MULTICHANNEL FEATURE ENHANCEMENT IN DISTRIBUTED MICROPHONE ARRAYS FOR ROBUST DISTANT SPEECH RECOGNITION IN SMART ROOMS

Seyedmahdad Mirsamadi and John H.L. Hansen[⋆]

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, U.S.A.
website: http://crss.utdallas.edu

ABSTRACT

Room reverberation and environmental noise present challenges for integration of speech recognition technology in smart room applications. We present a multichannel enhancement framework for distributed microphone arrays to mitigate the effects of both additive noise and reverberation on distant-talking microphones. The proposed approach uses techniques of nonnegative matrix and tensor factorization to achieve both noise suppression (through sparse representation of speech spectra) and dereverberation (through decomposition of magnitude spectra into convolutive components). Results of ASR experiments on the DIRHA-GRID corpus confirm that the proposed approach can achieve relative improvements of up to +20% in recognition accuracy in highly reverberant and noisy conditions using clean-trained models.

Index Terms— distant speech recognition, distributed microphone array, nonnegative matrix/tensor factorization

1. INTRODUCTION

The emerging interest in achieving smart interactive homes or office environments with automated appliances/technology has led to a growing interest in replacing the current modes of interaction (primarily touch screens or keyboards) with human speech. The ability to operate home technology systems by issuing voice commands is not only of critical importance to many elderly or disabled individuals, but also provides a degree of convenience which is desired by the general user. The challenge with employing audio and speech technology in the context of smart rooms is that such technologies require the use of close-talking (headset or lapel) microphones, which is in many cases not possible or too limiting. It is often desired to let an untethered user freely issue voice commands from any random location in the house, with one or more microphones capturing his/her voice from different locations in the smart room space.

Many current Automatic Speech Recognition (ASR) systems fail in such distant-talking scenarios, hindered by the corrupting effects of room reverberation and environmental noise. While these problems have conventionally been alleviated by the directional response of uniform arrays with closely-spaced microphones (i.e., by employing beamforming techniques), these solutions are limited to rooms that are equipped with such uniform arrays at fixed pre-determined positions (e.g., on the walls or ceiling). In addition, speaker location information is often not available and very difficult to estimate in reverberant rooms, further limiting the use of beamforming on conventional arrays as a general solution.

An attractive aim in automated home and office environments is the use of distributed microphone arrays for distant speech capture. In a distributed array, the microphones are far apart in random unknown locations, and there might be no synchrony among them due to the use of independent recording devices [1]. Such a distributed array is formed, for example, by the collection of microphones in smart phones or laptops within a room. In this flexible framework, using conventional time-delay based methods is not possible due to the lack of synchrony among channels, non-uniform array geometry, and the possibility of different signal-to-noise ratios and direct-to-reverberation ratios among the different channels. We should thus resort to feature enhancement techniques that are independent of signal phases and fuse the information of different channels at the magnitude spectrum level.

Feature enhancement techniques based on sparse representations of speech signals have recently accquired considerable popularity in ASR research [2, 3]. These approaches assume that the information contained in a speech spectrogram can be represented as a linear combination of a finite number of elementary bases, often referred to as *atoms* or *exemplars*. A collection of such atoms, called a *dictionary*, is created using clean training data. For feature enhancement, the closest representation of the noisy spectrum in terms of these clean exemplars is obtained and replaced by the original noisy spectrum (using techniques from the Nonnegative Matrix Factorization (NMF) field [4]). Such sparse representation (SR) approaches were originally used for speech sepa-

⋆ This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

ration [5], and later employed for single-channel noise-robust ASR through the use of fixed pre-trained speech and noise exemplars [2]. Although it has been shown in [2] that the noise suppression provided by the SR-based approach yields significant improvements in recognition accuracy in noisy environments, this approach is less effective in mitigating reverberation effects. While the SR-based approach achieves enhancement by weighting each time-frequency bin separately by a gain function (details in Sec. 3.2), the reverberation problem is a long-term effect involving multiple time frames [6] and thus cannot be handled by the SR frame-work.

In this study, we propose a multichannel feature enhancement framework based on the combined use of a generalized SR-based denoising approach and a previously proposed multichannel dereverberation algorithm based on Convolutive Nonnegative Tensor Factorization (CNTF) [7]. Unlike [2], we do not make use of a fixed pre-trained noise dictionary. We will instead use silence regions between utterances to build smaller local noise dictionaries, and show that this can provide a reasonable degree of noise suppression. This will in turn enable the system to effectively deal with previously unseen types of noise. Moreover, we propose to use a more general parametric gain function in SR-based denoising, which is shown to improve the recognition accuracy in low SNRs.

