JOINT SPARSE REPRESENTATION BASED CEPSTRAL-DOMAIN DEREVERBERATION FOR DISTANT-TALKING SPEECH RECOGNITION

Weifeng Li[†]*, Longbiao Wang*[♯]*, Fei Zhou*[†]*, and Qingmin Liao*[†]

[†] Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua University, China
[♯] Nagaoka University of Technology, Japan

ABSTRACT

In this paper we address reducing the mismatch between training and testing conditions for robust distant-talking speech recognition under realistic reverberant environments. It is well known that the distortions caused by reverberation, background noise, etc., are highly nonlinear in the cepstral domain. In this paper we propose to capture the complex relationships between clean and reverberant speech via joint dictionary learning. Given a test reverberant speech with a sequence of feature vectors we first find their sparse representations, and then estimate the underlying clean feature vectors using the dictionary of clean speech. Based on speech recognition experiments conducted under realistic reverberation conditions, the proposed method is shown to perform very well, resulting in an average relative improvement of 59.1% compared with the baseline front-ends.

Index Terms— sparse representation, reverberation-robust speech recognition, blind dereverberation, Mel-Frequency Cepstral Coefficients (MFCCs).

1. INTRODUCTION

The focus of this paper is on one-microphone robust distant-talking ASR under reverberation conditions, in which additive distortions and reverberation of the desired signal mainly hamper ASR. Reducing reverberation through deconvolution (inverse filtering) is one of the most common approaches. The main problem is that the reverberation energy must be very well estimated for successful deconvolution. Because both the speech signal and the reverberation are nonstationary signals, dereverberation to obtain clean speech from the convolution of nonstationary speech signals and impulse responses is far from trivial. Blind dereverberation (e.g., [1] [2]) is an extremely challenging task, since neither the room impulse response (RIR) describing the acoustic path between speaker and microphone nor the speaker signal is available. Dereverberation via suppression [3] and enhancement [4] treats the late reverberation as an additive noise, and spectral subtraction (SS) based techniques can be applied for reverberation compensation. However, their drawback is that the reverberation compensation is undesirable if the late reverberation is not estimated precisely. For robust speech recognition, it is sufficient to estimate a clean speech feature sequence. Cepstral Mean Normalization (CMN) [5] is a simple and effective way of reducing the channel distortion by normalizing the cepstrl feature vectors. Although CMN is effective for the early reverberation , it works not so well in a distant-talking environment in which the duration of the impulse response usually has a much longer tail. [6] and [7] increase the robustness of an ASR system to reverberation through variance compensation in cepstral domain and explicitly modeling reverberation in log-melspectral domain, respectively. However, they are mainly based on linear model. It is well known that in realistic distant-talking environments the distortions caused by reverberation and/or additive noise are highly nonlinear in the cepstral domain.

Recently sparse representation (SR) [8] have been growing interest in signal processing and pattern recognition. In SR a signal is approximated by a linear combination of a few atoms from a pre-defined dictionary. SR techniques have been applied to different fields, such as compressive sensing [9], face recognition [10], phone recognition [11], etc. However, most existing SR methods only consider sparse modeling in a single signal space, and few considers dictionary learning across different signal spaces. In reverberation-robust ASR, we have two coupled feature spaces (e.g., clean and reverberant feature spaces), and the two coupled spaces are usually related by some mapping function, which could be nonlinear as mentioned above. In such cases, it is often desirable to learn representations that can not only well represent each signal space individually, but also capture their relationships through the underlying sparse representations.

In this paper, we propose a joint sparse representation (JSR) technique, in which a joint dictionary learning is performed across the clean and reverberant feature spaces, for feature dereverberation with application to robust speech recognition. In the proposed joint dictionary learning, the two sparse codes are expected to capture possible complex relationships between the clean and reverberant feature spaces by sharing the same representation coefficients. For any given testing reverberant feature vectors we first find their sparse representation coefficients, and then estimate the underlying clean feature vectors, which are used for the input of the speech recog-

nition system. Our proposed method is learning-based and does not need any explicit reverberation model. Our experiments, conducted on realistic reverberation data, demonstrate that the proposed method is capable of significantly reducing the mismatches between the training and test conditions.

