SPEECH ENHANCEMENT USING BEAMFORMING AND NON NEGATIVE MATRIX FACTORIZATION FOR ROBUST SPEECH RECOGNITION IN THE CHIME-3 CHALLENGE

Thanh T. Vu, Benjamin Bigot, Eng Siong Chng

Rolls-Royce@NTU Corporate Lab, Nanyang Technological University, Singapore

ABSTRACT

In this paper we present our contribution to the third CHiME challenge on speech separation and recognition for noisy multi-channel recordings. The use-case of the challenge consists in single speaker utterances recorded in highly non-stationary noisy environments using a 6-microphone array mounted on a tablet computer. The front-end of our system is performing speech enhancement by cascading a cross-correlation-based channel selection, Signal Dependent MVDR beamforming and online source separation based on sparse NMF. The back-end module is a state-of-the-art speech recognition system with DNN acoustic models trained on fMLLR features and a RNN Language Model. Our system reaches an overall WER of 11.94% on real test recordings, achieving a relative improvement of 65% compared to the baseline system.

Index Terms— Speech Enhancement, Automatic Speech Recognition, MVDR Beamforming, Non Negative Matrix Factorization,CHiME challenge

1. INTRODUCTION

Automatic Speech Recognition (ASR) on far-field audio recordings remains challenging. Capturing speech with distant microphones increases significantly the contribution of environmental noise and reverberation and reduces performance of ASR. For several years, research challenges like CHiME [1, 2] and REVERB [3] have been providing challenging use-cases to an increasing number of participants. These campaigns have already successfully supported the development of novel methods with high capability in limiting the impact of noise and reverberation on ASR.

The third CHiME[1] challenge proposes to perform speech recognition on multi-channel speech recordings captured in various highly non-stationary noisy environments. Utterances of the Wall Street Journal corpus [4] have been read by a set of 12 speakers and recorded using a 6-microphone array mounted on a tablet computer. As shown in Fig.1, five microphones are turned towards the speaker and one microphone (top middle channel 2, noted backside channel) is on the back panel of the tablet. Users are free to hold the tablet as they want and to move around with the device while recording in one of the 4 noisy environments: bus, cafeteria, street, pedestrian area. Thus, one or several microphones may be partially or totally obstructed during the recording by user's fingers or by the support on which the device has been disposed.

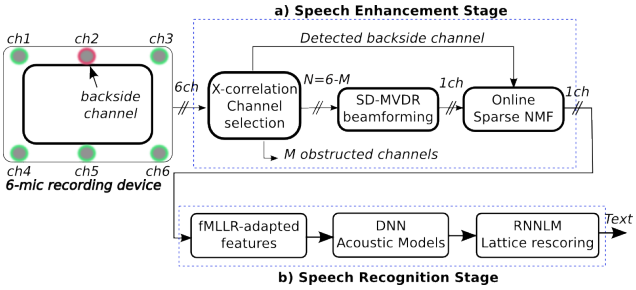

Fig. 1. Our framework for multi-channel robust ASR

Our contribution to CHiME is a flexible framework combining multi-channel speech enhancement with automatic speech recognition in order to handle recording artefacts and microphone failures. The system depicted Fig.1 is composed of a front-end Speech Enhancement (SE) module and a state-of-the-art back-end ASR. Our SE stage pre-process a 6-channel noisy recording into an enhanced single signal through 3 steps: channel selection based on cross-correlation, Minimum Variance Distortionless Response (MVDR) beamforming and SE source separation based on Non negative Matrix Factorization. The channel selection module automatically detects M obstructed microphones as well as the backside channel. At this step the M obstructed channels are discarded because the next module (MVDR beamformer) is acknowledged to be very sensitive to channel failures [5]. The detected backside channel will be used in a subsequent step by the NMF-based component. The N selected channels are processed in a second time by a MVDR beamformer in order to suppress additive noise and combine N channels into an enhanced single-channel signal. During the last step of the SE front-end, a sparse NMF-based source separation is used to remove residual noise from the output of the MVDR beamformer. The signal detected as the backside microphone is used by the NMF-based source separation module to adapt noise models to current test recordings. The enhanced speech

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/

signal produced by the SE front-end is then decoded by a state-of-the-art ASR including Deep Neural Network (DNN) Acoustic Models (AM) trained on speaker-adapted with feature space maximum likelihood linear regression (fMLLR) acoustic features. The word lattices produced by the speech decoder in an ultimate process are re-scored using a Recurrent Neural Network Language Model (RNN-LM).

