Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language 00 (2017) 1–18
www.elsevier.com/locate/csl

Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend☆

Takaaki Hori*[,a], Zhuo Chen[a,b], Hakan Erdogan[a,c], John R. Hershey[a], Jonathan Le Roux[a], Vikramjit Mitra[d], Shinji Watanabe[a]

[a] *Mitsubishi Electric Research Laboratories, Cambridge, MA, USA*
[b] *Columbia University, New York, NY, USA*
[c] *Sabanci University, Istanbul, Turkey*
[d] *SRI International, Menlo Park, CA, USA*

Received 11 April 2016; received in revised form 8 December 2016; accepted 19 January 2017

Abstract

This paper gives an in-depth presentation of the multi-microphone speech recognition system we submitted to the 3rd CHiME speech separation and recognition challenge (CHiME-3) and its extension. The proposed system takes advantage of recurrent neural networks (RNNs) throughout the model from the front-end speech enhancement to the language modeling. Three different types of beamforming are used to combine multi-microphone signals to obtain a single higher-quality signal. The beamformed signal is further processed by a single-channel long short-term memory (LSTM) enhancement network, which is used to extract stacked mel-frequency cepstral coefficients (MFCC) features. In addition, the beamformed signal is processed by two proposed noise-robust feature extraction methods. All 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 multi-channel noisy data training 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 set of techniques substantially reduced the word error rate (WER). Combining hypotheses from different beamforming and robust-feature systems ultimately achieved 5.05% WER for the real-test data, an 84.7% reduction relative to the baseline of 32.99% WER and a 44.5% reduction from our official CHiME-3 challenge result of 9.1% WER. Furthermore, this final result is better than the best result (5.8% WER) reported in the CHiME-3 challenge.

Keywords: 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

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: thori@merl.com (T. Hori).

http://dx.doi.org/10.1016/j.csl.2017.01.013

advancement of recognition accuracy, background noise and reverberation continue to pose problems even for the best systems. The presence of highly non-stationary noise is typical of public areas such as cafés, streets, or airports, 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 (Barker et al., 2013; Vincent et al., 2013; Kinoshita et al., 2013) 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) (Barker et al., 2015) is a new challenge task, which was designed around the well-studied Wall Street Journal corpus. In contrast with the previous CHiME challenges (Barker et al., 2013; Vincent et al., 2013), 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.

The CHiME-3 challenge has successfully finished in December 2015. 26 systems were submitted and various strategies that improved the recognition accuracy were proposed and discussed (Barker et al., 2015). We built the MERL/SRI system for CHiME-3 and achieved the 2nd best result among the 26 systems (Hori et al., 2015). 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. Although our system achieved a good level of performance for the challenge task, it was not yet complete as regards to exploiting all the component technologies in the best combination. In this paper, we further extend our previous work and present the complete system and new evaluation results, eventually achieving a better word error rate to that of the best system in the CHiME-3 challenge.

A noteworthy aspect of our system is the pervasive use of deep neural networks (DNNs) and recurrent neural networks (RNNs) at multiple levels throughout the system: the front-end speech enhancement based on long short-term memory (LSTM) RNNs, DNNs for acoustic modeling, and LSTM RNNs for language modeling. Furthermore, we apply noisy data training of DNN acoustic models, which improves the recognition accuracy as reported in Seltzer et al. (2013), Narayanan and Wang (2014) and Delcroix et al. (2015).

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 beamformed signal, or from the output of LSTM-based single-channel speech enhancement after beamforming; (3) DNN acoustic models, and large-scale LSTM RNN language models; and (4) system combination of different beamforming/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 5.05% WER for the real-test data, an 81.8% reduction relative to the baseline of 32.99% WER and a 44.5% reduction from our official CHiME-3 challenge result of 9.1% WER.

2. Proposed system

2.1. System overview

Fig. 1 describes our proposed system, which is separated into training and recognition stages. In the training stage, 6-channel microphone array signals $\{y_1, \cdots, y_6\}$ are processed independently by three feature extraction modules for mel-frequency cepstral coefficients (MFCCs), damped oscillator coefficient cepstrum (DOCC) and modulation of medium duration speech amplitudes (MMeDuSA), and converted to feature vector sequences $\{\mathbf{x}_1, \cdots, \mathbf{x}_6\}$, $\{\mathbf{x}_1^D, \cdots, \mathbf{x}_6^D\}$ and $\{\mathbf{x}_1^M, \cdots, \mathbf{x}_6^M\}$, respectively, where DOCC and MMeDuSA are noise robust features described in Section 2.4. After that, a DNN acoustic model for each feature extraction method is created by cross-entropy (CE) training followed by state-level Minimum Bayes Risk (sMBR) training. In the training phase, we do not use any speech enhancement techniques based on microphone arrays, and simply deal with the 6-channel signals independently to obtain a larger data set which is 6 times larger than the actually spoken data. The advantage of this architecture is demonstrated in Section 3.

In the recognition stage, we use three types of beamforming, a weighted delay-and-sum (WDAS) beamformer, a minimum variance distortionless response (MVDR) beamformer and a generalized eigenvector (GEV) beamformer to extract enhanced signals $\hat{y}$, $\hat{y}'$ and $\hat{y}''$ from 6-channel microphone array signals $\{y_1, \cdots, y_6\}$, as described in

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

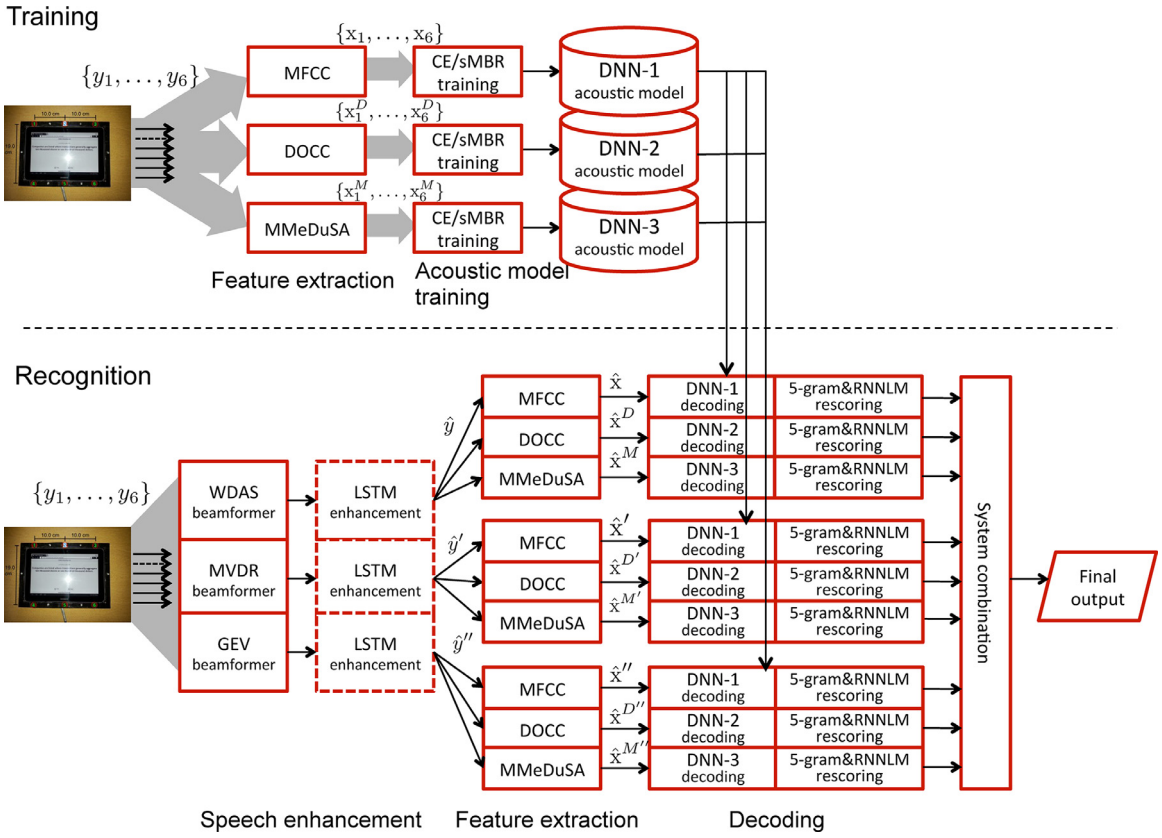

Fig. 1. Overview of the proposed system that is the extended version of our official CHiME-3 system.

