MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition

Hori, T.; Chen, Z.; Erdogan, H.; Hershey, J.R.; Le Roux, J.; Mitra, V.; Watanabe, S.

TR2015-135 December 2015

Abstract

This paper introduces the MERL/SRI system designed for the 3rd CHiME speech separation and recognition challenge (CHiME-3). Our proposed system takes advantage of recurrent neural networks (RNNs) throughout the model from the front speech enhancement to the language modeling. Two different types of beamforming are used to combine multimicrophone signals to obtain a single higher quality signal. Beamformed signal is further processed by a single-channel bi-directional long short-term memory (LSTM) enhancement network which is used to extract stacked mel-frequency cepstral coefficients (MFCC) features. In addition, two proposed noise-robust feature extraction methods are used with the beamformed signal. The features are used for decoding in speech recognition systems with deep neural network (DNN) based acoustic models and large-scale RNN language models to achieve high recognition accuracy in noisy environments. Our training methodology includes data augmentation and speaker adaptive training, whereas at test time model combination is used to improve generalization. Results on the CHiME-3 benchmark show that the full cadre of techniques substantially reduced the word error rate (WER). Combining hypotheses from different robust-feature systems ultimately achieved 9.10% WER for the real test data, a 72.4% reduction relative to the baseline of 32.99% WER.

IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)

THE MERL/SRI SYSTEM FOR THE 3RD CHIME CHALLENGE USING BEAMFORMING, ROBUST FEATURE EXTRACTION, AND ADVANCED SPEECH RECOGNITION

Takaaki Hori[1], *Zhuo Chen*[1,2], *Hakan Erdogan*[1,3], *John R. Hershey*[1],
Jonathan Le Roux[1], *Vikramjit Mitra*[4], *Shinji Watanabe*[1]

[1]Mitsubishi Electric Research Laboratories, Cambridge, MA, USA
[2]Columbia University, New York, NY, USA
[3]Sabanci University, Istanbul, Turkey
[4]SRI International, Menlo Park, CA, USA

ABSTRACT

This paper introduces the MERL/SRI system designed for the 3rd CHiME speech separation and recognition challenge (CHiME-3). Our proposed system takes advantage of recurrent neural networks (RNNs) throughout the model from the front speech enhancement to the language modeling. Two different types of beamforming are used to combine multi-microphone signals to obtain a single higher quality signal. Beamformed signal is further processed by a single-channel bi-directional long short-term memory (LSTM) enhancement network which is used to extract stacked mel-frequency cepstral coefficients (MFCC) features. In addition, two proposed noise-robust feature extraction methods are used with the beamformed signal. The features are used for decoding in speech recognition systems with deep neural network (DNN) based acoustic models and large-scale RNN language models to achieve high recognition accuracy in noisy environments. Our training methodology includes data augmentation and speaker adaptive training, whereas at test time model combination is used to improve generalization. Results on the CHiME-3 benchmark show that the full cadre of techniques substantially reduced the word error rate (WER). Combining hypotheses from different robust-feature systems ultimately achieved 9.10% WER for the real test data, a 72.4% reduction relative to the baseline of 32.99% WER.

Index Terms— CHiME-3, robust speech recognition, beamforming, noise robust feature, system combination

1. INTRODUCTION

With the wide-spread availability of portable devices equipped with automatic speech recognition (ASR), there is increasing demand for accurate ASR in noisy environments. Although great strides have been made in the advancement of recognition accuracy, background noise and reverberation continue to pose problems for the best of systems. The presence of highly non-stationary noise is typical of public areas such as a café, a street, or an airport, and tends to significantly degrade recognition accuracy in such situations. Such noises can be challenging to model and estimate due to their diverse and unpredictable spectral characteristics. Therefore, robust speech recognition in noisy environments has attracted increasing attention in ASR research and development.

Several challenge-based workshops focusing on related tasks have been recently held [1, 2, 3] to provide common data and benchmarks suitable for comparing and contrasting the performance of different methods. The 3rd CHiME speech separation and recognition challenge (CHiME-3) [4] is a new challenge task, which was designed around the well-studied Wall Street Journal corpus. In contrast with the previous CHiME challenges [1, 2], the CHiME-3 scenario focuses on typical use cases of portable devices. It features speakers talking in challenging noisy environments (cafés, street junctions, public transports and pedestrian areas), recorded using a 6-channel tablet mounted microphone array.