To our knowledge, this work is one of the few studies investigating NMF, traditionally used for single channel speech enhancement [6, 7, 8, 9, 10], in a task of speech recognition using a DNN-based ASR. We have integrated several SE technologies into a flexible framework able to handle efficiently corrupted input signals. Our best system achieves an overall WER of 11.94% on real noisy recordings for an overall relative WER improvement of 65% compared to the baseline of the challenge. We also provide a large set of experiments and discuss the benefit of the components of our system.

The remainder of this paper consists as follows. In Section 2 we describe briefly the CHiME baseline system. Our system is described and discussed in Section 3. In section 4 we present the corpus used for the experiments and the evaluations reported in Section 5, where we also discuss our proposal through a large set of contrastive experiments.

2. CHIME-3 BASELINE SYSTEM

CHiME's organizers have provided the baseline speech enhancement ASR system presented in Fig.2. This system is composed of a front-end SE based-on channel selection and MVDR beamformer. The ASR back-end is relying on DNN acoustic models trained on Mel Filterbank features. The speech enhancement baseline aims to transform the multi-channel noisy input signal into a single-channel enhanced output signal suitable for ASR processing.

MVDR beamforming is acknowledged to be very sensitive to channel failure, thus a first step consists in detecting the corrupted channels. The baseline channel selection module discards channels based on the computation of the energy captured by each microphone. M channels are removed from the set of 6 microphones if their energies are found lower than threshold. The N selected channels are then processed by a Time-Varying Minimum Variance Distortionless Response (TV-MVDR) beamformer with diagonal loading [11]. This method requires the computation of the noise covariance matrix and the estimation of the Time Difference Of Arrival (TDOA) of multichannel signals. To this purpose, in the baseline the spatial position of the target speaker in each time frame of speech is encoded by a non-linear SRP-PHAT pseudo-spectrum found to perform best among a variety of source localization techniques [12].

CHiME is providing two back-end ASR systems. A first system uses Gaussian Mixture Models acoustic models trained on MFCC acoustic features with LDA dimension reduction, Maximum likelihood linear transformation (MLLT),

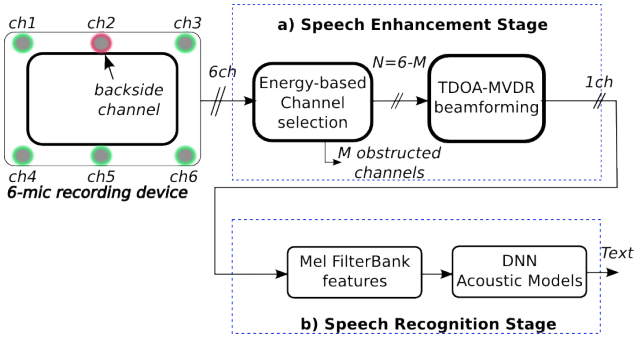

Fig. 2. Baseline framework for multi-channel robust ASR

and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT). A second ASR system is a DNN baseline providing state-of-the-art ASR performance. The DNN has 7 layers with 2048 neurons per hidden layer. The input layer has 5 frames of left and right context. The DNN training procedure consists in a pre-training using restricted Boltzmann machines, cross entropy training, and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion. We will compare our work to the baseline DNN-based ASR. In the next section we present our contribution to CHiME.

3. SYSTEM DESCRIPTION

3.1. Speech Enhancement Front-End

3.1.1. Cross-correlation-based channel Selection

As mentioned previously, during the recording of training and evaluation data, one or several microphones may have been obstructed by the user. Since MVDR beamforming performance suffers from such failure we have proposed an improved channel selection procedure. We assume all the microphones on the front of the recording device capture a larger contribution of speaker's speech and the microphone turned backwards captures mostly the contribution of background noise. Therefore we propose to use cross-correlation over the set of signals captured during one recording session in order to identify the obstructed microphones as well as the the backside channel. As presented on Algorithm 1, we first compute the average cross-correlation between the 6 available audio signals . The signals with a cross-correlation value lower than a threshold λ_1 are considered obstructed and are removed from the set of channels. Among the remaining signals, the channel with the lowest cross-correlation value is compared to a second threshold. If its cross-correlation is smaller than a threshold λ_2, the corresponding channel is identified as the backside microphone. Optimal normalized thresholds $\lambda_1 = 0.2$ and $\lambda_2 = 0.35$ have been estimated experimentally on the development set of real noisy recordings.