Section 2.2. After the beamforming, the beamformed signals are denoised in the signal or feature domains. LSTM-based single-channel speech enhancement is used to further enhance WDAS-processed signal $\widehat{y}$, MVDR-processed signal $\widehat{y}'$ and GEV-processed signal $\widehat{y}''$ as described in Section 2.3. However, this enhancement may be skipped according to the training and test conditions, i.e. the beamformed signals can be directly given to the feature extraction modules. In fact, the LSTM enhancement was included in our official CHiME-3 system, but finally excluded in the extended system. Then, three types of features, MFCC, DOCC and MMeDuSA, are extracted as in the training stage, where these features are not obtained from the raw signals of the microphone array but from the three types of beamformed signals $\widehat{y}$, $\widehat{y}'$ and $\widehat{y}''$. Accordingly, we obtain nine feature vector sequences $\{\widehat{\mathbf{x}}, \widehat{\mathbf{x}}^D, \widehat{\mathbf{x}}^M, \widehat{\mathbf{x}}', \widehat{\mathbf{x}}^{D'}, \widehat{\mathbf{x}}^{M'}, \widehat{\mathbf{x}}'', \widehat{\mathbf{x}}^{D''}, \widehat{\mathbf{x}}^{M''}\}$ by combining the three beamformers and the three feature extraction methods.

The extracted features are each processed using a pipeline consisting of: feature-space MLLR transformation (Section 2.5), DNN-HMM hybrid decoding with the DNN acoustic model trained in the corresponding feature space and 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, the nine different hypotheses are combined to provide the final result (Section 2.7).

2.2. *Beamforming*

We have experimented with weighted delay-and-sum (WDAS), minimum variance distortionless response (MVDR), and generalized eigenvector (GEV) beamforming.

2.2.1. *WDAS beamforming*

Weighted delay-and-sum beamforming uses GCC-PHAT (Brandstein and Silverman, 1997) 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 (Anguera et al., 2007).

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

Furthermore, weights for each microphone are determined based on the cross-correlation of each microphone signal with the other microphones (Anguera et al., 2007). After finally determining delays and weights for each microphone, the beamformed signal is obtained as $\widehat{y}(\tau) = \sum_{i=1}^{M} w_i y_i(\tau - \tau_i)$, 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 $y_i(\tau)$.

2.2.2. MVDR beamforming

An alternative beamforming method is the MVDR beamformer which minimizes the estimated noise level under the condition of no distortion in the desired signal. Our MVDR beamformer is based on Benesty et al. (2008) and Souden et al. (2010). This formulation is somewhat different than conventional MVDR and does not require an explicit steering vector estimation which we explain below.

Let $Y(t,f) = [y_1(t,f), \ldots, y_M(t,f)]^T$ be the spatial time−frequency signal composed of all microphone signals. MVDR seeks to find a spatial filter $h(f) = [h_1(f), \ldots, h_M(f)]^T$ to perform filter-and-sum beamforming with $\widehat{y}(t,f) = h(f)^H Y(t,f)$. We define channel factors for each microphone (in other words, the steering vector) with $b(f) = [b_1(f), \ldots, b_M(f)]^T$, where $Y(t,f) = s(t,f)b(f) + N(t,f)$ and where $s(t, f)$ is the STFT of the speech source and $\boldsymbol{N}(t, f)$ is the spatial time−frequency domain diffuse noise signal.

We define spatial covariance of noise as the complex $M \times M$ matrix $\Phi_{\text{noise}}(f) = E\left(N(t,f)N(t,f)^H\right)$ and spatial covariance of the noisy signal is similarly defined as $\Phi_{\text{noisy}}(f) = E\left(Y(t,f)Y(t,f)^H\right)$. We can estimate the latter one by averaging across frames. The noise spatial covariance requires a mask which will indicate speech and noise time−frequency bins, and hence we can estimate it by:

$$\Phi_{\text{noise}}(f) = \frac{1}{T}\sum_{t=0}^{T-1}\left(1 - \widehat{a}(t,f)\right)^2 Y(t,f)Y(t,f)^H, \tag{1}$$

where $\widehat{a}(t,f)$ is a predicted speech mask which we can obtain from an enhancement network. We assume spatial covariances do not change over time in an utterance due to stationarity. A similar estimate for the spatial covariance of speech $\boldsymbol{\Phi}_{\text{speech}}(f)$ can be calculated using the mask $\widehat{a}(t,f)$ as well. The covariance estimation method we use is an offline one which utilizes the whole utterance and the spatial covariance estimate is fixed throughout the utterance. In addition, we do not make use of a separate voice activity detector since the mask we obtain from the network provides an estimation of which time−frequency bins belong to speech and which belong to noise.

MVDR attempts to minimize the noise energy after beamforming while having no distortion for the speech part by solving the following problem:

$$\widehat{h}(f) = \ \text{argmin}_{h(f)} h(f)^H \Phi_{\text{noise}}(f) h(f) \tag{2}$$

$$\text{such that } h(f)^H b(f) = b_{\text{ref}}(f) \tag{3}$$

Here ref is the index of a chosen reference microphone. The solution to this problem is as follows:

$$\widehat{h}(f) = \frac{\Phi_{\text{noise}}^{-1}(f)b(f)b(f)^H e_{\text{ref}}}{b(f)^H \Phi_{\text{noise}}^{-1}(f)b(f)}, \tag{4}$$

which can be simplified by observing that $\Phi_{\text{speech}}(f) = \boldsymbol{\sigma}^2(f)b(f)b(f)^H$ where $\sigma^2(f)$ is the variance of the speech source. Using a matrix identity for the denominator and using the fact that $\Phi_{\text{noisy}}(f) = \Phi_{\text{speech}}(f) + \Phi_{\text{noise}}(f)$, this leads to the following result:

$$\widehat{h}(f) = \frac{1}{\lambda(f)}(G(f) - I_{M\times M})e_{\text{ref}}, \tag{5}$$

where $G(f) = \Phi_{\text{noise}}^{-1}(f)\Phi_{\text{noisy}}(f)$ is computed from the spatial covariance matrices and $\lambda(f) = \text{trace}(G(f)) - M$. $\boldsymbol{e}_{\text{ref}}$ is the standard unit vector for the reference microphone, which can be chosen using maximum posterior SNR given by:

$$\text{SNR}_{\text{post},r} = \frac{\sum_{f=0}^{F-1} h_r^H(f)\Phi_{\text{speech}}(f)h_r(f)}{\sum_{f=0}^{F-1} h_r^H(f)\Phi_{\text{noise}}(f)h_r(f)}, \tag{6}$$

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

where $\boldsymbol{h}_r(f)$ is the MVDR solution when using microphone r as the reference. We choose ref $=$ argmax$_r$SNR$_{\text{post},r}$ which gives the highest posterior SNR and our final estimated MVDR spatial filter is given by $\widehat{h}(f) = h_{\text{ref}}(f)$. Our MVDR beamformer is based on Benesty et al. (2008) and Souden et al. (2010) and does not explicitly use TDOA estimation, hence it is different from the one provided with the released CHiME-3 system (Barker et al., 2015).