The benefits of using CNTF together with SR-based enhancement is two-fold. First, it compensates for the inadequacy of the SR algorithm to address reverberation effects. Moreover, it provides an efficient way of combining information from different channels of a distributed array, providing a multichannel framework in which SR-based denoising can be employed individually on each channel before dereverberation. In addition, the CNTF algorithm has been shown to be robust against unequal levels of direct-to-reverberation ratio (DRR) among different channels, blindly increasing the contribution of higher DRR channels to the final output [7]. This would be particularly helpful in a smart room application where it is not known in advance which microphone is closer to the source.

The remainder of this paper is organized as follows. In Sec. 2, we describe the signal model for distant speech in the magnitude spectrum domain. In Sec. 3, we present the proposed multichannel enhancement approach for mitigating the effects of noise and reverberation. In Sec. 4, we provide results from ASR experiments performed to evaluate the proposed method, and in Sec. 5 we provide a summary and final conclusions.

2. TIME-FREQUENCY MODEL FOR DISTANT SPEECH

A speech signal captured by a distant-talking microphone is modeled by a convolution between the Room Impulse Response (RIR) and clean speech, on which an additive noise term is also imposed. For the purpose of feature enhancement in ASR, it is common to use the following approximate model for the magnitude spectrum of the received signal [6, 8],

$$X^{(i)}(m,k) = \sum_{p=0}^{L_H-1} H_k^{(i)}(p) S(m-p,k) + B^{(i)}(m,k), \quad (1)$$

where $S(m,k)$, $B^{(i)}(m,k)$ and $X^{(i)}(m,k)$ are the magnitude short-time Fourier transforms (STFTs) of the clean speech signal, the additive noise from the i'th microphone, and the received signal of the i'th microphone, respectively. Here, m and k are frame and frequency indices, and $H_k^{(i)}(m)$ represents the subband envelope of the RIR from the source location to the i'th microphone location. The superscript (i) will be used to denote channel index throughout this paper. The first term in Eq. 1 represents the reverberant signal and will be denoted by $Y^{(i)}(m,k)$. Arranging the STFT components in the spectrogram matrices will result in,

$$\mathbf{X}^{(i)} = \sum_{p=0}^{L_H-1} \mathbf{H}^{(i)}(p) \overset{p\rightarrow}{\mathbf{S}} + \mathbf{B}^{(i)}. \quad (2)$$

where $\mathbf{S}$, $\mathbf{B}^{(i)}$ and $\mathbf{X}^{(i)}$ are $K \times M$ spectrogram matrices of the correponding signals (K is the number of frequency bins, M is the number of time frames), and $\mathbf{H}^{(i)}(p)$ is a diagonal matrix of the form $\mathbf{H}^{(i)}(p) = \text{diag}([H_0^{(i)}(p), \ldots, H_{K-1}^{(i)}(p)])$. The operator $p \rightarrow$ shifts the rows of its argument matrix by p positions to the right, filling in zeros from the left. The matrices $\mathbf{X}^{(i)}$, $\mathbf{B}^{(i)}$ and $\mathbf{H}^{(i)}(p)$ for different channels ($i = 1, \ldots, Q$) can be considered as the frontal slices of the corresponding third order tensors $\mathcal{X}$, $\mathcal{B}$, and $\mathcal{H}(p)$ [7].

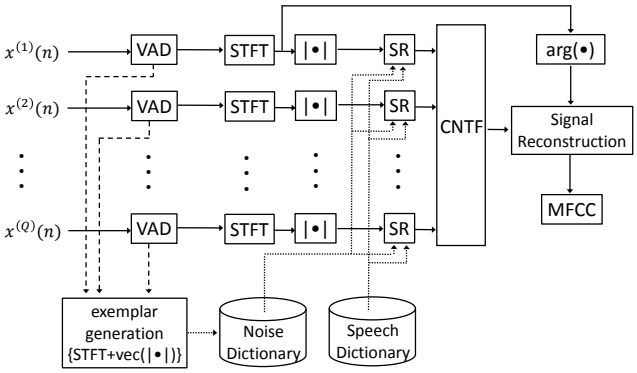

Fig. 1. The proposed front-end for noise and reverberation robust ASR. The dashed ouputs from VAD units represent non-speech (noise-only) intervals.