Algorithm 1 Channel Selection algorithm

procedure CHANNELSELECTION
  // Compute cross-correlation for pairs of channels
  for each $i \in$ [1 ... 6] **do**
    Score[ch_i] $\leftarrow 0$
    for each $j \in$ [1 ... 6] **do**
      Score[ch_i]:=Score[ch_i]+$X_{CORR}(ch_i, ch_j)$
    end for
  end for
  // Normalize score between 0 and 1
  for each $i \in$ [1 ... 6] **do**
    Score[ch_i] = Score[ch_i] / $\max$(Score)
  end for
  // Remove error channel
  for each $i \in$ [1 ... 6] **do**
    if Score[ch_i] $< \lambda_1$ **then**
      Score[ch_i] $\leftarrow$ null
    end if
  end for
  // Select backside noisy channel
  if $\min$(Score) $< \lambda_2$ **then**
    back_channel = ch_i with $\min_{i\in[1...6]}(Score[ch_i])$
  else
    back_channel $\leftarrow 0$
  end if
end procedure

3.1.2. Signal-Dependent MVDR Beamforming

Some preliminary experiments using the baseline MVDR beamformer have shown that this method is particularly good when applied on simulated noisy multichannel data, but its efficiency remains very small on real noisy recordings and can even degrade ASR results. In order to better handle multiple channel mismatch occurring in real recordings we have preferred the signal-dependent MVDR beamformer initially presented in [13]. This MVDR beamformer does not make any assumption about TDOA and therefore releases the MVDR from TDOA estimation errors. This alternative beamformer also introduces a trade-off factor $\tau \geq 0$ to balance the level of noise removal versus speech distortion in the output. A trade-off equals to zero means a more noisy but less distorted output. We studied the impact of this parameter on speech recognition. We observed a very small impact of this parameter on the ASR performance and evaluated its optimal value to $\tau = 5$ (less noisy but more distorted output).

3.1.3. Sparse NMF-based speech enhancement

We have used the sparse NMF SE method presented in [7] to remove residual noise from the enhanced signal produced by the MVDR beamformer. Sparse NMF assumes the Short Term Fourier Transform of a noisy signal $V \in \mathcal{R}^{F\times T}$ (F the number of frequency bins and T the number of time frames) is a linear combination of basis vectors $W \in \mathcal{R}^{F\times B}$ (with B the number of bases) and activation coefficients $H \in \mathcal{R}^{B\times T}$. NMF can estimate W and H by minimizing the sparseness of H in the L_1 norm. The distance between V and WH is computed in the Euclidean space by:

$$W, H = \min_{W,H} D(V||WH) + \mu||H||_1 \tag{1}$$

W and H are estimated using iterative multiplicative update rules as in Eq.2 and Eq.3, with $\overline{W}$ the column-wise L_2-normalized version of W, and $.*$ and $/$ the Hadamard product and division.

$$H \leftarrow H. * \frac{\overline{W}^T \frac{V}{\overline{W}H}}{\overline{W}^T 1 + \mu} \tag{2}$$

$$W \leftarrow W. * \frac{\frac{V}{\overline{W}H} H^T + 1(1H^T . * \overline{W}) . * \overline{W}}{1H^T + 1(\frac{V}{\overline{W}H} . * \overline{W}) . * \overline{W}} \tag{3}$$

We note W_S and W_N the speech and noise bases estimated using training recordings of clean speech and background noise. NMF-speech enhancement [10] is achieved by estimating the noise and speech coefficients H_S and H_N of a noisy speech signal using fixed speech and noise bases $[W_S\, W_N]$ and iterating on the multiplicative update rules.