This paper presents the MERL/SRI system designed for CHiME-3 and its evaluation results. The goal of the study was to create an advanced system by determining the best combination of the leading methods on development data and testing their generalization to the evaluation data.

A noteworthy aspect of our system is the pervasive use of DNNs and RNNs at multiple levels throughout the system: the front end speech enhancement based on bidirectional long short-term memory (BLSTM) RNNs, DNNs for acoustic modeling, and RNNs for language modeling.

For the CHiME-3 task, our system relies on the following key technologies: (1) beamforming to enhance the target speech from the multi-channel signals; (2) noise-robust feature extraction, either directly from the signal, or by extracting standard features on the output of BLSTM-based single-channel speech enhancement; (3) DNN and LSTM acoustic models, and large-scale RNN language models; and (4) system combination of different robust-feature systems. Through a series of experiments with different combinations of these techniques, we investigate the relative contributions of the methods, and show that in combination they are surprisingly effective for the CHiME-3 task, ultimately achieving 9.10% WER for the real test data from the noisy speech model baseline of 32.99%.

2. PROPOSED SYSTEM

2.1. System overview

Figure 1 describes our proposed system. In the initial stage, we use a weighted delay-and-sum beamformer, as well as a minimum variance distortionless response (MVDR) beamformer to extract enhanced signals $\hat{y}$ and $\hat{y}'$ from 6-channel microphone array signals $\{y_1, \dots, y_6\}$, as described in Section 2.2. After the beamforming, the beamformed signals are denoised in the signal or feature domains. In one system, BLSTM-based single-channel speech enhancement is used to further enhance the weighted delay-and-sum beamformed signal $\hat{y}$, as described in Section 2.3, and MFCC features ($\mathbf{x}^B$) are extracted from the enhanced signal. As an alternative to enhancement, two types of noise robust features $\mathbf{x}^D$ and $\mathbf{x}^M$ are extracted directly from the weighted delay-and-sum beamformed $\hat{y}$ signal, as described in Section 2.4. These three methods are processed in parallel, in addition to the extraction of standard MFCC features $\mathbf{x}'$ from the MVDR-beamformed signal $\hat{y}'$.

The extracted features ($\mathbf{x}'$, $\mathbf{x}^B$, $\mathbf{x}^D$, $\mathbf{x}^M$) are each processed using a pipeline consisting of: feature-space MLLR transformation (Section 2.5), DNN-HMM hybrid decoding with the standard WSJ0 5k trigram language model (Section 2.5), and re-scoring with a 5-gram language model and RNNLM (Section 2.6). Finally, four different hypotheses are combined to provide the final result (Section 2.7).

2.2. Beamforming

We have experimented with weighted delay-and-sum (WDAS) and minimum variance distortionless response (MVDR) beamforming. Weighted delay-and-sum beamforming uses GCC-PHAT [5] cross correlation to determine candidate time delays of arrival (TDOA) between each microphone and a reference microphone. The reference microphone is chosen based on pairwise cross correlations. These time delay candidates are calculated for each segment of the signal and reconciled across segments using a Viterbi search [6]. Furthermore, weights for each microphone are determined based on cross-correlation of each microphone signal with the other microphones [6]. After finally determining delays and weights for each microphone, the beamformed signal is obtained as $\hat{y}(\tau) = \sum_{i=1}^{M} w_i y_i(\tau - \tau_i)$, where M is the number of microphones, $y_i(\tau)$ is the time-domain signal at microphone i, and w_i and τ_i are the corresponding weights and delays. We use $y_i(t, f)$ to indicate the short-time Fourier transform (STFT) of the time-domain signal $y_i(\tau)$.