2. PROPOSED JSR METHOD

2.1. Single Sparse Representation (SR)

Let $\mathbf{x}^c$ and $\mathbf{x}^r$ be the feature vectors (e.g., Mel-Frequency Cepstral Coefficients (MFCC)) of clean and reverberation speech, respectively. In our case, a single sparse representation (SR) could approximate $\mathbf{x}^c$ and $\mathbf{x}^r$ by a linear combination of a few atoms from their corresponding dictionary $\mathbf{D}^c = [\mathbf{d}_1^c, \mathbf{d}_2^c, \cdots, \mathbf{d}_J^c]$ and $\mathbf{D}^r = [\mathbf{d}_1^r, \mathbf{d}_2^r, \cdots, \mathbf{d}_J^r]$[1] respectively, which is learned via

$$\min \sum_{i=1}^{N} ||\mathbf{x}_i^c - \mathbf{D}^c \mathbf{y}_i^c||^2 + \gamma^c ||\mathbf{y}_i^c||_1 \quad (1)$$

and

$$\min \sum_{i=1}^{N} ||\mathbf{x}_i^r - \mathbf{D}^r \mathbf{y}_i^r||^2 + \gamma^r ||\mathbf{y}_i^r||_1 \quad (2)$$

where i is the frame index in our case and N is the total number of training examples. $\{\mathbf{y}_i^c\}$ and $\{\mathbf{y}_i^r\}$ are the sparse representation coefficients of $\mathbf{x}_i^c$ and $\mathbf{x}_i^r$, respectively. γ^c and γ^r denote a penalty weight on sparsity. $||\cdot||_1$ denotes the ℓ_1-norm, respectively.

SR learns representations of the input data that have only few components that are significantly non-zero, i.e. that are sparse. In a sparse representations, the dictionary $\mathbf{D}^c$ or $\mathbf{D}^r$ should be overcomplete, when the size of the dictionary is higher than the dimensionality of the input. The sparse overcomplete representations has been shown to be robust to noise and partial image occlusion [10].

2.2. Joint Dictionary Training

In order to capture the complex relationships between the clean and reverberant speech spaces and form a common representation across the two spaces, we need to learn the dictionaries $\mathbf{D}^c$ and $\mathbf{D}^r$ jointly, and then estimate the underlying $\mathbf{x}^c$ from $\mathbf{x}^r$. The flow chart is illustrated in Fig. 1.

Given the coupled training feature sequences $\{\mathbf{x}_i^c\}_{i=1}^N$ and $\{\mathbf{x}_i^r\}_{i=1}^N$, the problem of jointly learning the dictionaries can be formulated as follows:

$$\min \sum_{i=1}^{N} (||\mathbf{x}_i^c - \mathbf{D}^c \mathbf{y}_i||^2 + ||\mathbf{x}_i^r - \mathbf{D}^r \mathbf{y}_i||^2) + \gamma ||\mathbf{y}_i||_1. \quad (3)$$

Let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^c \\ \mathbf{x}^r \end{bmatrix}, \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}^c \\ \mathbf{D}^r \end{bmatrix}. \quad (4)$$

[1] $\{\mathbf{d}_j^c\}$ and $\{\mathbf{d}_j^r\}$ are representation basis sets of clean and reverberant speech, respectively.

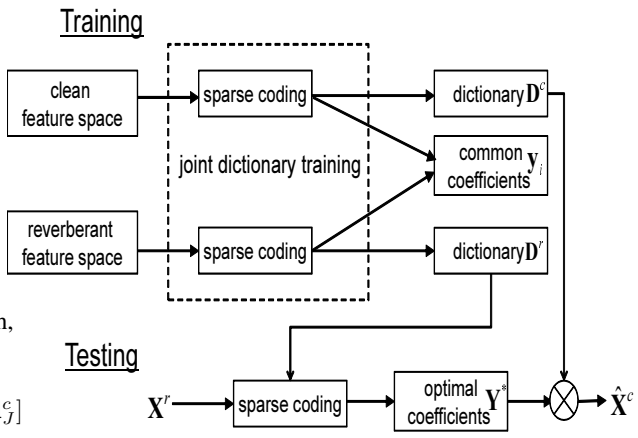

Fig. 1. Block diagram of the proposed joint sparse representation (JSR) method.

Then Eq. (3) reduces to a standard sparse code problem:

$$\min \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{D}\mathbf{y}_i||^2 + \gamma ||\mathbf{y}_i||_1. \quad (5)$$

Equation (5) is not convex in both $\mathbf{D}$ and $\mathbf{y}_i$, however it is convex in one of them with the other fixed[2].