3. MULTICHANNEL SPECTRUM ENHANCEMENT

3.1. Overview

Fig. 1 shows the overall proposed system for suppressing noise and reverberation in distant-talking speech. A Voice

Activity Detection (VAD) unit initially detects whether the input signal is speech or pure noise corresponding to a silence period. If silence is detected, the magnitude STFTs of the noise are computed and used to update a noise dictionary which will be employed to suppress noise in the forthcoming speech utterance. If speech is detected, sparse representation based denoising is performed on the magnitude STFTs, using the current noise dictionary as well as a speech dictionary which is created in advance using the training database (further details on dictionary creation in the next section).

The denoised spectra from all channels are then jointly processed by CNTF dereverberation algorithm to produce an estimate of the clean speech spectrum. The phase values from one of the microphones is used together with the estimated clean spectrogram to produce an enhanced time domain signal which is finally submitted to a conventional Mel Frequency Cepstral Coefficient (MFCC) extraction unit.

3.2. Environmental Noise Suppression based on Sparse Representaion (SR)

It has been shown that a speech spectrogram can be represented as a nonnegative linear combination of a few basic *example* spectrograms, and that this linear combination is often very sparse [9]. To obtain representative vectors that can be decomposed into such linear combinations, we divide the magnitude spectrogram matrix into overlapping windows of length T and stack the spectrogram portion under the window into a column vector of length $K \cdot T$. We can create a clean speech dictionary by sliding this window on the spectrogram matrices of the training data and arranging the resulting exemplars ($\mathbf{a}_j^s$) as the columns of a matrix. A noise dictionary can also be constructed in a similar fashion using the noise samples collected by the microphones during silence periods.

$$\mathbf{A}_s = [\mathbf{a}_1^s, \ldots, \mathbf{a}_{N_s}^s]. \tag{3}$$

$$\mathbf{A}_n = [\mathbf{a}_1^n, \ldots, \mathbf{a}_{N_n}^n]. \tag{4}$$

Here, N_s and N_n are the number of speech and noise exemplars, respectively.

In order to enhance a received speech utterance, we use the same windowing and column-stacking procedure described above to obtain representative vectors from the magnitude spectrograms of the received signals, and express the resulting vectors as a linear combination of the speech and noise exemplars (here we drop the channel index (i) for simplicity of notation):

$$\tilde{\mathbf{x}}_l = \text{vec}([\mathbf{x}_{l\Delta}, \ldots, \mathbf{x}_{l\Delta+T-1}]), \tag{5}$$

$$\tilde{\mathbf{x}}_l = \sum_{j=1}^{N_s} v_{j,l}^s \mathbf{a}_j^s + \sum_{j=1}^{N_n} v_{j,l}^n \mathbf{a}_j^n = \mathbf{A}\mathbf{v}_l. \tag{6}$$

In the above equations, $\mathbf{x}_m$ is the magnitude DFT of the m'th frame of the received signal, vec(.) is the column stacking operator, Δ is the skip rate used in windowing the spectrogram, $v_{j,l}^s$ and $v_{j,l}^n$ are the non-negative weights (*activations*) in the linear combination, $\mathbf{v}_l = [v_{1,l}^s, \ldots, v_{N_s,l}^s, v_{1,l}^n, \ldots, v_{N_n,l}^n]^T$, and $\mathbf{A} = [\mathbf{A}_s, \mathbf{A}_n]$.

Based on Eq. (6), the collection of all vectors resulting from sliding the window on the spectrogram of i'th microphone signal can be written as,

$$\tilde{\mathbf{X}}^{(i)} = \mathbf{A}\mathbf{V}^{(i)}, \tag{7}$$

where $\tilde{\mathbf{X}}^{(i)} = [\tilde{\mathbf{x}}_0^{(i)}, \ldots, \tilde{\mathbf{x}}_{L-1}^{(i)}]$, $\mathbf{V}^{(i)} = [\mathbf{v}_0^{(i)}, \ldots, \tilde{\mathbf{v}}_{L-1}^{(i)}]$, and L is the total number of spectrogram windows within the utterance.

The decomposition in Eq. (7) can be achieved by means of performing nonnegative matrix factorization on $\tilde{\mathbf{X}}^{(i)}$, keeping the base matrix $\mathbf{A}$ fixed and only updating $\mathbf{V}^{(i)}$ using the multiplicative update rules introduced in [10]:

$$\mathbf{V}^{(i)} \leftarrow \mathbf{V}^{(i)} . * \frac{\mathbf{A}^T \frac{\mathbf{X}^{(i)}}{\mathbf{A}\mathbf{V}^{(i)}}}{\mathbf{A}^T \mathbf{1}}. \tag{8}$$

In Eq. (8), '$.*$' represents elementwise multiplication and the matrix divisions are elementwise as well, and $\mathbf{1}$ is a vector of all ones (of length $N_s + N_n$).