2.2.3. GEV beamforming

The GEV beamformer (Warsitz and Haeb-Umbach, 2007; Heymann et al., 2016) attempts to find spatial filters that will be used in filter-and-sum beamforming by maximizing posterior SNR for each frequency bin. The SNR for frequency f after beamforming can be expressed as follows when we use a spatial filter $\boldsymbol{h}(f)$:

$$\text{SNR}_{\text{post}}(f) = \frac{h^H(f)\Phi_{\text{speech}}(f)h(f)}{h^H(f)\Phi_{\text{noise}}(f)h(f)}. \tag{7}$$

In GEV beamforming, the filters are chosen to maximize this posterior SNR separately for each frequency. The solution that maximizes this quantity is given by the solution to the following generalized eigenvalue problem:

$$\Phi_{\text{speech}}(f)h(f) = \boldsymbol{\lambda}\Phi_{\text{noise}}(f)h(f). \tag{8}$$

So, we simply choose $\widehat{h}(f)$ as the eigenvector corresponding to the largest eigenvalue of the generalized eigenvalue problem. In GEV, we do not choose a reference microphone.

The STFT of the filter-and-sum beamformed signal for both MVDR and GEV beamformers can then be obtained using the spatial filter $\widehat{h}(f)$ as:

$$\widehat{y}(t,f) = \sum_{i=1}^{M} \widehat{h}_i^*(f)y_i(t,f). \tag{9}$$

As shown in Fig. 1, we utilize all three types of beamforming, namely WDAS, MVDR and GEV in our system. All beamformer outputs were used to extract three kinds of features for each one of them to yield nine separate decoding hypotheses.

2.3. Speech enhancement using LSTM

We have shown in previous work (Weninger et al., 2014a; Erdogan et al., 2015b) that LSTMs and BLSTMs are particularly effective 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 LSTMs trained with phase-sensitive signal approximation (PSA) loss function (Erdogan et al., 2015b).

The observed signal in a single channel speech enhancement problem can be mathematically expressed in the STFT domain as follows:

$$\widehat{y}(t,f) = g(f)s(t,f) + n(t,f), \tag{10}$$

where $\widehat{y}(t,f)$, $s(t,f)$ 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.

A 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 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.

In earlier speech separation studies (Erdogan et al., 2015b; Weninger et al., 2014b), we have shown that (B) LSTMs perform better than DNNs with spliced inputs for source separation. The reason for this result may be due to the fact that LSTMs make use of contextual information in a better way than DNNs and they are better suited for sequential inputs.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

2.3.1. Mask prediction

It has been shown in earlier studies on source separation that it is beneficial to predict a mask that multiplies the STFT of the mixed signal for estimating the target signal (Weninger et al., 2014a; Wang et al., 2014; Narayanan and Wang, 2013). In such approaches, the output of the network is a mask or filter function $[\widehat{a}(t,f)]_{(t,f)\in B} = f_W(\widehat{y})$, 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 $\widehat{s}(t,f) = \widehat{a}(t,f)\widehat{y}(t,f)$. The input to the network is usually a set of features extracted from the STFT of the noisy signal $\widehat{y}$. 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 (Weninger et al., 2014a).

In case of mask prediction, the network's loss function $\mathcal{L}(\widehat{a}) = \sum_{(t,f)\in B} D\Big(\widehat{a}(t,f)\Big)$ can be a mask approximation (MA) or a magnitude spectrum approximation (MSA) loss. They correspond to using distortion measures $D_{\text{ma}}(\widehat{a}) = |\widehat{a} - a^*|^2$ and $D_{\text{msa}}(\widehat{a}) = (\widehat{a}|\widehat{y}| - |s|)^2$ 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 Erdogan et al. (2015a), which is the complex domain distance between the reconstructed and the clean speech signals, namely $D_{\text{psa}}(\widehat{a}) = |\widehat{a}\widehat{y} - s|^2$ which is equivalent to using $D_{\text{psa}}(\widehat{a}) = \Big(\widehat{a}|\widehat{y}| - |s|\cos(\theta)\Big)^2$ where θ is the difference angle between phases of $\widehat{y}$ and s. The PSA loss function yielded better performance in source separation as compared to MSA or SA objectives in Erdogan et al. (2015a).

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 (Mitra et al., 2013). 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.

2.4.1. Damped Oscillator Coefficients (DOC)

DOC features are motivated by studies that claim auditory hair cells exhibit damped oscillations in response to external stimuli (Neiman et al., 2011) and such oscillations result in enhanced sensitivity and sharper frequency responses. DOC processing, the incoming speech signal is analyzed by a bank of bandpass gammatone filters that split the time-domain signal into subband signals. Based on prior observations (Mitra et al., 2015), we have used 40 gammatone filters that were equally spaced on the equivalent rectangular bandwidth (ERB) scale. The bandlimited subband signals from these filters serve as forcing functions to an array of 40 forced damped oscillators (FDO) whose response was used as the acoustic feature (see Mitra et al., 2013 for details). The DOC feature extraction block diagram is shown in Fig. 2. The damped oscillator response is smoothed using a modulation filter with cutoff frequencies at 0.9 Hz and 100 Hz, which helps to reduce the background subband noise, please refer to Fig. 3. Additionally the fine-grained temporal resolution of forced damped oscillator processing coupled with the long term memory associated with FDOs help to reduce late reverberation effects in the speech signal, please refer to Fig. 4.

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 $\widehat{\mathbf{x}}^D$, $\widehat{\mathbf{x}}^{D'}$ and $\widehat{\mathbf{x}}^{D''}$ used in our experiments are their cepstral

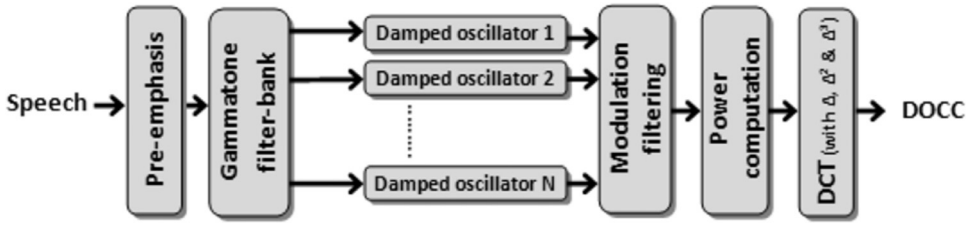

Fig. 2. Block diagram of damped oscillator-based feature extraction.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

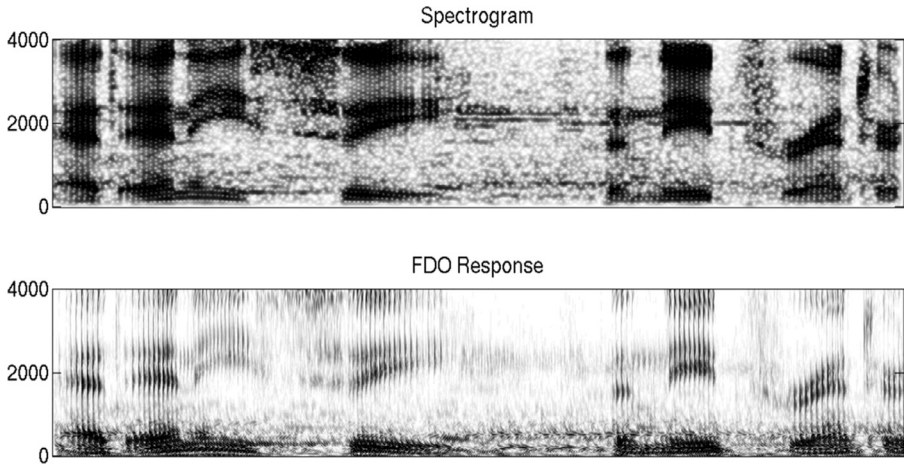

Fig. 3. Spectrogram vs. FDO response for speech corrupted with babble noise.

version, the DOC Cepstrum (DOCC), which is generated by performing a Discrete Cosine Transform (DCT), keeping only the first 13 coefficients including C_0.