An alternative beamforming method is the MVDR beamformer which minimizes the estimated noise level under the condition of no distortion in the desired signal. MVDR is a filter-and-sum beamformer whose filters can be obtained in the frequency domain as

$$[h_1(f), \dots, h_M(f)]^T = \frac{1}{\lambda(f)} \left(\boldsymbol{G}(f) - \boldsymbol{I}_{M\times M}\right) \boldsymbol{e}_{\text{ref}},$$

where $\boldsymbol{G}(f) = \boldsymbol{\Phi}_{\text{noise}}^{-1}(f)\boldsymbol{\Phi}_{\text{noisy}}(f)$ is computed from the $M \times M$ spatial covariance matrices $\boldsymbol{\Phi}_{\text{noise}}(f)$ of the noise and $\boldsymbol{\Phi}_{\text{noisy}}(f)$ of the noisy signal, and $\lambda(f) = \text{trace}(\boldsymbol{G}(f)) - M$. $\boldsymbol{e}_{\text{ref}}$ is the standard unit vector for the reference microphone, which can be chosen using maximum posterior expected SNR. Our MVDR beamformer is based on [7] and does not explicitly use TDOA estimation, hence it is different from the one provided with the released CHiME-3 system [4]. The STFT of the filter-and-sum beamformed signal can then be obtained as: $\hat{y}(t, f) = \sum_{i=1}^{M} h_i(f) y_i(t, f)$.

As shown in Figure 1, weighted delay-and-sum beamforming output was used as the input to extract features in three systems. The beamformed signal was enhanced using a BLSTM network and MFCC features of both were stacked together, and two robust feature extractions described in Section 2.4, DOC and MMeDuSA were extracted from the beamformed signal. The MVDR beamforming was used in a single extra system, as an input to MFCC feature extraction.

2.3. Speech enhancement using bidirectional LSTM

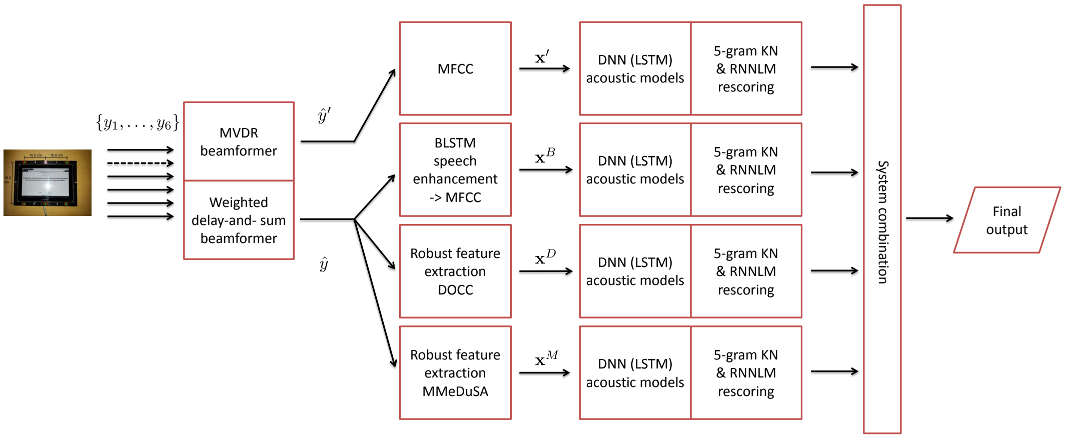
We have shown in previous work [8, 9] that LSTMs and BLSTMs are particularly efficient at dealing with highly challenging non-stationary noises for speech enhancement. Here, in one of our systems, we perform speech enhancement to deal with the noise remaining in the beamformed signals, using BLSTMs trained with phase-sensitive signal approximation (PSA) loss function [9].

Speech enhancement problem can be mathematically expressed in the STFT domain as follows:

$$\hat{y}(t, f) = g(f)s(t, f) + n(t, f),$$

where $\hat{y}(t, f)$, $s(t, f)$ and $n(t, f)$ are the STFTs of noisy, clean and noise signals respectively and $g(f)$ is the reverberation filter. We would like to recover the reverberant clean signal from the noisy signal. We can use a neural network when we are given noisy and clean signal pairs for training.