Following the NMF iterations of Eq. (8), the resulting activation values can be used to separate the speech and noise contributions:

$$\tilde{\mathbf{Y}}_s^{(i)} = \mathbf{A}_s \mathbf{V}_s^{(i)} \tag{9}$$

$$\tilde{\mathbf{Y}}_n^{(i)} = \mathbf{A}_n \mathbf{V}_n^{(i)} \tag{10}$$

In the above equations, $\tilde{\mathbf{Y}}_s^{(i)}$ and $\tilde{\mathbf{Y}}_n^{(i)}$ are the speech and noise components in $\tilde{\mathbf{X}}^{(i)}$, and $\mathbf{V}_s^{(i)}$ and $\mathbf{V}_n^{(i)}$ represent the upper N_s rows (speech activations) and the lower N_n rows (noise activations) of $\mathbf{V}^{(i)}$, respectively.

Denoting the columns of $\tilde{\mathbf{Y}}_s^{(i)}$ by $\tilde{\mathbf{y}}_{s,l}^{(i)}$, $(l = 0, \ldots, L-1)$, an estimate of the denoised spectrogram segments can be obtained by

$$[\mathbf{y}_{s,l\Delta}^{(i)}, \ldots, \mathbf{y}_{s,l\Delta+T-1}^{(i)}] = \text{unvec}(\tilde{\mathbf{y}}_{s,l}^{(i)}), \tag{11}$$

where the unvec(.) operator reshapes a vector of length $K \cdot T$ into a $K \times T$ matrix. The overall speech spectrogram matrix $\mathbf{Y}_s^{(i)}$ can be constructed by overlap-adding the segments obtained from Eq. (11). Additionally, we can obtain an estimate for the spectrogram of the additive noise on channel i in a similar fashion to (11), but using the columns of $\tilde{\mathbf{Y}}_n^{(i)}$ instead.

Although it is possible to directly use the above estimated $Y_s^{(i)}(m,k)$ as the denoised spectral components, we have found that doing so results in poor recognition performance, stemming from the residual speech energy not expressed by the linear combination of the exemplars (similar results have been reported in [2, 11]). We thus use the following parametric gain function instead,

$$G^{(i)}(m,k) = \left(\frac{\xi^{(i)}(m,k)}{\alpha + \xi^{(i)}(m,k)} \right)^{\beta}, \tag{12}$$

where,

$$\xi^{(i)}(m,k) = \frac{Y_s^{(i)}(m,k)}{Y_n^{(i)}(m,k)}, \quad (13)$$

can be interpreted as a measure of the *a priori* SNR in the spectral amplitude estimation [12]. Note that by selecting $\alpha = \beta = 1$, the Wiener amplitude estimator used in [2] will be obtained. However, it is known that in low SNRs the optimal gain function given by Minimum Mean-Square Error (MMSE) estimation of the spectral amplitudes deviates from the Wiener solution [12], and thus the added degrees of freedom provided by α and β will lead to more precise amplitude estimates. Similar improvements provided by such generalized gain functions have been reported in other enhancement techniques as well (e.g. in [13]). Using the gain function of Eq. (12) and the original magnitude STFT values $X^{(i)}(m,k)$, the denoised spectral amplitudes of the i'th channel can be obtained by the following spectral weighting:

$$Y_d^{(i)}(m,k) = G^{(i)}(m,k)X^{(i)}(m,k) \quad (14)$$

3.3. Multichannel Dereverberation based on Convolutive Nonnegative Tensor Factorization (CNTF)

Based on the model of Eq. (1) and the estimated reverberant spectral amplitudes from Sec. 3.2, the dereverberation task can be stated as finding the nonnegative factors $H_k^{(i)}(m)$ (for all i) and $S(m,k)$ (common among all channels) which minimize the total error defined by,

$$E = \sum_{i,m,k} \left(Y_d^{(i)}(m,k) - \sum_{p=0}^{L_H-1} H_k^{(i)}(p) S(m-p,k) \right)^2. \quad (15)$$