2.4.2. *Modulation of medium duration speech amplitudes (MMeDuSA)*

Studies (Drullman et al., 1994; Ghitza, 2001) have indicated that amplitude modulation (AM) of the speech signal plays an important role in speech perception and recognition. Hence, recent studies (Mitra et al., 2012; Potamianos and Maragos, 2001) have treated the speech signal as a sum of amplitude-modulated (AM) narrow-band signals. MMeDuSA (Mitra et al., 2014) tracks the subband amplitude modulation (AM) signals of speech by using the Teager's energy operator (Teager, 1980). In the MMeDuSA pipeline the speech signal is first pre-emphasized and then analyzed using a gammatone filterbank with 40 channels equally spaced on the ERB scale. A medium duration Hamming window of length 51 ms with a 10 ms frame rate is used to extract the subband AM energies. The magnitudes were then compressed using the 15th root; please refer to Fig. 5 for details regarding the MMeDuSA feature extraction pipeline. 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. The MMeDuSA acoustic features $\widehat{\mathbf{x}}^{M}$, $\widehat{\mathbf{x}}^{M'}$ and $\widehat{\mathbf{x}}^{M''}$ used in our experiments are 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

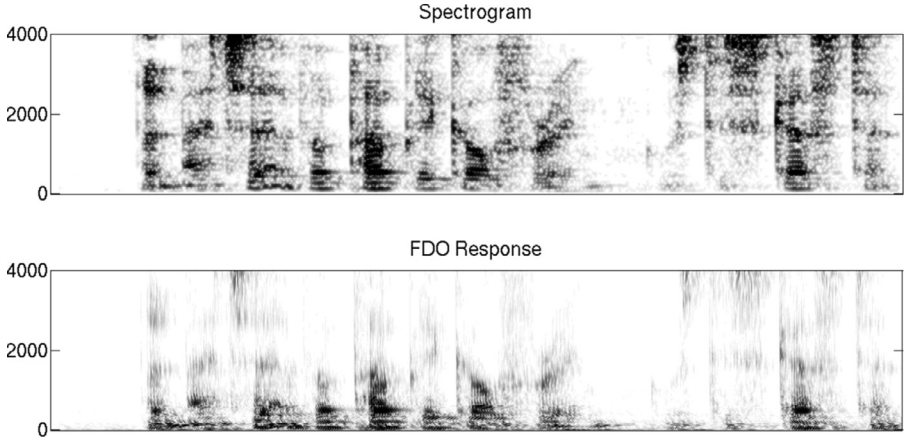

Fig. 4. Spectrogram vs. FDO response for speech corrupted with room reverberation.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

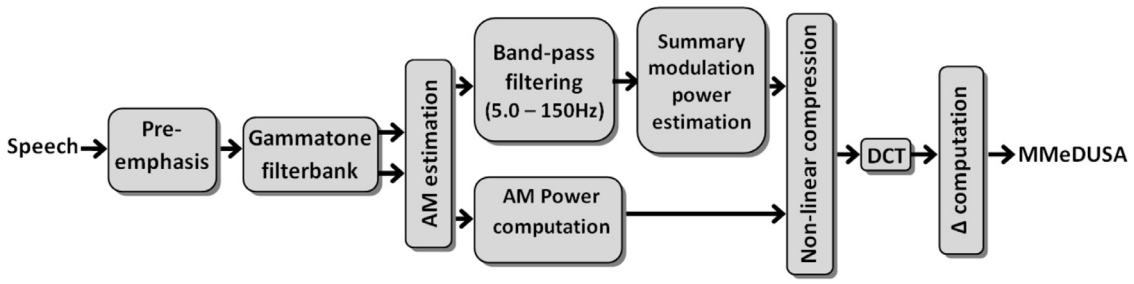

Fig. 5. Flow diagram of MMeDuSA feature extraction from speech.

concatenating them both (for details, please refer to Mitra et al., 2014). The Teager energy operator based AM extraction step filters out band-limited high-pass noise in every subband, reducing the overall noise level in the signal, contributing to MMeDuSA's robustness to noisy conditions (Mitra and Franco, 2015).

2.5. Acoustic modeling

Unlike our previous challenge submission system that uses the *enhanced* speech data, the acoustic models are trained by using the original *noisy* speech data, which avoids having to re-train a new system using enhanced speech for each enhancement method. The effectiveness of using noisy speech data is reported in various articles (Seltzer et al., 2013; Narayanan and Wang, 2014; Delcroix et al., 2015). Also, instead of training speech data recorded by a particular microphone, we use speech data recorded by all 6 microphones (i.e., 6 times more data). This aims to cover speech data variations over microphones and to make the amount of training data sufficient for the DNN-based acoustic modeling.

All the other configurations are similar to our previous challenge submission system (Hori et al., 2015), i.e., they use standard (11 frame context) DNN acoustic models followed by state-level sequence discriminative training, as prepared by the CHiME-3 baseline script. 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) (Gales, 1998), i.e., $\overline{\mathbf{x}}_t = \mathbf{A}_s\widehat{\mathbf{x}}_t + \mathbf{b}_s$, 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) (Mikolov et al., 2010) and a 5-gram language model with a modified Kneser–Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1996), which are trained using the WSJ0 text corpus. The RNNLM is an effective language model, which is represented as a recurrent neural network (RNN) 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 (Mikolov et al., 2010).

Given word sequence $W = w_1, w_2, \cdots, \mathrm{w_L}$, the occurrence probability of W is computed as a product of word probabilities given by the RNN

$$P(W) = \prod_{i=1}^{L} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{L} y_i[w_i], \tag{11}$$

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

where w_1^{i-1} denotes word sequence $w_1, \cdots, w_{i-1}$, y_i is the ith output vector of the RNN, and $y_i[w_i]$ is an element of y_i, which corresponds to the predicted probability of word w_i.

For $i = 1, \cdots, L$, output vector y_i is computed as follows:

$$\overline{w}_{i-1} = W_p \cdot \text{onehot}(w_{i-1}) + b_p \tag{12}$$

$$h_i = \sigma(W_{ih}\overline{w}_{i-1} + W_{hh}h_{i-1} + b_h) \tag{13}$$

$$y_i = \text{softmax}(W_{ho}h_i + b_o). \tag{14}$$

In Eq. (12), word w_{i-1} is projected to a low dimensional vector $\overline{w}_{i-1}$ using weight matrix W_p and bias vector b_p, where onehot(w) denotes a function that obtains a one-hot vector representation of word w. In Eq. (13), hidden activation vector h_i is obtained from $\overline{w}_{i-1}$ and previous hidden activation vector h_{i-1} using weight matrices W_{ih} and W_{hh} and bias vector b_h, where $\sigma(\cdot)$ is the sigmoid function. In Eq. (14), output vector y_i is obtained using weight matrix W_{ho}, bias vector b_o, and the softmax function. In the above computation, we assume that w_0 is a special word that indicates the beginning of the word sequence.