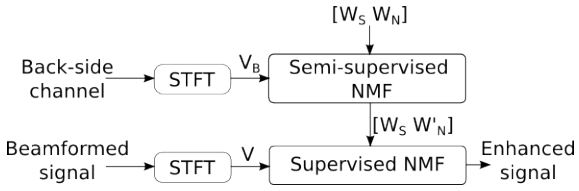

Fig. 3. Semi supervised NMF using the backside channel

We assume the backside channel, if not obstructed, may contain a significant contribution of background noise. Therefore, as presented Fig.3, this channel may help to refine the noise bases W_N for the current signal to enhance. Obviously the backside channel also contains some contribution of the speaker and it's likely that the updated NMF speech coefficients become sparser. We limit the sparseness by introducing factors θ and μ respectively used in the estimation of noise and speech coefficients ($\theta > \mu$). We set $\theta = 1.5$ and $\mu = 1$ based on perceptual evaluation using PESQ [14]. A semi-supervised MNF is achieved with the iterative multiplicative updates rules in Eq.4, Eq.5 and Eq.6. The noise base is finally updated in $W = [W_S\, W'_N]$ and used in the supervised NMF enhancement step.

$$H'_S \leftarrow H'_S . * \frac{\overline{W_S}^T \frac{V}{\overline{W}H}}{\overline{W_S}^T 1 + \mu} \tag{4}$$

$$H'_N \leftarrow H'_N . * \frac{\overline{W_N}^T \frac{V}{\overline{W}H}}{\overline{W_N}^T 1 + \theta} \tag{5}$$

$$W'_N \leftarrow W'_N . * \frac{\frac{V}{\overline{W}H} H' T_N + 1(1 H'^T_N . * \overline{W'_N}) . * \overline{W'}_N}{1 H'^T_N + 1(\frac{V}{\overline{W}H} . * \overline{W'_N}) . * \overline{W'_N}} \quad (6)$$

3.2. Automatic Speech Recognition Back-end

3.2.1. fMLLR-based DNN acoustic models

Our ASR back-end is based on Kaldi speech recognition tool kit. The speech decoder is using context-dependent DNN-HMM AM. Acoustic models have been trained on MFCC acoustic features with LDA dimension reduction, Maximum likelihood linear transformation (MLLT), and feature space maximum likelihood linear regression (fMLLR). The DNN has 5 frames of left and right context for the input layer and 7 layers with 2048 neurons per hidden layer. The DNN is pre-trained using restricted Boltzmann machines, cross entropy training, and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion [15]. To overcome some important limitations of the baseline due mostly to acoustic mismatch between the enhanced training data and the real noisy recordings, we have produced several acoustic models based on different combinations of training data.

3.2.2. RNN Language Model lattice rescoring

RNN-LM [16] has been proven able to improve the language model perplexity on various datasets by accurately capturing long-term word dependency. We have prepared a RNN-LM on the WSJ-0 $5k$ training dataset for ASR task as required by the challenge. The final perplexity is equal to 44.41.

4. CHIME CHALLENGE CORPUS

The corpus provided by CHiME challenge is composed of real and simulated noisy audio recordings. The target noisy environments are bus (BUS), cafeteria (CAF), pedestrian zone (PED) and street (STR). The simulated noisy data have been constructed with clean utterances taken from the WSJ-0 corpus [4] mixed with real background noises and microphone-specific Signal-over-Noise ratio, then convolved with a time-varying filter modelling the direct sound between the speaker and the microphones estimated using SRP-PHAT. The real noisy training data are 6-channel recordings of 4 speakers reading 100 utterances (selected from WSJ-0) in a sound proof booth (BTH) and in each of the four environments ($4 \times 4 \times 100 = 1,600$ utterances). This set of training data is supplemented by $7,138$ simulated utterances produced by mixing every WSJ-0 training audio files with each of the four background noises. A development (DEV) set and a test (TEST) set have also been provided. The set of real noisy recordings is composed of 410 (DEV) and 330 (TEST) utterances (the same that can be found in the DEV and TEST sets in the WSJ-0 5k ASR task) distributed among 4 speakers and recorded in the 4 target environments. The DEV and TEST real test datasets finally contain respectively $1,640$ (410×4) and $1,320$ (330×4) utterances. The recordings of the same set of sentences recorded in the recording booth have been artificially mixed with the real noise to prepare equivalent simulated DEV and TEST datasets.

5. EXPERIMENTS

We present the experiments we have done and the corresponding performances reached by our system on the CHiME evaluation data. We have prepared several set-ups in order to discuss the main features of our system as well as their impact on ASR performances. These set-ups consist in several variants of the front-end SE stage on the channel selection, beamforming and sparse NMF modules, as well as variants of the back-end ASR especially by comparing different acoustic features and training datasets. The performances are reported in Table 1 in terms of Word Error Rate (WER) and discussed with regards to the baseline system provided by the challenge.