Long short-term memory (LSTM) neural network is a type of recurrent neural network (RNN) that utilizes memory cells that can potentially remember their contents for an indefinite amount of time. In recurrent networks such as LSTMs, information is passed from one layer both to the layer above as well as to the corresponding layer in the next time frame. LSTMs additionally feature a cell structure that avoids problems of vanishing or exploding gradients that commonly arise in regular RNN training. In bidirectional



$\{y_1, \ldots, y_6\}$
MVDR beamformer
Weighted delay-and- sum beamformer
$\hat{y}'$
$\hat{y}$
MFCC
BLSTM speech enhancement -> MFCC
Robust feature extraction DOCC
Robust feature extraction MMeDuSA
$\mathbf{x}'$
$\mathbf{x}^B$
$\mathbf{x}^D$
$\mathbf{x}^M$
DNN (LSTM) acoustic models
5-gram KN & RNNLM rescoring
System combination
Final output

Fig. 1: Overview of the system proposed to CHiME-3.

LSTMs (BLSTMs), there are two sequences of layers at each level, one running forward as in classical RNNs, and another running backwards, both feeding at each time step to the layers above.

2.3.1. Mask prediction

It has been shown in earlier studies of source separation that it is beneficial to predict a mask that multiplies the STFT of the mixed signal for estimating the target signal [8, 10, 11]. In such approaches, the output of the network is a mask or filter function $[\hat{a}(t, f)]_{t,f \in B} = f_W(\hat{y})$, where B is the set of all time-frequency bins and W represents neural network parameters. In this case the enhanced speech is obtained by $\hat{s}(t, f) = \hat{a}(t, f)\hat{y}(t, f)$. The input to the network is usually a set of features extracted from the STFT of the noisy signal $\hat{y}$. In earlier studies, it was shown that using logarithm of mel-filterbank energies with 100 mel-frequency bins gave good results in a task of interest [8].

In case of mask prediction, the network's loss function $\mathcal{L}(\hat{\boldsymbol{a}}) = \sum_{t,f \in B} D(\hat{a}(t, f))$ can be a mask approximation (MA) or a magnitude spectrum approximation (MSA) loss. They correspond to using distortion measures $D_{\mathrm{ma}}(\hat{a}) = |\hat{a} - a^*|^2$ and $D_{\mathrm{msa}}(\hat{a}) = (\hat{a}|\hat{y}| - |s|)^2$ respectively, where a^* is the ideal ratio mask.

2.3.2. Phase-sensitive loss function

We introduced a phase-sensitive spectrum approximation (PSA) loss function in [12], which is the complex domain distance between the reconstructed and the clean speech signals, namely $D_{\mathrm{psa}}(\hat{a}) = |\hat{a}\hat{y} - s|^2$ which is equivalent to using $D_{\mathrm{psa}}(\hat{a}) = (\hat{a}|\hat{y}| - |s|\cos(\theta))^2$ where θ is the difference angle between phases of $\hat{y}$ and s. The PSA loss function yielded better performance in source separation as compared to MSA or SA objectives in [12].

2.3.3. Channel adaptation

It is problematic to use the real training data since the close talking microphone had artifacts and did not represent the true clean speech. We thus used only the simulated training data for training, validating on the simulated development data. For BLSTM enhancement system, WDAS beamforming was performed with a fixed reference microphone, and we used the reverberated clean signal of the same microphone as the clean target during training.

As beamforming can change the scale and possibly the channel of the beamformed speech, training the network so that it reconstructs the reverberated clean speech simply by masking the beamformed noisy speech is not likely to work. To compensate for the potential scale and channel mismatch, we linearly filter the reverberated clean speech to match the beamformed speech obtained by the beamformer. This multi-frame channel adaptation filter $q_{d,f}$ is obtained by minimizing the following linear least squares loss function

$$\mathcal{L}(\boldsymbol{q}) = \frac{1}{2} \sum_{t,f} \left| \left\{ \sum_{d=-T}^{T} q(d, f) s(t + d, f) \right\} - \hat{y}(t, f) \right|^2 .$$

After channel adaptation, the mask prediction network can be trained to perform enhancement on noisy beamformed data [8, 9].

2.4. Robust feature extraction

We experimented with two main robust feature extraction techniques: *damped oscillator coefficients* (DOC) and *modulation of medium duration speech amplitudes* (MMeDuSA). In DOC processing, the auditory hair cells within the human ear are modeled as forced damped oscillators [13]. The DOC features model the dynamics of the hair-cell oscillations to auditory stimuli within the human ear. The hair cells transduce the motion of incoming sound waves and excite the neurons of the auditory nerves, which then convey the relevant information to the brain.