We also introduce LSTM RNN language models (Sundermeyer et al., 2012) to further improve the system performance. It is well known that the standard RNNs cannot hold hidden activation patterns for a long time because the activation pattern at a certain time is exponentially decaying according as iterative propagation through time, and it is difficult to train interdependence between distant events (Bengio et al., 1994). To solve this problem, the LSTM RNN has a memory cell in each hidden unit instead of a regular network unit. An LSTM cell can remember a value for an arbitrary length of time. It contains input, forget, and output gates that determine when the input is significant enough to remember, when it should continue to remember or forget the value, and when it should output the value. Accordingly, it is expected that the LSTM language models (LSTMLMs) can predict the next word more accurately than the standard RNNLMs by exploiting the longer contextual information. In the case of LSTMLMs, hidden activation vector h_i in Eq. (13) is computed using the LSTM activation function.

In the decoding phase, word lattices are first generated using the baseline language model for CHiME-3, which is the standard 5k WSJ trigram downsized with an entropy pruning technique (Stolcke, 2000). After that, N-best lists are generated from the lattices using the 5-gram model. Finally, the N-best lists are rescored using a linear combination of the 5-gram and RNNLM (or LSTMLM) probabilities in log domain, i.e.,

$$\log P(W) = \sum_{i=1}^{L} \{\lambda \log P_{rnn}(w_i|w_1^{i-1}) + (1-\lambda)\log P_{5gkn}(w_i|w_{i-4}^{i-1})\}, \tag{15}$$

where W is assumed to be each sentence hypothesis and λ denotes the interpolation weight. The best-rescored hypothesis is selected as the result of each single system, and the rescored N-best lists are also used for system combination.

2.7. System combination

In the proposed system, nine feature vector sequences are obtained for all pairs of beamforming and feature extraction methods, and they are separately processed by a WFST-based decoder to output word lattices. After rescoring with the RNN language models, nine lists of N-best hypotheses are obtained, which are then used for system combination.

System combination is a technique to improve recognition accuracy by combining different recognition hypotheses from different systems (Evermann and Woodland, 2000). First, the multiple hypotheses are combined by taking their union after reweighting each hypothesis with its posterior probability. After that, minimum Bayes risk (MBR) decoding is performed on the combined hypotheses (nine N-best lists in our system) using an algorithm in Xu et al. (2011). With this decoding, we can find the hypothesis with the minimum expected word error rate from among all the hypotheses obtained by the multiple systems. The decoded result is the final output of the system.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

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. Experiments

3.1. CHiME-3 task

The 3rd CHiME challenge consists of two types of data, real and simulated. The real data were recorded with a 6-channel tablet-mounted microphone array in four real noisy environments (on buses, in cafés, in pedestrian areas, and at street junctions) uttered by actual talkers. The simulated data were 6-channel 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 the 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. There is no linguistic overlapping between the train, development and test sets. The WSJ0 text corpus is also available to train language models, which consists of 37M words from 1.6M sentences.

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 (Povey et al., 2011). 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. To train DNN acoustic models for individual conditions, we used GMM-based state alignments obtained in each condition, and therefore did not use any cross alignment information over different conditions.

3.3. Beamforming and speech enhancement

We used the "Beamformit" beamforming toolkit for implementing weighted delay-and-sum (WDAS) beamforming (Anguera et al., 2007). 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.

We implemented an MVDR beamformer which does not require explicit calculation of delays as discussed in Section 2.2 (Souden et al., 2010). This beamforming requires a good estimation of the noise spatial covariance matrix, which we obtained using the masks predicted by the LSTM speech enhancement network. Speech

[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.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

enhancement network was trained using CHiME-3 training set and CHiME-3 development set was used as the validation data. The noisy signal's spatial covariance matrix was estimated from the whole utterance directly. 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 spatial covariance estimation and MVDR beamforming. The reference microphone was chosen as the one obtaining highest estimated posterior SNR, except microphone 2 which was never chosen.

We implemented a GEV beamformer as well. Both MVDR and GEV beamformers use the same estimated spatial covariance matrices. The differences between MVDR and GEV are explained in Section 2.2. We used all channels for GEV beamforming unlike MVDR beamforming. The reason for this ad-hoc choice, which could be revisited in the future, is to have more diversity in the results, since we perform ASR system combination and more diversity helps in system combination.

Table 2 compares several combinations of noisy and enhancement data in training and test phases. First, we observed that all multi-channel enhancement methods (WDAS, MVDR, and GEV) significantly improved the performance from (non-enhanced) noisy speech data. Also, by comparing 2-b and 2-c, we confirmed that the acoustic model trained with (non-enhanced) noisy speech data performed better than that trained with enhanced speech data, as we discussed in Section 2.5. Table 2 also shows that the performance of MVDR and GEV was better than WDAS, especially for the simulation test sets, but the difference of the performance of these for the real test sets were marginal. In the following experiments, we perform experiments with WDAS for further analysis since WDAS is used in the official DNN baseline version 2.[2] MVDR and GEV are used with WDAS in the system combination stage (Section 3.7).

In addition, we examined the deviation of WERs when using different random initializations for DNN acoustic model training in the WDAS condition (2-c). With 6 different random seeds, the mean and the standard deviation of WERs were 9.02 $\pm$ 0.09, 8.18 $\pm$ 0.03, 14.00 $\pm$ 0.21 and 14.81 $\pm$ 0.25 for sim-dev, real-dev, sim-test, and real-test sets, respectively. Since the deviations are relatively small and have no big impact on the results, we use DNN acoustic models trained from random initialization with a fixed random seed in all the experiments.

3.4. Effect of 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), an RNN language model (RNNLM), and two LSTM RNN language models (LSTMLMs) with different sizes of hidden layers are trained on the WSJ0 text corpus. The RNNLM is a class-based model (Mikolov et al., 2011a) with 200 word classes and 500 hidden units. The LSTMLMs are word-based models with 500 and 1000 hidden units, respectively. For RNNLMs, we use the RNNLM toolkit (Mikolov et al., 2011b), which is a standard toolkit to train RNNLMs. Since the toolkit runs only on a single CPU, a class-based model is a reasonable choice in terms of computational cost and recognition accuracy because it significantly reduces the CPU time for the softmax layer. For the LSTMLMs, we built a GPU-based implementation for training LSTM RNNs using Chainer deep learning framework (Tokui et al., 2015). In this case, we no longer need the class-based architecture, because the GPU implementation is fast enough to compute the softmax of word-based models. Moreover, we can increase the number of hidden units without a big increase in computation time due to the highly parallel GPU computation. When using an

Table 2
The effect of noisy and enhancement data for training and test. The results are obtained with DNN sMBR training for MFCC fMLLR features.

	Training	Test	Sim-dev	Real-dev	Sim-test	Real-test
2-a	noisy (ch5)	noisy (ch5)	11.28	12.67	13.73	23.66
2-b	enh. (WDAS)	enh. (WDAS)	9.68	8.41	16.16	16.17
2-c	noisy (ch5)	enh. (WDAS)	9.10	8.20	14.23	14.86
2-d	noisy (ch5)	enh. (MVDR)	**6.00**	**7.31**	**6.44**	14.81
2-e	noisy (ch5)	enh. (GEV)	6.94	7.62	7.17	**14.52**

[2] http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/results.html .

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

Table 3
Language model effect from the result of 2-c in Table 2.

	LM	# hidden units	Sim-dev	Real-dev	Sim-test	Real-test
2-c	3-gram	N/A	9.10	8.20	14.23	14.86
3-a	5-gram KN	N/A	7.82	6.88	12.29	13.02
3-b	RNNLM	500	6.50	5.46	10.33	10.86
3-c	LSTMLM	500	6.00	5.08	9.92	10.23
3-d	LSTMLM	1000	**5.57**	**4.47**	**9.17**	**9.57**

RNNLM or an LSTMLM, the word probability is computed as a linear combination of the 5-gram KN and the RNNLM (or LSTMLM) as described in Section 2.6, where the best combination weight is chosen using the development set.