We have shown in [7] that by considering the channel spectrogram matrices as the frontal slices of a third order tensor $\mathcal{X}$, the factorization provided by Eq. (15) can be interpreted as a special case of convolutive nonnegative tensor factorization, in which the base matrices are constrained to be diagonal. It was shown that the nonnegative factors can be obtianed by iterating the following multiplicative update rules:

$$\hat{H}_k^{(i)}(p) \leftarrow \hat{H}_k^{(i)}(p) \frac{\sum_m Y_d^{(i)}(m,k)\hat{S}(m-p,k)}{\sum_m Z^{(i)}(m,k)\hat{S}(m-p,k)}, \quad (16)$$

$$\hat{S}(l,k) \leftarrow \hat{S}(l,k) \frac{\sum_i \sum_m Y_d^{(i)}(m,k)\hat{H}_k^{(i)}(m-l)}{\sum_i \sum_m Z^{(i)}(m,k)\hat{H}_k^{(i)}(m-l)}, \quad (17)$$

where $Z^{(i)}(m,k)$ is the estimate of the reverberant spectral amplitudes based on the current values of the factors. Note that in the case of a single microphone ($Q = 1$), the CNTF algorithm simplifies to the single-channel convolutive nonnegative matrix factorization (CNMF) algorithm of [14].

Table 1. Average T_{60}s and SNRs for the selected channels*.

Microphone (location/name)	Room T_{60}	Avg. T_{60} of cross-room RIRs	Avg. SNR of utterances
LivingRoom/LA2	0.77 s	1.23 s	13.8 dB
Kitchen/KA2	0.82 s	1.40 s	11.6 dB
Bedroom/B1L	0.62 s	1.66 s	9.8 dB
Corridor/C1L	0.70 s	1.05 s	11.9 dB

* Values based on development dataset (for which this information is provided). The test dataset is expected to have similar values to those reported here.

4. EXPERIMENTAL RESULTS

4.1. Evaluation Database

We evaluate the performance of the proposed method on the DIRHA-GRID corpus [15]. The DIRHA-GRID corpus contains multichannel 16 kHz recordings of different acoustic scenes collected in an apartment with multiple rooms (living room, kitchen, bedroom, corridor). Each acoustic scene has a duration of 1 minute and includes a variable number of speech utterances (short commands derived from the GRID corpus [16]), as well as different non-speech sources representing typical noises in domestic environments such as radio, TV, knocking, ringing, vacuuming, etc. Multiple microphones are attached to the walls or ceiling of each room to capture the commands which are issued randomly from different rooms. This creates a realistic multiroom scenario in which speech and noise instances can occur in one room and captured by microphones both in the same and other rooms. As a result, many of the RIRs we are dealing with in this corpus are cross-room RIRs with very high reverberation times (T_{60}) exceeding 1 second. These large reverberation times together with the variety of noise sources in each scene make for a very challenging ASR task. In our experiments, we use only one microphone from each room, limiting ourself to a distributed array scenario in which no closely-spaced microphones are available. Table 1 lists the microphones used and the corresponding average reverberation times and SNRs. The corpus contains a training set (containing 17,000 clean utterances from 34 different speakers in the GRID corpus), as well as a development dataset and two test sets. The ASR experiments reported here are all performed on test set 1. (Further details on the DIRHA-GRID corpus are provided in [15]).

4.2. ASR experiments

We used HTK toolkit to conduct ASR experiments on the DIRHA-GRID corpus. We used 13-dimensional MFCCs along with their delta and double-delta coefficients as speech features. Cepstral Mean Normalization (CMN) was used in

Table 2. word error rates in ASR experiments*.

Acoustic Models	No. of channels	Channel(s)	No Enhancement	SR, Wiener gain func. ($\alpha = \beta = 1$)	SR, General gain func. ($\alpha = 2$, $\beta = 1.2$)	SR, General gain func. ($\alpha = 2$, $\beta = 1.2$) +CNMF/CNTF
Clean-trained	1	LA2	80.6	77.1	74.7	69.3
	1	KA2	81.7	77.2	75.4	71.9
	2	LA2+KA2	80.1	76.1	73.8	67.5
	4	LA2+KA2+B1L+R1C	78.1	75.2	73.0	62.7
MLLR-adapted	1	LA2	70.3	67.9	64.6	53.7
	1	KA2	73.6	71.9	70.2	60.2
	2	LA2+KA2	67.4	64.2	60.6	49.2
	4	LA2+KA2+B1L+R1C	68.1	65.1	60.5	42.6