2.3. Clean Feature Estimation

Once we obtain the coupled dictionaries $\mathbf{D}^c$ and $\mathbf{D}^r$, for any given testing noisy feature vector $\mathbf{x}_i^r$, we first find its sparse representation in terms of $\mathbf{D}^r$

$$\mathbf{y}_i^* = \arg\min_{\mathbf{y}_i} ||\mathbf{x}_i^r - \mathbf{D}^r \mathbf{y}_i||^2 + \gamma ||\mathbf{y}_i||_1, \quad (6)$$

and then estimate its corresponding clean feature vector $\mathbf{x}_i^c$ in terms of $\mathbf{D}^c$ via

$$\hat{\mathbf{x}}_i^c = \mathbf{D}^c \mathbf{y}_i^*. \quad (7)$$

In terms of a test utterance with a sequence of feature vectors $\mathbf{X}^r = \{\mathbf{x}_1^r, \mathbf{x}_2^r, \cdots, \mathbf{x}_M^r\}$ with M frames, the sparse representations of $\mathbf{X}^r$ can be found via

$$\mathbf{Y}^* = \arg\min_{\mathbf{Y}} ||\mathbf{X}^r - \mathbf{D}^r \mathbf{Y}||^2 + \gamma ||\mathbf{Y}||_1, \quad (8)$$

where $\mathbf{Y}^* = \{\mathbf{y}_1^*, \mathbf{y}_2^*, \cdots, \mathbf{y}_M^*\}$, and then the estimated clean feature vector sequence $\mathbf{X}^c$ is obtained by

$$\hat{\mathbf{X}}^c = \mathbf{D}^c \mathbf{Y}^*. \quad (9)$$

[2] In our experiments, we used a Matlab package developed in [12].

2.4. Related to other works

[5] and [6] obtained their robustness to reverberation in cep stral domain through mean substraction and variance com pensation, respectively, and they are mainly based on linea model. However, when the impulse response has a longe tail the distortions caused by reverberation are nonlinear i cepstral domain. Our method captures the non-linear rela tionships between the clean and reverberant feature space through joint sparse representations. Moreover, our metho does not need any explicit reverberation model.

3. EXPERIMENTAL RESULTS

3.1. Database

The proposed approach was evaluated on a realistic speec recognition task under reverberant environments. The train ing and testing datasets was taken from the CENSREC-4 data [13]. The clean training data, in which the total number o the utterance was 8,440 by 110 speakers (55 females and 5 males), was selected for training the acoustic model. Th testset D was recorded in real reverberant environments b 10 human speakers (five females and five males) using tw microphones (close-talking and distant-talking), in which th speech recorded by a distant microphone was selected for th evaluation. There were four reverberant environments (in-ca lounge, meeting room, and office) inside the testset D. In each environment, the total number of the utterances was 493.

The speech signal was sampled at 16 kHz and windowed with a 20-ms Hamming window every 10 ms (with a pre-emphasis $1 - 0.97z^{-1}$). A 24-channel mel-filter bank (MFB) analysis was applied, and finally the log MFB outputs were converted into 12 MFCCs through Discrete Cosine Transformation (DCT). The acoustic models consist of 18 phone models that have five states (three states for 'sp'). The baseline system was trained using 39-dimensional feature vectors consisting of 12-dimensional MFCC parameters and log-energy, along with their delta and delta-delta parameters.

3.2. Experimental Settings

As shown in Fig. 1, we need training and testing data set. Inside the testset D of CENREC-4 corpus, there was a separate adaptation data consisting of 110 utterances for each of the four environments. The speech was also recorded by two microphones (close-talking and distant-talking). In our experiments, for each environment we adopted this adaptation data for training the two coupled dictionaries on MFCC feature vectors using Eq. (5), and then estimated the clean MFCC sequence of each utterance inside testset D using Eq. (9). The sparse size and the penalty weight γ were empirically set to 1,024 and 0.4, respectively. Figure 2 shows the probability density functions (PDFs) of the first Mel-frequency cepstral coefficient of the close-talking speech, distant speech, and the estimated one using JSR for each reverberant environment. It can be found that compared with the distant speech, using JSR reduces the difference from the close speech, which will be helpful for the speech recognition system.

