalization, (b) with normalization.

higher than others. Figure 3 shows the difference between the non-normalized and normalized visual features of the clean dictionary, it shows that the difference between the first dimension and other dimension is reduced after normalization. Therefore we create a new set of baseline visual features (29-dimension features) which only consists of from 2nd to 30th parameters to reduce the negative effect of the normalization. A normalization method [7] was used in CENSREC-1-AV to calculate the visual features, when we change the normalization method only using the mean value, we can get the features (new-30-dimension) without the problem we described above.

Table II shows the recognition accuracy results for the proposed method with sparse imputation noise reduction. From the result, we can know that the sparse imputation method achieved a recognition rate of 98.39%, 96.34%, 88.63%, and

TABLE II
RECOGNITION ACCURACY OF SPARSE IMPUTATION METHOD WITH NO NORMALIZATION(METHOD 2.1), WITH NORMALIZATION (2.2) AND BASELINE WHICH THE VISUAL FEATURE IS 30-DIMENSION.

	SNR	Baseline	Method 2.1	Method 2.2
Audio only	Clean	99.67	99.61	99.13
	5dB	86.84	77.74	98.39
	0dB	65.05	65.02	96.34
	-5dB	53.79	55.77	88.63
Visual only	Clean	42.32	42.65	14.94
	Gamma	7.20	28.16	23.27

TABLE III
RECOGNITION ACCURACY OF SPARSE IMPUTATION METHOD WITH NORMALIZATION (METHOD 2.2.2) AND BASELINE WHICH THE VISUAL FEATURE IS 29-DIMENSIONAL(BASELINE29); BASELINE30 IS THE NEW-30-DIMENSIONAL FEATURES; METHOD2.2.30 IS BASED ON BASELINE30.

	Baseline29	Method2.2.29	Baseline30	Method2.2.30
Clean	44.69	28.98	37.82	36.32
Gamma	9.42	31.77	5.97	33.21

0.84%, 3.26%, 6.24% better than the additive noise reduction method on the noisy condition 5dB, 0dB and -5dB respectively. Table III shows the visual results of the 29-dimensional features and new-30-dimensional features of baseline and the sparse imputation noise reduction method. Method2.2.29 with gamma features achieved 31.77%, it is better than the 30-dimensional gamma SR features (Method2.1, 2.2). But with clean features only achieved 28.98% which is lower than the result of Baseline29. Although the accuracy of Baseline30 is lower than baseline and Baseline29, Method2.2.30 achieved 36.32% with clean data and 33.21% with gamma noise. We can know the effectiveness of our proposed method when using the Baseline30 features.

V. CONCLUSIONS

In this paper, we have proposed two methods to reduce the noise both for audio signal and visual signal. Our results show effectiveness of the additive noise reduction method and the sparse imputation noise reduction method. The additive noise reduction method can be employed easily for various noise environment and sparse imputation noise reduction has achieved better performance when the noise environment is confirmed beforehand. And we discussed the difference of the audio and visual features and made some changes and improved the performance of the sparse imputation noise reduction on visual features.

REFERENCES

[1] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," Proc. 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. ICSLP2000, vol.2, pp.1023-1026 (2000).
[3] E.J. Candes and M.B. Wakin, "An Introduction To Compressive Sampling," Signal Processing Magazine, IEEE, vol.25, no.2, pp.21-30 (2008).
[4] T.N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian Compressive Sending for Phonetic Classification," Proc. ICASSP, pp.4370-4373 (2010).
[5] Z.H Wu, Y. Shen, Q. Wang, J. Liu, and B. Li, "Blind Source Separation based on Compressed Sensing," Communications and Networking in China (CHINACOM), 6th International ICST Conference, pp.794-798 (2011).
[6] M.N. Schmidt, J. Larsen, and Hsiao Fu-Tien, "Wind Noise Reduction using Non-Negative Sparse Coding," Machine Learning for Signal Processing, IEEE Workshop, pp.431-436 (2007).
[7] S.Tamura et al., "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP2010, pp.85-88 (2010).
[8] S.G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," Signal Processing, IEEE Transactions, vol.41, no.12, pp.3397-3415 (1993).