In DOC processing, the incoming speech signal is analyzed by a bank of bandpass gammatone filters that split the time-domain signal into subband signals. We used 40 gammatone filters that were equally spaced on the equivalent rectangular bandwidth (ERB) scale. The bandlimited subband signals from these filters serves as forcing functions to an array of 40 damped oscillators whose response was used as the acoustic feature (see [13] for details). We analyzed the damped oscillator response by using a Hamming window of 25 ms with a frame rate of 10 ms. The power signal from the damped oscillator response was computed, then compressed using the 15th root, resulting in 40-dimensional DOC features. The DOC features $\mathbf{x}^D$ used in our experiments are either these original features, or their cepstral version, which is generated by performing a Discrete Cosine Transform (DCT), keeping only the first 13 coefficients including C0.

MMeDuSA [14] tracks the subband amplitude modulation (AM) signals of speech by using a medium duration analysis window. On top of tracking the subband AM signals, MMeDuSA also tracks the overall summary modulation information. The summary modulation plays an important role in both tracking voiced speech and locating events such as vowel prominence/stress, etc. MMeDuSA directly uses the nonlinear Teager energy operator [15] to crudely estimate the AM signal from the bandlimited subband signals. The MMeDuSA generation pipeline used a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. A Hamming analysis window of 51 ms with a 10 ms frame rate was used. The magnitudes were then compressed using the 15th root. The MMeDuSA acoustic features $\mathbf{x}^M$ used in our experiments are either these original features, or their cepstral version, which is obtained by performing DCT separately over subband AM signals, keeping only the first 13 coefficients, and over summary AM signals, keeping only the first 3 coefficients, finally concatenating them both (for details, please refer to [14]).

2.5. Acoustic modeling

In this challenge, our proposed system mainly uses standard (11 frame context) DNN acoustic models followed by state-level sequence discriminative training, as prepared by the CHiME-3 baseline script. In addition, we also investigated the use of long-range speech/noise characteristics in the acoustic model by using a longer context (15 frames) DNN or using an LSTM. Both attempts often improved the WERs (e.g., 15-frames DNN reduced the WER by 0.5% absolutely in the real-data evaluation set, and LSTM by 1% absolutely). However, these improvements were not consistent with the development set results, and their training time was increased drastically. Therefore, the experimental section only reports the results using the standard DNN.

As the CHiME-3 challenge rules allow for the use of speaker label information, we investigated transforming the features using feature-space maximum likelihood linear regression (fMLLR) [16], i.e., $\bar{\mathbf{x}}_t = \mathbf{A}_s\hat{\mathbf{x}}_t + \mathbf{b}_s$, where s denotes a speaker index. Accounting for speaker variability using fMLLR is convenient in a DNN-based framework because fMLLR is applied directly to the features, and the structure of the system thus does not need to be modified. The fMLLR transform $(\{\mathbf{A}_s, \mathbf{b}_s\})$ was estimated using a Gaussian Mixture Model (GMM) based ASR system by iteratively maximizing the likelihood of the data given the transcription alignments for the training data, and the one-best hypothesis alignment obtained by the system for the test data. The DNN-based systems were then trained on or applied to the fMLLR-transformed features.

2.6. Language modeling

Our CHiME-3 system employs a recurrent neural network language model (RNNLM) [17] and a 5-gram language model with a modified Kneser-Ney smoothing [18, 19], which are trained using the WSJ0 text corpus. The RNNLM is an effective language model, which is represented as a neural network including a hidden layer with re-entrant connections to itself with one-word delay. The activations of the hidden units play a role of *memory* keeping a history from the beginning of the speech. Accordingly, the RNNLM can robustly estimate word probability distributions by representing the histories smoothed in the continuous space and by taking long-distance interword dependencies into account. Mikolov et al. reported that RNNLMs yielded a large gain in recognition accuracy when combined with a standard n-gram model [17].