5.1. ASR acoustic models and training data

In a first experiment we have trained DNN Acoustic Models (AM) with static and dynamic MFCC features normalized per-speaker by Cepstral Mean and Variance Normalization (CMVN) and speaker-adapted with feature space Maximum Likelihood Linear Regression (fMLLR). The effectiveness of fMLLR features for distant talk speech recognition was shown in [17]. The WER obtained using this AM significantly outperforms the baseline DNN-based ASR trained on Mel Filter bank features. As reported lines (3) and (4) of Table 1, our system reaches 27.21% WER on real test data compared 33.76% WER with the baseline (6.55% better). We keep this fMLLR-based AM in the next experiments.

In a second experiment we have investigated the performance of the baseline beamformer by comparing a 6-channel set-up with a mono-channel set-up (i.e. no speech enhancement) using only the audio captured by the central bottom microphone (channel 5). As expected on simulated DEV and TEST, the baseline beamformer is very efficient on the 6-channel set-up and outperforms the single-channel WER. For instance WER is equal to 14.47% on simulated DEV data for the single-channel case as shown line (1) of Table 1, and is much better with 8.46% WER on the beamformed 6-channel recordings on line (3) using similar ASR AM. We have noted that here as well our fMLLR-based AM achieves better WER on both DEV and TEST simulated data compared to baseline Mel filter bank AM (referring to lines (2) and (4)). But surprisingly we have found that the baseline beamformer on real DEV and TEST noisy recordings does not help compared to a single-channel set-up. For example, WER is equal to 16.64% (line (1)) for single-channel on real DEV data and is increased to 18.24% (line (3)) for the beamformed 6-channel on the same dataset. We assume the methods used to col-

Table 1. ASR Result in WER (%) of our system submitted to CHiME-3 Challenge.

Exp	Description	SE Front-end		ASR Back-end		DEV (WER)		TEST (WER)	
		#ch	Enhancement	Feature	Training	Simu	Real	Simu	Real
(1)	DNN baseline	1	-	Mel FB	ch5(real+simu)	14.47	16.64	20.80	33.14
(2)	DNN fMLLR	1	-	fMLLR	ch5(real+simu)	11.23	12.24	14.02	23.21
(3)	DNN baseline	6	SRP-PHAT MVDR	Mel FB	enh(real+simu)	8.46	18.24	11.19	33.76
(4)	DNN fMLLR	6	SRP-PHAT MVDR	fMLLR	enh(real+simu)	7.55	17.71	6.00	27.21
(5)	(4) + real training	6	SRP-PHAT MVDR	fMLLR	real (ch5)	4.95	13.46	6.36	23.16
(6)	(5)+Xcorr Ch.Sel. $\lambda_1 = \lambda_2 = 0.35$	6	Xcorr Ch. select., SRP-PHAT MVDR	fMLLR	real (ch5)	4.89	11.72	6.41	18.93
(7)	(6) + SD-MVDRS $\lambda_1 = 0.2$, $\lambda_2 = 0.35$	6	Xcorr Ch. select., SD-MVDR ($\tau = 5$)	fMLLR	real (ch5)	7.64	9.8	10.36	14.28
(8)	(7) + Sparse NMF	6	Xcorr Ch. select., SD-MVDR, SNMF	fMLLR	real (ch5)	8.64	11.09	11.50	16.34
(9)	(8) + NMF SE	6	Xcorr Ch. select., SD-MVDR, SNMF	fMLLR	ch5 (noisy + NMF SE)	7.47	9.32	9.96	13.48
(10)	(9) + RNN-LM	6	Xcorr Ch. select., SD-MVDR, SNMF	fMLLR	ch5 (noisy + NMF SE)	6.43	8.14	8.52	11.94

