The THU-SPMI CHiME-4 system : Lightweight design with advanced multi-channel processing, feature enhancement, and language modeling

Hongyu Xiang, Bin Wang, Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, Beijing, China

Contact: ozj@tsinghua.edu.cn

Abstract

In this paper, we describe our lightweight system designed for CHiME-4. For multi-channel processing, we experiment with a bundle of beamforming methods, including minimum variance distortionless response (MVDR), parameterized multi-channel wiener filter (PMWF), generalized sidelobe canceller (GSC), spectral mask estimation (ME), and compare these techniques with the same back-end. Combining MVDR's distortionless and reliable estimation of the steering vector by ME is found to be most effective. We propose to applying histogram equalization (HEQ) to compensate for the residual noise in the MVDR beamformed speech. We apply the recently introduced trans-dimensional random field (TRF) language model and confirm its superiority in rescoring. In combination these techniques are surprisingly effective in the CHiME-4 task, achieving 6.55% word error rate (WER) for the real evaluation data while keeping low system complexity. Applying multi-channel training further reduces the WER to 5.81%.

1. Background

The performance of automatic speech recognition (ASR) has been significantly improved in recent years. However, robust ASR in everyday environments remains a challenge. Research efforts can be roughly decomposed into developing more powerful front-ends (e.g. microphone array signal processing, feature enhancement) and back-ends (e.g. acoustic modeling, language modeling).

For front-ends, some widely used beamforming techniques are minimum variance distortionless response (MVDR) [1], parameterized multi-channel wiener filter (PMWF) [2], generalized sidelobe canceller (GSC) [3], and weighted delay and sum (WDAS) [4]. Beamforming filters could be designed based on different criteria, representing different trade-offs between distortion and noise reduction. For example, MVDR minimizes the output energy subject to no distortion in the desired direction. It is known that the effectiveness of beamformers heavily relies on the estimation of the spatial correlation matrix, the steering vector or time delays, which are usually difficult to estimate in practice. Researchers have explored to estimate the spatial correlation matrix using time-frequency masks, which are obtained either by complex Gaussian mixture models (GMMs) [5] or advanced neural networks [6]. For back-ends, neural network based acoustic models have become the state-of-the-art in speech recognition [7]. Neural network based language models (LMs) have also begun to surpass the classic n-gram LMs [8, 9].

The CHiME-4 challenge [10] revisits the CHiME-3 data [11], i.e., WSJ0 corpus sentences spoken by talkers situated in challenging noisy environments recorded via a 6-microphone tablet device. The aim is to provide a new benchmark task for evaluating and promoting far-field speech recognition in everyday environments.

The CHiME-3 baseline uses MVDR beamformer with diagonal loading [12] as the front-end. The back-end is based on the Kaldi toolkit [13] and consists of a GMM-HMM using fMLLR transformed features to provide senone state alignment and a DNN using fbank features. The DNN is trained using sequence discriminative training with state-level mimimum Bayes risk (sMBR) criterion. After CHiME-3, an upgraded Kaldi-based baseline script was made available for CHiME-4 task, which further incorporates multichannel enhancement using WDAS based BeamformIt [4], fMLLR features for the DNN stage, interpolated 5-gram LM and RNN LM for rescoring. The CHiME-4 baseline produces an average WER of 11.57% for the real evaluation data (obtained by our own run).

This paper presents the THU-SPMI system designed for CHiME-4. For time constraint, we only submit results for 6-channel track, although the techniques developed in this submission could be applied to 1-channel track and 2-channel track.

2. Contributions

The goal of this study is to create a lightweight advanced system for far-field multi-channel speech recognition, which can achieve a good trade-off between system complexity and system performance, and is practically useful. To this end, we do not rely on feature fusion (e.g. extracting multiple types of features) or hypothesis fusion (e.g. training multiple systems and doing ROVER), though these are provably beneficial. We are selective to integrate front-end and back-end techniques and stay simple. Specifically, we identify the following three key techniques which enable us to significantly improve over the baseline while keeping low system complexity.

1) For multi-channel processing, after experiments with a bundle of beamforming methods, the MVDR beamformer with the steering vector being estimated by time-frequency masks, as proposed in [5], is found to be most effective. Our contribution is extensive comparisons between various beamformers with the same back-end.

2) Note that the MVDR beamformer reduces the noise under the distortionless constraint of any signal from the source direction. There are few artifacts in the beamformed speech, but there still exists considerable residual noise. We propose to apply histogram equalization (HEQ) technique for feature normalization, which is originally studied for single-channel feature enhancement [14]. WERs are found to be significantly reduced by using HEQ after the MVDR beamformer, which is an important empirical finding from this study.