In the decoding phase, word lattices are first generated using the baseline language model for CHiME-3, which is the standard 5k WSJ trigram with entropy pruning. After that, N-best lists are generated from the lattices using the 5-gram model. Finally, the N-best lists are reranked using a linear combination of the 5-gram and RNN LMs, i.e., $P(W) = \Pi_{i=1}^{L}(\lambda P_{rnn}(w_i|h_i) + (1-\lambda)P_{5gkn}(w_i|h_i)$ for each sentence hypothesis $W = w_1, w_2 \ldots, w_L$, where λ denotes the interpolation weight and h_i is the history of w_i. The best-ranked hypothesis is selected as the result of each single system. The N-best lists are also used for system combination.

2.7. System combination

System combination is a technique to improve recognition accuracy by combining different recognition outputs [20]. For our CHiME-3 system, each feature extraction output is separately processed by the recognizer to output lattices or N-best lists. These multiple hypotheses are then combined by taking into account the word posterior probabilities. Finally, a confusion network is constructed and the sequence of words that has the best posterior probability in each confusion set is selected. This procedure results in the word sequence with the minimum Bayes risk (MBR), i.e., the minimum expected word error.

3. EXPERIMENTS

3.1. CHiME-3 Task

The 3rd CHiME challenge consists of two types of data, real and simulated. The real data were recorded in four real noisy environments (on buses, in cafés, in pedestrian areas, and at street junctions) uttered by actual talkers. The simulated data were noisy utterances generated by artificially mixing clean speech data convoluted with estimated impulse responses of an environment, with background noises separately recorded in that environment.

To evaluate systems, training, development, and test sets are provided by the CHiME-3 organizers. The training set consists of 1600 real noisy utterances from 4 speakers, and 7138 simulated noisy utterances from the 83 speakers forming the WSJ0 SI-84 training set, in the 4 noisy environments. The transcriptions are also based on those of the WSJ0 SI-84 training set. The development set consists of 410 real and 410 simulated utterances in each of the 4 environments, for a total of 3280 utterances from 4 other speakers than those in the training set. The test set contains 330 real and 330 simulated utterances in each of the 4 environments, for a total of 2640 utterances from 4 other speakers than those in the training and development sets. The WSJ0 text corpus is also available to train language models.

3.2. Baseline ASR results

The organizers also provided baseline software to perform data simulation, speech enhancement, and ASR. The ASR baseline uses the Kaldi ASR toolkit [21]. Table 1 shows the baseline performance given by the software without speech enhancement, where acoustic models based on GMMs and DNNs were trained. The DNNs were trained based on Cross Entropy (CE) and state-level Minimum Bayes Risk (sMBR) criteria. We consider the baseline WER for the real test set to be 32.99%, which was generated by the GMM-based system[1].

In the following experiments, we investigate the performance gains of the different techniques used in our system and their combinations.

Table 1: CHiME-3 baseline WERs.

method	sim-dev	real-dev	sim-test	real-test
GMM	18.46	18.55	21.84	**32.99**
DNN (CE)	16.23	18.45	25.00	38.47
DNN (sMBR)	**14.30**	**16.13**	**21.51**	33.43

3.3. Beamforming and speech enhancement

We used the "Beamformit" beamforming toolkit for implementing weighted delay-and-sum (WDAS) beamforming [6]. Beamformit was performed by using only the 5 microphones that are facing the speaker. We excluded microphone 2 since it faces the other direction and contains less speech. Experiments showed that this leads to better performance. The Beamformit algorithm was run in segment mode to provide weighted delay-and-sum beamforming every half a second.

Table 2: CHiME-3 WERs using various beamforming and enhancement methods with MFCC features in a GMM-based system trained on *clean* speech.

method	sim-dev	real-dev	sim-test	real-test
w/o BF	50.31	55.65	63.32	79.80
WDAS	36.91	31.55	57.81	57.06
MVDR	34.63	49.89	39.98	68.34
WDAS+BLSTM enh.	**17.51**	**15.68**	**27.16**	**29.74**

Table 3: CHiME-3 WERs using various beamforming and enhancement methods with MFCC features in a GMM-based system retrained on *enhanced* speech.

method	sim-dev	real-dev	sim-test	real-test
w/o BF	18.46	18.55	21.84	32.99
WDAS	15.10	12.53	22.96	**22.88**
MVDR	15.73	17.97	**15.47**	26.07
WDAS+BLSTM enh.	**13.34**	**12.36**	19.49	23.21