lect and to prepare the simulated and the real recordings contained in the training and test dataset differ a lot, and may generate a significant distortion and mismatch among the audio documents of the training and evaluation datasets. The training data used to prepare the DNN-based AM used in the multichannel set-ups (lines (3) and (4)) consist in $1,600$ real and $7,138$ simulated multichannel recordings processed by the baseline speech enhancement method. We assume that basically the baseline DNN-based AM may be overfitting the simulated enhanced recordings corresponding to the largest proportion of data in the training set. We have chosen to remove any simulated enhanced recordings from the multichannel training data and have prepared new DNN-based AM using only the real recordings of channel 5 available in the training dataset. Channel 5 is assumed to be less noisy and should reduce the mismatch between DNN-based AM and the enhanced real test recordings. In this process the size of the training dataset has been drastically reduced but as we can see the line (5) ASR performance is significantly improved for both real DEV and TEST data compared to previous results on line (4). On real DEV data, WER is reduces from 17.71% to 13.46% (4.25% better). On real TEST data, WER is improved by 4% with a WER score now equal to 23.16%.

Globally the adaptation of the ASR training set has allowed an absolute improvement of 4.78% (26% relative) on the real DEV set and 10.6% (31.3% relative) on the real TEST compared to the baseline system. The next set of experiments focuses on the speech enhancement components.

5.2. Channel selection, SD-MVDR and Sparse NMF

We have proposed a cross-correlation-based channel selection module in order to discard the corrupted signals and to detect the backside microphone in the 6-channel set of test recordings before applying any beamforming method. The cross-correlation values, once computed for every channels, are compared to a threshold conditioning the decision to keep or to discard the channels. In this experiment we set the threshold to these optimal values $\lambda_1 = \lambda_2 = 0.35$ in order to remove both corrupted and backside channels, assumed to be the noisiest channels of the set. The efficiency of the channel selection is absolutely crucial since MVDR beamforming is sensitive to channel failure. The ASR performances obtained using the cross-correlation channel selection are presented line (6) of Table 1. Our proposal helps to improve significantly WER on DEV and TEST real recordings compared to the results reached using the baseline channel selection approach. On DEV real data WER is improved by 1.7%, passing from 13.46% (line (5)) to 11.72% (line (6)). On real TEST data, the gain is even higher and the WER is decreased by 4.2%, from 23.16% to 18.93%.

In the next experiment we first substitute the baseline SRP-PHAT MVDR beamforming with an implementation of a TDOA-independent (or Signal Dependent) MVDR (noted SD-MVDR) initially presented in [13]. We also adapt the threshold of the cross-correlation-based channel selection to $\lambda_1 = 0.2$ to remove the corrupted channels and $\lambda_2 = 0.35$ to detect and to keep the signal detected as the backside microphone. As discussed previously the trade-off factor τ of the SD-MVDR is set to $\tau = 5$ since the value reached the best WER on the real DEV set. Globally we observed that the impact on ARS performance of the SD-MVDR trade-off τ is very small. Again, the substitution of the baseline MVDR with our proposal improves significantly ASR performance in terms of WER for both real DEV and TEST evaluation data. The results of this set-up are reported line (7) of Ta-

ble 1. On the real DEV data, WER is reduced by 1.2%, with a new score of 9.8% WER, and is reduced by 3.7% on the real TEST data with an updated WER value equal to 14.28%. As expected, changing the MVDR beamforming method impacts negatively the ASR performances on simulated DEV and TEST data since it increases the mismatch between the output of simulation process and the product of our enhancement method during the test runtime.

In a third experiment we have applied the sparse NMF-based source separation method to the beamformed mono-channel signals. We have estimated speaker-independent clean speech NMF basis from 5% of the WSJ-0 training set preliminary processed by the Short Term Fourier Transform magnitude weighted by a 32ms Hamming window and a 16ms-shift. The noise basis W_N are estimated on 4×15 minutes of background noise (bus, caf, street and pedestrian). The dimension of each basis matrix has been set to 100 and the dimension of the noise base $W_N = [W_{BUS}\, W_{CAF}\, W_{PED}\, W_{STR}]$ is therefore equal to 400. We set the sparseness constrain factor to 5 as recommended in [9] where this value yielded good objective scores. The number of iterations of the sparse NMF algorithm is set to 200. As reported line (8) of Table 1, sparse NMF decreases ASR performance by 1% to 2% absolute WER both on real and simulated evaluation data. This negative result can be explained because performing sparse NMF on evaluation data drastically distorts the output signal and increases their mismatch with the DNN-based AM used by the ASR. One possible solution reported on line (9) consists in augmenting the ASR training data with real training recordings of the channel 5 processed by the same supervised sparse NMF SE. This method will enable the capture of the NMF distortion in the acoustic models of the ASR, reducing the mismatch between training and test data. This process globally improved ASR for every evaluation data set. The WER values are now equal to 9.32% on real DEV data and to 13.48% on the real TEST data. We have also evaluated the benefit of using semi-supervised noise basis estimation by injecting the backside channel in the NMF SE step. We found that globally the backside channel only contribute for 0.5% WER to the improvement brought by the NMF process. In other words, not considering the backside channel in the NMF SE increases only by 0.5% the WER reported on line (9) of Table 1.