* In multichannel experiments, including CNTF is necessary to combine information from different channels and obtain a single set of features. Those WERs belonging to multichannel experiments without CNTF have been obtained by simply selecting the channel with the highest SNR, according to Eq. (13).

all experiments. Using the clean utterances in the training set, left-to-right HMMs with 12 mixtures per state and a variable number of states (ranging from 3 to 10, based on the number of syllables) were trained for each of the 51 words in the speech commands and used as speaker-independent acoustic models. The front-end processing scheme illustrated in Fig. 1 was used to obtain features for speech recognition. A basic GMM-based voice activity detector trained on the development set was used to obtain VAD information for all channels, and further refined manually to minimize the effects of VAD errors on recognition accuracy. The STFT analysis for each channel uses a Hamming window of length 64 ms, a skip rate of 16 ms, and DFT length of 1024 samples. For each 1 minute recording in the test set, the channel spectrograms are first enhanced by the SR-based denoising algorithm, using 4000 speech exemplars created from the training set and a variable number of noise exemplars created using the silence regions in the recording. The noise dictionary for each 1-minute scene is created solely based on the silence regions within the same scene. The sliding window used for windowing the spectrograms is of length $T = 20$ and uses a skip rate of $\Delta = 10$ frames (both in dictionary creation and the enhancement of noisy spectra). The activation values in $\mathbf{V}^{(i)}$ were all initialized by 1, followed by 20 NMF iterations (Eq. 8) to obtain the sparse representations. In the CNTF algorithm, a filter length of $L_H = 16$ was used for all subband filters $H_k^{(i)}(m)$, which were initialized with $H_k^{(i)}(m) = 1 - m/2L_H, (m = 0, \ldots, L_H - 1)$. 20 iterations of the update rules (16) and (17) were used to obtain an estimate of the clean speech spectrogram. The estimated amplitudes were then used together with the phase values from channel 1 (living room channel) to reconstruct an enhanced time domain signal, which is finally submitted to a conventional MFCC extraction unit with a frame length of 25 ms and a frame skip rate of 10 ms.

Table 2 shows the word error rates (WERs) obtained from ASR experiment on test set 1 from the DIRHA-GRID corpus. We report recognition results both by using clean-trained models as well as models that are MLLR-adapted towards the development set. To facilitate separate analysis of the contribution from each sub-system, we also report WERs obtained by applying SR-based denoising only (skipping dereverberation). Moreover, we use the SR algorithm both with the Wiener gain function (i.e., with $\alpha = \beta = 1$), as well as the general gain function of Eq. (12) (values of $\alpha = 2$ and $\beta = 1.2$ were selected based on experiments performed on the development dataset). It is observed from the table that the proposed enhancement method achieves considerable improvement in both single-channel and multichannel scenarios. The relative WER improvements for 1-channel, 2-channel and 4-channel cases are +13.0%, +15.7% and +19.7% using clean-trained models and +20.9%, +27% and +37.4% using adapted models. Note that a considerable improvement is provided by CNTF in the 4-channel scenario, because in this case there is a microphone present in all of the 4 rooms (In single-channel and dual-channel experiments, speech events may happen in rooms where there is no microphone used, hence relying only on cross-room recordings).

5. CONCLUSIONS

In this study, we presented a front-end processing strategy for noise and reverberation robust distant speech recognition in distributed microphone arrays. The proposed approach is independent of speaker location and adapts to varying levels of SNR and DRR among different channels. Using clean-trained models, relative WER improvements of +13-19.7% are achieved, with improvements increasing to +20.9-37.4% when models are MLLR adapted. The proposed enhancement strategy effectively addresses reverberation and noise in diverse smart room scenarios.

6. REFERENCES

[1] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09.*, Oct 2009, pp. 161–164.

[2] J.F. Gemmeke, T. Virtanen, and A Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, Sept 2011.

[3] T.N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From timit to lvcsr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2598–2613, Nov 2011.

[4] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shunichi Amari, *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov 2012.

[7] S. Mirsamadi and J.H.L. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Interspeech*, 2014.

[8] Y. Avargel and I Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[9] J.F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272–287, April 2010.

[10] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[11] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Interspeech*, 2010, pp. 717–720.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[13] S.O. Sadjadi and J.H.L. Hansen, "Blind reverberation mitigation for robust speaker identification," in *ICASSP*, 2012, pp. 4225–4228.

[14] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *ICASSP*, 2009, pp. 45–48.

[15] M. Matassoni, R.F. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Interspeech*, 2014.

[16] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.