We also implemented an MVDR beamformer which does not require explicit calculation of delays [7]. This beamforming requires a good estimation of the noise spatial covariance matrix, which we obtained using data from the beginning and end parts of each utterance. The noisy signal's spatial covariance matrix was estimated from the whole utterance. MVDR beamforming was performed using only reliable channels. We automatically determined channel reliability based on ad-hoc measures such as high frequency energy content and also whether the energy profile of the signal was changing too fast. The channels deemed unreliable were left out of MVDR beamforming. The reference microphone was chosen as the one obtaining highest estimated posterior SNR, except microphone 2 which was never chosen.

Tables 2 and 3 compare the beamforming methods using WDAS and MVDR with the GMM systems. Overall, both methods improved the performance from the GMM baseline without enhancement (for both trained and retrained systems), with a significantly better performance for WDAS on the real dev set. In addition, the BLSTM enhancement was applied to the WDAS-beamformed signals, leading to a large improvement for the GMM-based system trained on clean speech (30-50% absolutely), and to comparable results for the retrained system. We thus mainly use the simpler WDAS system in the following experiments, but consider MVDR and BLSTM as well for system combination, as they lead to comparable performance on the simulation dev set and/or real dev set. Since retraining with enhanced features improves performance drastically, we always report results after retraining in the rest of the paper.

[1] The WERs in the table are slightly different from the official CHiME-3 results, but their trend is very similar. These differences are likely to come from parameter initialization and machine specific issues.

3.4. Noise-robust features and fMLLR adaptation

Table 4 reports recognition results with noise robust features. Our proposed noise robust features (MMeDuSA, DOC, and their cepstrum versions) were extracted from WDAS-beamformed signals. All the results are obtained with DNN-based recognition systems. First, we compared the result of MMeDuSA and DOC features with log mel filterbank features (log mel), and MMeDuSA and DOC consistently improved the performance between 0.48-4.60% absolute. In addition, we also used fMLLR transformation for the cepstrum versions of these features, and we obtained further improvement. In this experiment, MMeDuSA performed better than DOC, but both systems were used for system combination. Hereafter, we only report results using fMLLR transforms without explicitly referring to it in the tables.

3.5. Large-scale language models

In the above experiments, we used only the baseline 3-gram language model. Hereafter we introduce large-scale language models to further improve ASR performance. A Kneser-Ney smoothed 5-gram model (5-gram KN) and an RNN language model (RNNLM) were trained on the WSJ0 text corpus. The RNNLM is a class-based model with 200 word classes and 500 hidden units. The 5-gram and RNN LM probabilities were linearly combined, where the best combination weights were chosen using the development set. Table 5 shows the result of using the above advanced language models for the MMeDuSA DNN system, explained at the previous section, and the DNN system with WDAS BLSTM enhancement, as explained in Section 3.3. Note that we stacked two types of MFCC features w/ and w/o BLSTM enhancement for the input of the DNN to make the feature smooth. Both 5-gram KN and RNN LM further improved the performance from the 3-gram model. Large-scale language models are also effective for the other systems (DOC and MVDR-MFCC) but we omit them from the paper for brevity.

3.6. System combination

Finally, the outputs of different enhancement/feature systems were combined. In all the systems, we used 5-gram KN and RNN language models. The N-best lists rescored by the language model were combined into one list and MBR decoding is performed for the list to obtain the minimum Bayes risk word sequence hypothesis.

Table 4: CHiME-3 WERs with a DNN-based recognizer for WDAS-beamformed signals using noise-robust features and fMLLR-transformed cepstrum versions of the robust features.

features	sim-dev	real-dev	sim-test	real-test
log mel	12.58	10.66	23.86	20.17
DOC	12.00	10.18	20.35	18.53
MMeDuSA	**10.83**	**9.54**	**19.26**	**18.27**
DOC-fMLLR	10.06	8.68	17.10	15.28
MMeDuSA-fMLLR	**9.73**	**8.39**	**16.30**	**14.96**

Table 5: Effect of 5-gram KN smoothing and RNNLM on WER for MMeDuSA and WDAS BLSTM systems (DNN).