Table 3 shows the result of re-scoring with the above language models. The 3-gram baseline result comes from Table 2-c, where noisy (ch5) data were used for acoustic model training and enhanced data (WDAS) were used for testing. With 5-gram KN, RNNLM and LSTMLM, the WER was steadily reduced, e.g., from 14.86% to 10.23% for the real-test set. The LSTMLM yielded a lower WER than the RNNLM in the case of 500 hidden units, and further reduced the WER from 10.23% to 9.57% when the number of hidden units increased to 1000. In the following experiments, we only show the WERs after re-scoring with the linear combination of 5-gram KN and LSTMLM with 1000 hidden units.

3.5. Acoustic model training with noisy multi-channel data

We investigated the use of noisy multi-channel data directly for training acoustic models. Compared to using only one channel or one enhanced signal, this approach increases the amount of training data by the number of channels, which may help cover various acoustic variations over microphones. Table 4 compares the cases of using only the 5th channel, the 5 channels excluding the backside microphone (2nd channel), and all 6 channels. The amount of training data is 18, 90, 108 h, respectively.

Table 4 clearly shows that the results using 5 and 6 channels outperformed that using a single channel by 0.5−2.5% absolutely. The results using 5 and 6 channels are almost comparable, but the average development-set WER of the 6-channel results is slightly better than that of the 5-channel results, and the following experiments thus use the acoustic model trained with 6 channels.

3.6. LSTM single-channel enhancement

We investigated the effect of LSTM single-channel enhancement after beamforming. Table 5 reports recognition results with/without LSTM single-channel enhancement for the test data, where acoustic models were trained with noisy multi-channel data and the test data were enhanced by WDAS beamforming.

Table 4
Increasing the amount of training data by using multi-channel signals from the result of 3-d in Table 3.

	Mic. ids	Hour	Sim-dev	Real-dev	Sim-test	Real-test
3-d	5	18	5.57	4.47	9.17	9.57
4-a	1,3,4,5,6	90	4.77	**3.93**	**6.52**	8.15
4-b	1,2,3,4,5,6	108	**4.54**	3.99	7.63	**7.73**

Table 5
The effect of LSTM single-channel enhancement for enhanced signals.

	Beamforming	LSTM	Sim-dev	Real-dev	Sim-test	Real-test
4-b	enh. (WDAS)	w/o	**4.54**	**3.99**	7.63	**7.73**
5-a	enh. (WDAS)	w/	6.61	6.13	**7.05**	7.74

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

LSTM enhancement after WDAS beamforming did not improve the results on the development sets, so we do not use it in our final system. Enhancement did not improve after MVDR beamforming as well in our earlier trials, but we do not report the results since we slightly changed our implementation of MVDR. The reason we think LSTM enhancement does not help in this case is because it causes distortion in the speech signals and this hurts the performance of the speech recognizer. It seems that beamforming already provides a good enough signal that can be used for speech recognition.

These results are in contrast with our earlier results on CHiME-2 (Weninger et al., 2015) where LSTM enhancement significantly improved recognition results. The nature of the CHiME-2 and CHiME-3 data is different in that CHiME-2 used living room noises containing speech-like elements, only two microphones were used and the mixing of speech and noise was simulated. However, CHiME-3 has real mixing with 6 microphones and noise environments which are arguably less speech-like. These factors may make a difference regarding whether machine-learning based front-end speech enhancement after beamforming improves ASR accuracy or not. In addition, we had a system which used stacked enhanced and noisy features in our earlier work on CHiME-3 (Hori et al., 2015). When one uses stacked or concatenated features from noisy and enhanced wave files, one needs to train a different model for each enhancement type which causes delay in experimentation. In this version of our system, we found out that training once only using noisy data and decoding with various types of enhanced data yielded good results. Hence we did not use stacked features in this work. We ended up only using LSTM speech enhancement for getting the statistics for MVDR and GEV beamforming. In the future, we plan to continue working on methods to use speech enhancement for speech recognition.

3.7. Noise-robust features and system combination

Table 6 reports recognition results with noise robust features extracted from different beamformed signals and their system combination. In addition to the standard MFCC, our proposed noise robust features (MMeDuSA and DOCC) were extracted from three beamformed signals of WDAS, MVDR and GEV, where none of the feature parameters were tuned or optimized to the development or test data, i.e. their default configurations were used. All the results are obtained with DNN-based recognition systems.

First, we compare the real-test results of MMeDuSA and DOCC features for the WDAS-beamformed signal. In this case, DOCC (8.08%) was better than MMeDuSA (8.52%), but the both robust features were slightly worse than MFCC (7.73%). On the other hand, for the MVDR-beamformed signals, the both DOCC (5.91%) and MMeDuSA (6.62%) outperformed better than MFCC (8.83%), where DOCC was slightly better than MMeDuSA as well. For the GEV-beamformed signal, we obtained a result similar to the MVDR result. The best combination of beamforming and feature extraction was MVDR and DOCC (6-d), which gave 5.91% WER.

Looking at the WERs in Table 6, there are no big differences between feature types when using WDAS beamformer while MFCC is worse than DOCC and MMeDuSA for real recording data (real-dev and real-test) when using MVDR or GEV beamformer. We think that MVDR and GEV beamformers reduce background noise more directly than WDAS beamformer, but they may be introducing some distortion into the noise floor especially in real recording data, and increase the channel mismatch between training and test data. Since both DOCC and MMeDuSA can cope with such channel mismatches effectively, a combination of MVDR (or GEV) and DOCC (or MMeDuSA) can

Table 6
Effect of robust features and system combination.

	Beamforming	Feature type	Sim-dev	Real-dev	Sim-test	Real-test
4-b	WDAS	MFCC	4.54	3.99	7.63	7.73
6-a	WDAS	DOCC	4.77	4.05	8.49	8.08
6-b	WDAS	MMeDuSA	5.08	4.19	8.98	8.52
6-c	MVDR	MFCC	**2.49**	3.73	**2.83**	8.83
6-d	MVDR	DOCC	2.66	**3.18**	3.10	**5.91**
6-e	MVDR	MMeDuSA	2.85	3.27	3.32	6.62
6-f	GEV	MFCC	3.06	4.01	3.48	8.94
6-g	GEV	DOCC	2.82	3.50	3.24	7.42
6-h	GEV	MMeDuSA	3.13	3.79	3.64	8.11
6-i	System combination		**2.37**	**2.50**	**2.66**	**5.05**

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

be a good choice for a single system. This could be the reason why we obtained the best result for the combination of MVDR and DOCC.

Finally, the outputs of different beamforming and feature extraction systems were combined. In all the systems, we used 5-gram KN and LSTM language models. The *N*-best lists rescored by the language model were combined into a single list and MBR decoding was performed for the list to obtain the minimum Bayes risk word sequence hypothesis.

The last row of Table 6 shows the results of combining all the nine systems. With the system combination, the WER is significantly reduced from those of the individual systems. Thus, the beamforming techniques and robust features we introduced had complementary properties that yielded substantial improvements by system combination. Our final system achieved 5.05% WER.

4. Discussion

The previous section showed the effects of noisy-data training, beamforming, LSTM language models, multi-channel data training, robust feature extraction, and system combination. In this section, we discuss the results of beamforming and robust feature extraction in more details. Furthermore, we compare the extended system described in this paper with our official CHiME-3 system submitted to the challenge. Specifically, we clarify which component contributed to the improvement in recognition accuracy.