5.3. RNN-LM lattice re-scoring

In this last section we report the performance reached by the previous system (line (10)) after re-scoring of the hypothesis word lattice using a Recurrent Neural Network Language Model prepared on the training dataset. This WER score obtained on the real DEV dataset has been improved by 1.2% and is now equal to 8.14%. On real TEST data the WER is equal to 11.94% with a similar improvement of 1.54%.

5.4. Submitted system

Detailed WER scores obtained by our best system for every evaluation subsets and every types of noise are reported Table 2. Our system combines cross-correlation-based Channel Selection, Signal Dependent MVDR, Sparse NMF, fMLLR DNN-based AM and RNNLM re-scoring and yields an overall WER of 11.94% on real TEST evaluation data. It achieves its best results on the pedestrian environment.

Table 2. Detailed ASR results (% WER) obtained by our best system combining cross-correlation-based Channel Selection, Signal Dependent MVDR, Sparse NMF, fMLLR DNN-based AM and RNNLM re-scoring

Environment	DEV set		TEST set	
	Sim	Real	Simu	Real
BUS	6.14	10.03	6.18	17.57
CAF	7.98	7.74	9.28	12.10
PED	**5.27**	**6.46**	**8.72**	**8.48**
STR	6.36	8.33	9.90	9.62
Overall	**6.43**	**8.14**	**8.52**	**11.94**

6. CONCLUSION

We have presented our contribution to the third CHiME challenge on speech separation and recognition for noisy multi-channel recordings. The front-end of our system is performing speech enhancement by cascading a cross-correlation-based channel selection, Signal Dependent MVDR beamforming and source separation based on sparse NMF. The back-end module is a state-of-the-art ASR system with DNN acoustic models learnt on fMLLR features, and a RNN Language Model. To our knowledge this work is one of the few studies investigating NMF-based, traditionally used for single channel speech enhancement, in a task of speech recognition using DNN-based ASR. We have integrated several speech enhancement technologies into a flexible framework able to handle efficiently corrupted input signals. We also provide a large set of experiments and discuss the benefit of the component of our contribution. We show that correlation-based Channel Selection, Signal Dependent MVDR, Sparse NMF, fMLLR DNN-based AM and RNNLM re-scoring are all useful and can improve the performances of the overall system. Our best system achieves an overall WER of 11.94% on real noisy recordings and 65% relative WER improvement compared to the challenge baseline.

7. ACKNOWLEDGEMENT

This work was conducted within the Rolls-Royce@NTU Corp Lab with support from the National Research Foundation Singapore under the Corp Lab@University Scheme.

8. REFERENCES

[1] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 162–167.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, pp. 621–633, 2013.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[4] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," in *Linguistic Data Consortium*, 2007.

[5] Y. Obuchi, "Multiple-microphone robust speech recognition using decoder-based channel selection," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004, pp. 52–55.

[6] P. O. Hoyer, "Non-negative sparse coding," in *12th IEEE Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 557–565.

[7] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, pp. 88–101, 2008.

[8] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for singale-channel speech separation," in *ICASSP 2014*, 2014, pp. 3749–3753.

[9] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *ISCA Interspeech 2014*, 2014.

[10] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised seperation of sounds from single-cahnel mixtures," in *ICA 07*, 2007, pp. 414–421.

[11] X. Mestre and M. A. Lagunas, "On diagonal loading for minimum variance beamformers," in *International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003, pp. 459–462.

[12] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 41–48.

[13] E. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 21, pp. 945–958, 2013.

[14] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[15] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, 2013, pp. 2345–2349.

[16] T. Mikolov, *Statistical Language Models Based on Neural Networks*, Ph.D. thesis, Brno University of Technology, 2012.

[17] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*, 2013, pp. 19–24.