LM	sim-dev	real-dev	sim-test	real-test
MMeDuSA 3-gram	9.73	8.39	16.30	14.96
+ 5-gram KN	8.76	7.33	14.56	13.48
+ RNN LM	**7.20**	**6.70**	**12.39**	**11.23**
WDAS BLSTM 3-gram	9.15	8.70	15.14	18.44
+ 5-gram KN	8.42	7.71	13.64	17.28
+ RNN LM	**6.59**	**6.23**	**11.39**	**14.55**

Table 6: CHiME-3 WERs for various features using a DNN recognizer, and system combination result using MBR decoding.

system	sim-dev	real-dev	sim-test	real-test
a: DOC	7.21	5.88	13.03	11.36
b: MMeDuSA	7.20	5.76	12.39	11.23
c: BLSTM	6.62	5.80	10.69	12.42
d: MVDR[2]	6.71	6.60	5.69	10.90
MBR(a,b,c,d)	**5.44**	**4.63**	**8.55**	**9.10**

Table 6 shows the results of combining 4 systems (WDAS-DOC, WDAS-MMeDuSA, WDAS-BLSTM-MFCC, MVDR-MFCC). By increasing the number of systems, the WER is consistently reduced from those of the individual systems. Thus, the enhancement techniques and robust features we introduced have complementary properties that yield substantial improvements after MBR system combination. We also confirmed the effectiveness of the BLSTM enhancement by excluding it from system combination, i.e., the WER result for MBR(a,b,d) was worse for the real-test set than MBR(a,b,c,d). Our final system achieved 9.10%.

4. SUMMARY

We presented the MERL/SRI system proposed to address the 3rd CHiME speech separation and recognition challenge (CHiME-3). To achieve high speech recognition accuracy in that scenario, we extended our recurrent neural network-based system by applying (1) beamforming, (2) enhancement and noise-robust feature extraction, (3) advanced speech recognition back-end including large-scale RNN language models, and (4) system combination of different enhancement/robust-feature systems. We reported the results on the CHiME-3 benchmark, showing substantial reduction of word error rate (WER) from the baseline. By combining multiple hypotheses from the different robust-feature systems, we finally achieved 9.10% WER for the real test data, a 72.4% reduction over the noisy speech model baseline of 32.99%.

Acknowledgement: Hakan Erdogan was partially supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the BIDEB 2219 program.

[2] MVDR results are boosted by using the alignment obtained from the MMeDuSA 3-gram system.

5. REFERENCES

[1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[2] 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 *Proc. ICASSP*, May 2013, pp. 126–130.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, “The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Proc. WASPAA*, Oct 2013, pp. 1–4.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ’CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. ASRU*, 2015.

[5] M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in *Proc. ICASSP*, vol. 1, 1997, pp. 375–378.

[6] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[7] M. Souden, J. Benesty, and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 260–276, 2010.

[8] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Proc. GlobalSIP Symposium on Machine Learning Applications in Speech Processing*, Dec. 2014.

[9] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. ICASSP*, Apr. 2015.

[10] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, pp. 1849–58, 2014.

[11] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7092–7096.

[12] H. Erdogan, J. R. Hershey, J. Le Roux, and S. Watanabe, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. ICASSP*, Apr. 2015.

[13] V. Mitra, H. Franco, and M. Graciarena, “Damped oscillator cepstral coefficients for robust speech recognition,” in *Proc. Interspeech*, 2013, pp. 886–890.

[14] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, “Medium duration modulation cepstral feature for robust speech recognition,” in *Proc. ICASSP*, 2014.

[15] H. Teager, “Some observations on oral air flow during phonation,” *IEEE Trans. ASSP*, 1980.

[16] M. J. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[17] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, , and S. Khudanpur, “Recurrent neural network based language model,” in *Proc. Interspeech*, 2010, pp. 1045–1048.

[18] R. Kneser and H. Ney, “Improved backing-off for M-gram language modeling,” in *Proc. ICASSP*, 1995, pp. 181–184.

[19] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” in *Proc. ACL*, 1996, pp. 310–318.

[20] G. Evermann and P. C. Woodland, “Posterior probability decoding, confidence estimation and system combination,” in *Proc. NIST Speech Transcription Workshop*, 2000.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.