4.1. Contribution of enhancement/feature-extraction methods

As shown in Table 6, we have already confirmed that combining systems with three beamformers and three feature extraction methods substantially reduced the recognition errors, which resulted in 5.05% WER. Our single best system that achieved 5.91% WER used the MVDR beamformer and the DOCC feature. Although the WER was further reduced by system combination, the contributions of the other beamformers and features to WER reduction remain unclear. Therefore, we show more detailed results in Table 7, which includes WERs in each recording environment for nine individual systems, a system combination of three systems within each beamformer, and a combination of all the systems.

In single systems with the WDAS beamformer (4-b, 6-a, 6-b), MFCC features outperformed robust features in most environments. This implies that the effect of the robust features was absorbed by the noisy-data-trained acoustic models, because both DOCC and MMeDuSA yield a certain error reduction when using enhanced-data-trained acoustic models in our official system (Hori et al., 2015). However, combining WDAS-based systems with MFCC, DOCC and MMeDuSA (7-a) reduced the WER substantially from 7.73% to 6.83%. This means that complementary properties are included in DOCC and MMeDuSA and that they are important for improving the recognition accuracy.

Table 7
Effect of beamforming and robust features for each recording environment for the real-test data.

	Beamforming	Feature type	BUS	CAF	PED	STR	Ave.
4-b	WDAS	MFCC	**10.60**	7.36	5.94	7.02	7.73
6-a	WDAS	DOCC	11.76	7.47	6.22	6.87	8.08
6-b	WDAS	MMeDuSA	12.25	7.53	7.03	7.27	8.52
6-c	MVDR	MFCC	18.86	6.09	4.32	6.07	8.83
6-d	MVDR	DOCC	10.62	5.04	**3.38**	**4.59**	**5.91**
6-e	MVDR	MMeDuSA	12.06	5.44	3.70	5.29	6.62
6-f	GEV	MFCC	18.73	5.72	4.65	6.65	8.94
6-g	GEV	DOCC	15.33	**4.78**	3.92	5.66	7.42
6-h	GEV	MMeDuSA	16.69	5.32	4.32	6.11	8.11
7-a	System comb.(WDAS)	4-b + 6-a + 6-b	9.95	6.00	5.29	6.11	6.83
7-b	System comb.(MVDR)	6-c + 6-d + 6-e	10.90	4.45	3.06	4.48	5.72
7-c	System comb.(GEV)	6-f + 6-g + 6-h	14.84	4.45	3.44	5.60	7.08
6-i:	System comb.	All	**8.69**	**4.05**	**3.01**	**4.46**	**5.05**

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

Table 8
Word error rate of each speaker for the real-test data. F01, F04, M03, and M04 indicate speaker Ids.

	Beamforming	Feature type	F01	F04	M03	M04	Ave.
4-b	WDAS	MFCC	8.58	10.09	5.77	6.48	7.73
6-d	MVDR	DOCC	6.73	7.51	3.23	6.18	5.91
6-i	System comb.	All	**5.70**	**6.28**	**2.91**	**5.16**	**5.05**

In contrast, in the case of the MVDR and GEV beamformers, the effect of robust features is substantial since DOCC and MMeDuSA outperform the MFCC features in all the environments. In cafés (CAF), pedestrian (PED), and street (STR) environments, the DOCC feature with the MVDR or GEV beamformer resulted in the lowest WERs. Only in the bus (BUS) environment, MFCC with the WDAS beamformer was the best.

In addition, combining MVDR-based systems (7-b) reduced the WER from 5.91% to 5.72%, and combining GEV-based systems (7-c) reduced the WER from 7.42% to 7.08%. These results also indicate that MFCC, DOCC, and MMeDuSA have properties that are complementary to each other. Finally, combining all the systems (6-i) resulted in the best performance. Accordingly, all the beamformers and robust features including MFCC are indispensable for achieving the final result.

We also investigated the WERs of each speaker. Table 8 shows the WERs of four speakers included in the real-test data, for which the results were obtained from WDAS-MFCC (4-b), MVDR-DOCC (6-d), and the all-combined (6-i) systems. Looking at the speaker-dependent WERs, ASR accuracy improvement by beamforming and robust features does not seem to depend on the speaker since the WERs are consistently reduced for each speaker.

4.2. Improvement from the official CHiME-3 system

Here, we summarize the differences between the presented system and our official CHiME-3 system. Fig. 6 shows WER reductions by the official system from the baseline, in which we applied WDAS and MVDR beamformers, LSTM-based single-channel enhancement, acoustic model retraining, robust features, fMLLR, 5-gram+RNNLM, and system combination, and our official result was 9.1% WER. We also plot WERs of the official DNN baseline version 2 (12.8%)[3] and our final system (5.05%). This final result is better than the best result (5.8% WER) in the CHiME-3 challenge (Yoshioka et al., 2015).

As shown in the previous sections, the error reduction from 9.1% to 5.05% was achieved by the following extensions:

(1) *Multi-channel noisy data training:* The official system basically employed acoustic models retrained with the enhanced data. As shown in Fig. 6, this retraining was effective to obtain a better performance (22.88%) than training with only clean speech (29.7%). Although this matched model worked well in the official system, multi-channel noisy data training further increased the robustness especially against the enhanced speech signals.
(2) *Large-scale LSTMLM vs. medium-sized RNNLM:* The RNNLM was quite effective for CHiME-3 task, where we obtained a 25% relative WER reduction (14.96% to 11.23%) as in Fig. 6. By introducing LSTMLM with 1000 hidden units, we further obtained a 12% relative error reduction (10.86% to 9.57%) as shown in Table 3.
(3) *Increase from four systems to nine systems:* Our official system included only four sub-systems, in which MFCC, DOCC and MMeDuSA features were extracted from WDAS-beamformed signal while only MFCC feature was extracted from MVDR-beamformed signal. In the extended system, MFCC, DOCC and MMeDuSA features were all extracted from both WDAS and MVDR plus GEV-beamformed signals, and thus their combinations resulted in the nine sub-systems. By increasing the number of sub-systems, we could include better sub-systems, one of which already achieved 5.91% WER using the MVDR beamformer and the DOCC feature. In addition, every sub-system had a better WER as a single system than the baseline version 2 using the WDAS

[3] The official DNN baseline version 2 was provided by the organizers after the CHiME-3 challenge, where the system is a part of our official CHiME-3 system, in which only the WDAS beamformer is used without robust features and system combination.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

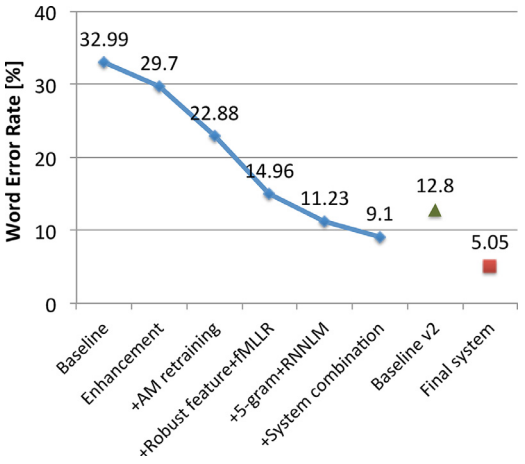

Fig. 6. Improvement from the official CHiME-3 system submitted to the challenge.

beamformer and the MFCC feature based on our official system. Consequently, the nine-system combination greatly outperformed the four-system combination of the official system.

4.3. Processing time in decoding phase

In this work, we focused on improving the recognition accuracy regardless of computational cost. But for reference, we show the processing time needed to complete each module in the decoding phase, which is summarized in Table 9.