3) Recently, we have shown in previous work [15, 16] with open source code [17] that a new trans-dimensional random

This work is supported by NSFC grant 61473168.

Table 1: Average WER (%) for the CHiME-4 baseline system obtained by our own run.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	GMM	12.90	14.35	21.55	21.09
	DNN sMBR	8.12	9.37	14.84	14.38
	KN5+RNN	5.89	6.97	11.57	10.66

field (TRF) LM achieves superior performance. In the CHiME-4 task, we confirm that interpolated TRF and LSTM performs better than using LSTM alone, and produces significantly better rescoring performance than interploted 5-gram-KN and RNN provided in the CHiME-4 baseline. This represents an advance of the state-of-the-art of language model rescoring.

3. Experimental evaluation

3.1. System overview

Basically, the proposed system follows the pipeline of the CHiME-4 baseline, and is strengthened with the three techniques which are highlighted before and will be introduced and evaluated in the following. Table 1 shows the WER results for the baseline, which are obtained by our own run. Starting from the baseline, we incrementally investigate the relative contribution of each technique from front-end to back-end, and show that in combination they are surprisingly effective for the CHiME-4 task, ultimately achieving 6.55% WER for the real evaluation data. Applying multi-channel training further reduces the WER to 5.81%, which was conducted after the CHiME-4 submission.

3.2. Beamforming

We experiment with a bundle of beamforming methods, which will be briefly introduced below. The experimental results are shown in Table 2, with the same back-end.

3.2.1. Signal model

In the time domain, most beamforming methods assume the following signal model:

$$x_i(t) = s(t) * h_i(t) + n_i(t) \tag{1}$$

where $x_i(t)$ is the i-th microphone signal, $s(t)$ is the source signal, $h_i(t)$ is the impulse response from the source to the i-th microphone, and $n_i(t)$ is the additive noise.

In frequency domain, we have

$$\mathbf{X}(t,\omega) = S(t,\omega)\mathbf{d}(\omega) + \mathbf{N}(t,\omega) = \mathbf{G}(t,\omega) + \mathbf{N}(t,\omega) \tag{2}$$

where $S(t,\omega)$, $\mathbf{X}(t,\omega)$, $\mathbf{N}(t,\omega)$ are the STFT coefficients of the desired source signal, the microphone signal vector and the noise signal vector respectively. $\mathbf{d}$ denotes the steering vector. For convenience, we omit t and ω in the following description.

3.2.2. Weighted delay and sum (WDAS)

WDAS simply aligns different channels in time and sums them together as follows:

$$y(t) = \sum_i w_i x_i(t - \tau_i) \tag{3}$$

where τ_i is the time delay from the source to the i-th micophone, w_i is the weight. The CHiME-4 baseline BeamformIt [4] is based on WDAS, where time delays are estimated by use of generalized cross correlation with phase transform (GCC-PHAT) [18] and two-step Viterbi postprocessing.

3.2.3. Minimum variance distortionless response (MVDR)

MVDR is designed to minimize the output energy subject to no distortion in the desired direction:

$$\min_{\mathbf{W}} \mathbf{E}||\mathbf{W^H X}||^2 \; s.t. \mathbf{W^H d} = 1 \tag{4}$$

which has the well-known closed-form solution

$$\mathbf{W} = \frac{\mathbf{\Phi_{NN}^{-1} d}}{\mathbf{d^H \Phi_{NN}^{-1} d}} \tag{5}$$

where $\mathbf{\Phi_{NN}}$ is the noise correlation matrix, $\mathbf{H}$ denotes conjugate transposition.

The performance of MVDR relies heavily on the estimation of the noise correlation matrix $\mathbf{\Phi_{NN}}$ and the steering vector $\mathbf{d}$. The steering vector could be estimated by time delays τ_i, $\mathbf{d} = [e^{-j\omega\tau_1}, e^{-j\omega\tau_2}, ...]$, as did in the CHiME-3 baseline. A recent method studied in [5], denoted as MVDR-EV, is to obtain the steering vector from the principal eigenvector of the estimated spatial corelation matrix of clean signal $\mathbf{\Phi_{GG}} = \mathbf{\Phi_{XX}} - \mathbf{\Phi_{NN}}$, and use complex GMM based spectral mask estimation (ME) method to esimate $\mathbf{\Phi_{NN}}$.