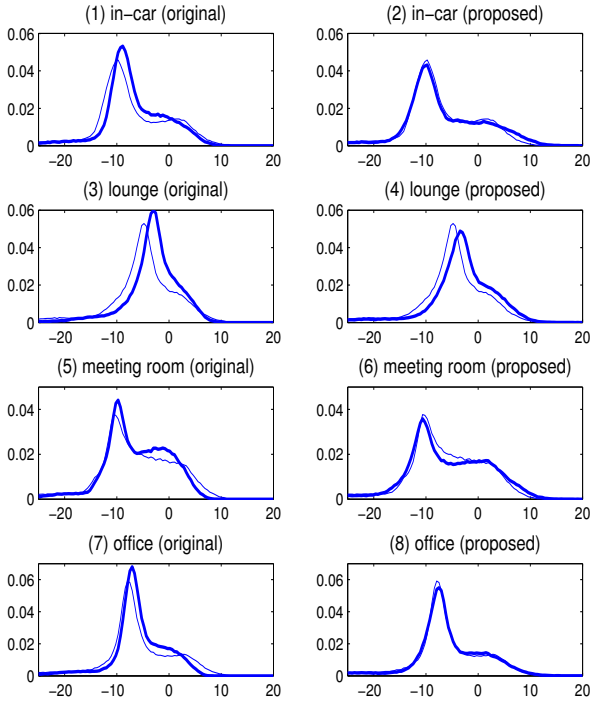

Fig. 2. Probability density functions (PDFs) of the first Mel-frequency cepstral coefficient of the close speech, distant speech, and the estimated one using JSR for each reverberant environment. Inside each sub-figure, thin line is for the close-talking speech and bold line is for the distant speech or the estimated one.

For comparisons, a *parametric formulation of the generalized spectral subtraction* (GSS) [14] were applied and we retrained the acoustic model after this processing. Cepstral Mean Normalization (CMN) [5], one traditional dereverberation technique adopted in many current systems, was also performed for comparison.

3.3. Results

Table 1 shows the recognition results obtained from the different methods. "CT" denotes the recognition performance of the speech recorded by close-talking microphone. The recognition performance of "baseline" (distant speech) depend on the reverberant environments. When the recording environment is seriously reverberant (e.g., in the case of "lounge"), the recognition accuracy can degrade into less than 50%. Due

Table 1. Word accuracies (as percentages) for different methods. "meeting" denotes the meeting room environment.

	in-car	lounge	meeting	office	Average
baseline	76.27	43.83	89.12	85.18	73.60
CT	97.16	96.45	97.24	98.07	97.23
GSS	86.32	76.67	85.65	83.29	82.98
CMN	84.86	59.76	90.66	93.50	82.19
GSS+CMN	88.88	83.07	89.32	92.52	88.45
JSR	84.71	67.32	92.55	95.23	**84.95**
JSR+CMN	87.62	79.74	90.19	93.30	**87.71**

to its capacity of reducing the channel distortion, Cepstral Mean Normalization (CMN) is helpful for improving the speech recognition performance. "GSS" performs significantly improve the ASR performance in lounge environment, which demonstrates the effectiveness of spectral subtraction for suppressing the late reverberations. "GSS" performs not as well as "CMN" in meeting room and office environments, where there mainly exit the early reverberations. When the joint sparse representation (JSR) is employed, the average recognition performance is better than "GSS" and "CMN", which demonstrates the effectiveness of the proposed method. Moreover, with a subsequent CMN post-processing, the recognition performance is further improved in relatively serious environments (i.e., "in-car" and "lounge"), and the average recognition accuracy is achieved with a relative improvement of 59.1% compared with "baseline".

4. CONCLUSIONS

In this paper, we have proposed a joint sparse representation (JSR) technique for cepstral-domain dereverberation for realistic distant-talking speech recognition. In our proposed JSR, a joint dictionary learning is performed across the clean and reverberant cepstral features in order to capture possible complex relationships between the two feature spaces. Given a sequence of reverberant feature vectors (Mel-frequency cepstral coefficients (MFCCs)), their corresponding cepstral vectors of clean speech are estimated through the use of the dictionary of clean speech. Compared with the spectral subtraction and the Cepstral Mean Normalization (CMN), the proposed method shows its superiority in terms of a significant improvement in recognition performance in the speech recognition experiments conducted in four realistic reverberant environments.

In our experiments, the reverberant environments are assumed to be known. In order to carry out data-driven recognition, we need to develop an effective algorithm for the unseen reverberant environments. Moreover, when the system encounters a new reverberant environment, automatic adapting the sparse representations to different reverberant environments is desirable, and this should be one of future works.

5. REFERENCES

[1] B. Yegnanarayana and P.S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267 –281, 2000.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717 –1731, 2010.

[3] Qin Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2023 –2032, 2007.

[4] Mingyang Wu and DeLiang Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774 – 784, 2006.

[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254 – 272, 1981.

[6] M. Delcroix, T. Nakatani, and S. Watanabe, "Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 324 –334, 2009.

[7] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1676 – 1691, 2010.

[8] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031 – 1044, 2010.

[9] E.J. Candes and M.B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21 –30, 2008.

[10] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 –227, 2009.

[11] T.N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4370 –4373.

[12] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2007, pp. 801–808.

[13] T Nishiura, R Gruhn, and S Nakamura, "Evaluation framework for distant-talking speech recognition under reverberant environments," in *Interspeech-2008*, 2008.

[14] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328 –337, 1998.