Every processing time is represented in real-time factor (RTF) when we run the program as a single-thread process on Intel(R) Xeon(R) E5-2670 machines operating at 2.60 GHz, but we used E7-4830 machines at 2.13 GHz only for MVDR and GEV beamformers. Those two CPUs had a similar performance.

It is shown that it roughly takes $3.5 \sim 4 \times$ RTF to obtain a recognition result with a single system including beamforming, feature extraction, ASR decoding and LM re-scoring. In beamforming and feature extraction, one of the three methods is used for a single system. In ASR decoding, the 1st-pass GMM SI decoding obtains state alignments using speaker-independent (SI) GMM acoustic models, and estimates fMLLR transformation matrices for test speakers. The 2nd-pass GMM SD decoding obtains better alignments using the speaker-dependent (SD) fMLLR features, and updates the fMLLR matrices. Then, the DNN decoding is performed with the updated fMLLR features. In LM re-scoring, *N*-best re-scoring with either RNNLM or LSTMLM is performed following 5-gram lattice re-scoring.

To obtain the final result including system combination, the whole processing time becomes roughly 9 times larger than that for a single system. However, all the processes are performed offline and can be parallelized over multiple CPU cores by dividing the data set into small subsets. Accordingly, we did not have any serious computational problems when conducting the experiments.

Table 9
Processing time of each module in decoding phase. Each number corresponds to the elapsed time taken for the process, which is represented in real-time factor.

Processing module	Real-time factor					
Beamforming	WDAS	1.8	MVDR	1.5	GEV	1.3
Feature extraction	MFCC	3.5×10^{-3}	DOCC	3.8×10^{-3}	MMeDuSA	5.0×10^{-3}
ASR decoding	1st-pass GMM SI decoding				0.23	
	2nd-pass GMM SD decoding				0.49	
	fMLLR feature transformation				7.1×10^{-3}	
	DNN decoding				0.98	
LM re-scoring	5-gram KN				0.12	
	RNNLM	0.22		LSTMLM	0.71	
System combination	MBR decoding with nine *N*-best lists				0.05	

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

5. Conclusion

We presented the multi-microphone speech recognition system we submitted to the 3rd CHiME speech separation and recognition challenge (CHiME-3) and its extended system. 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 LSTM 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 5.05% WER for the real-test data, an 84.7% reduction relative to the baseline of 32.99% WER and a 44.5% reduction from our official CHiME-3 challenge result of 9.1% WER.

Acknowledgment

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

References

Anguera, X., Wooters, C., Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. 15, 2011–2022.
Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines. In: Proceedings of Automatic Speech Recognition and Understanding (ASRU).
Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. 27, 621–633.
Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing. Springer-Verlag, Berlin, Germany.
Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166.
Brandstein, M.S., Silverman, H.F., 1997. A robust method for speech signal time-delay estimation in reverberant rooms. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 375–378.
Chen, S.F., Goodman, J., 1996. An empirical study of smoothing techniques for language modeling. In: Proceedings of Association for Computational Linguistics (ACL), pp. 310–318.
Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., Nakatani, T., 2015. Strategies for distant speech recognitionin reverberant environments. EURASIP J. Adv. Signal Process. 2015, 1–15.
Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680.
Erdogan, H., Hershey, J.R., Le Roux, J., Watanabe, S., 2015a. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J., 2015b. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Evermann, G., Woodland, P.C., 2000. Posterior probability decoding, confidence estimation and system combination. In: Proceedings of NIST Speech Transcription Workshop.
Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12, 75–98.
Ghitza, O., 2001. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. 110, 1628–1640.
Heymann, J., Drude, L., Haeb-Umbach, R., 2016. Neural network based spectral mask estimation for acoustic beamforming. In: Proceedings of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP).
Hori, T., Chen, Z., Erdogan, H., Hershey, J.R., Roux, J., Mitra, V., Watanabe, S., 2015. The MERL/SRI system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding (ASRU).
Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The reverb challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–4.
Kneser, R., Ney, H., 1995. Improved backing-off for M-gram language modeling. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 181–184.
Mikolov, T., Deoras, A., Kombrink, S., Burget, L., Cernocky, J., 2011a. Empirical evaluation and combination of advanced language modeling techniques. In: Proceedings of Interspeech, pp. 605–608.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S., 2010. Recurrent neural network based language model. In: Proceedings of Interspeech, pp. 1045–1048.
Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Cernocky, J., 2011b. RNNLM − recurrent neural network language modeling toolkit. In: Proceedings of Automatic Speech Recognition and Understanding (ASRU) Demo, pp. 196–201.
Mitra, V., Franco, H., 2015. Time−frequency convolutional networks for robust speech recognition. In: Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 317–323.
Mitra, V., Franco, H., Graciarena, M., 2013. Damped oscillator cepstral coefficients for robust speech recognition. In: Proceedings of Interspeech, pp. 886–890.
Mitra, V., Franco, H., Graciarena, M., Mandal, A., 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In: Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4117–4120.
Mitra, V., Franco, H., Graciarena, M., Vergyri, D., 2014. Medium duration modulation cepstral feature for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Mitra, V., Van Hout, J., Wang, W., Graciarena, M., McLaren, M., Franco, H., Vergyri, D., 2015. Improving robustness against reverberation for automatic speech recognition. In: Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 525–532.
Narayanan, A., Wang, D., 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7092–7096. Vancouver, Canada
Narayanan, A., Wang, D., 2014. Investigation of speech separation as a front-end for noise robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 826–835.
Neiman, A.B., Dierkes, K., Lindner, B., Han, L., Shilnikov, A.L., 2011. Spontaneous voltage oscillations and response dynamics of a Hodgkin−Huxley type model of sensory hair cells. J. Math. Neurosci. 1, 11. https://link.springer.com/article/10.1186/2190-8567-1-11.
Potamianos, A., Maragos, P., 2001. Time−frequency distributions for automatic speech recognition. IEEE Trans. Speech Audio Process. 9, 196–200.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlček, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. In: Proceedings of ASRU.
Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7398–7402.
Souden, M., Benesty, J., Affes, S., 2010. On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Lang. Process. 18, 260–276.
Stolcke, A., 2000. Entropy-based pruning of backoff language models. arXiv preprint cs.CL/0006025.
Sundermeyer, M., Schlüter, R., Ney, H., 2012. LSTM neural networks for language modeling. In: Proceedings of Interspeech.
Teager, H., 1980. Some observations on oral air flow during phonation. IEEE Trans. Acoust. Speech Signal Process. 28, 599–601.
Tokui, S., Oono, K., Hido, S., Clayton, J., 2015. Chainer: a next-generation open source framework for deep learning. In: Proceedings of Workshop on Machine Learning Systems (LearningSys) in the Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS).
Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second CHiME speech separation and recognition challenge: datasets, tasks and baselines. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 126–130.
Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 1849–1858.
Warsitz, E., Haeb-Umbach, M.R., 2007. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio Speech Lang. Processs. 15, 1529–1539 http://dx.doi.org/10.1109/TASL.2007.898454 .
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA).
Weninger, F., Le Roux, J., Hershey, J.R., Schuller, B., 2014a. Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proceedings of GlobalSIP Symposium on Machine Learning Applications in Speech Processing.
Weninger, F.J., Hershey, J.R., Le Roux, J., Schuller, B., 2014b. Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proceedings of GlobalSIP Machine Learning Applications in Speech Processing Symposium.
Xu, H., Povey, D., Mangu, L., Zhu, J., 2011. Minimum Bayes risk decoding and system combination based on a recursion for edit distance. Comput. Speech Lang. 25, 802–828.
Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W.J., Espi, M., Higuchi, T., et al., 2015. The NTT chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 436–443.

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013