Allowing the desired direction gain to be the reference component of $\mathbf{d}$, we obtain MVDR with relative transfer function (MVDR-RTF) [2]. Assuming the first channel to be the reference channel, MVDR-RTF can be expressed as

$$\min_{\mathbf{W}} \mathbf{E}||\mathbf{W^H X}||^2 \; s.t. \mathbf{W^H d} = d_1 \tag{6}$$

where d_1 is the first component of $\mathbf{d}$. The solution is

$$\mathbf{W} = \frac{\mathbf{\Phi_{NN}^{-1} \Phi_{GG}}}{tr(\mathbf{\Phi_{NN}^{-1} \Phi_{GG}})} \mathbf{u_1} \tag{7}$$

where $\mathbf{u_1}$ is vector $[1, 0, 0, ..., 0]$.

3.2.4. Generalized sidelobe canceller (GSC)

Generalized sidelobe canceller is composed of three parts: a fixed beamformer, a block matrix and a noise canceller. The fixed beamfomer and the block matrix are normally fixed filters. Using $\mathbf{b}$ and $\mathbf{z}$ to represent the output of the fixed beamformer and block matrix respectively, GSC aims at finding the filter minimizing the output of the noise canceller,

$$\min_{\mathbf{R}} ||\mathbf{b} - \mathbf{R^H z}||^2 \tag{8}$$

where $\mathbf{R}$ is the noise canceller filter and is normally implemented by an adaptive filter.

3.2.5. Parameterized multi-channel Wiener filter (PMWF)

PMWF explicitly expresses the trade-off between noise reduction and distortion. The PMWF filter is defined by

$$\min_{\mathbf{W}} \mathbf{E}(||\mathbf{W^H X} - G_1||^2 + \beta||\mathbf{W^H N}||^2) \tag{9}$$

where G_1 is the first element of $\mathbf{G}$, assuming the first microphone to be the reference microphone, and β is the parameter. The first term $\mathbf{E}(||\mathbf{W^H X} - G_1||^2$ represents distortion and the second term $\mathbf{E}||\mathbf{W^H N}||^2$ represents noise reduction. The solution is

$$\mathbf{W} = (\mathbf{\Phi_{XX}} + \beta\mathbf{\Phi_{NN}})^{-1}\mathbf{\Phi_{GG}}\mathbf{u_1} \tag{10}$$

Table 2: Average WER (%) of different beamformers with the CHiME-4 baseline back-end but without RNN.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	WDAS	8.19	9.40	15.59	15.61
	MVDR	14.31	5.97	25.89	6.99
	GSC	10.99	15.77	19.79	24.17
	GSC+WDAS	9.46	11.73	16.61	19.00
	PMWF	10.82	9.90	19.58	14.18
	ME+direct	8.75	6.98	15.05	7.74
	ME+PMWF	8.87	6.51	15.52	7.33
	ME+MVDR-EV	8.04	6.07	13.59	7.32
	ME+MVDR-RTF	11.07	6.90	18.99	8.53

3.2.6. Results and Discussions

In Table 2, WDAS denotes the BeamformIt in the CHiME-4 baseline [4]; MVDR denotes the one released at CHiME-3 [11]; GSC is a standard one with a fixed beamformer and a simple fixed block matrix. GSC+WDAS means using WDAS to relpace the beamformer block of GSC. When applying MVDR, MVDR-RTF and PMWF, noise correlation matrix is estimated using a limited context immediately before the utterance as in the CHiME-3 baseline. After complex GMM based mask estimation (ME), we apply the estimated masks directly to separate the source ("ME+direct") or to estimate the spatial correlation matrices which are fed to different beamformers (the last three rows in Table 2). We use all 6 channels with energy based microphone failure detection, except in the case of running WDAS where we do not use channel 2.

Several points can be drawn from Table 2. (1) CHiME-3 baseline MVDR performs best on the simulated data but worst on the real data. Presumably this is because that the steering vector estimation in the CHiME-3 baseline MVDR is similiar to the generation of the simulated data and is not matched to the real data. The CHiME-4 baseline WDAS (BeamformIt) performs well. (2) Different beamforming methods pursue tradeoff between reducing noise and avoiding source distortion from different perspectives. MVDR-RTF and PMWF contain distortion even if with perfect estimation of spatial correlation matrix. The MVDR-EV beamformer is attractive since it explicitly enforces distortionless in the desired source direction. (3) The MVDR-EV beamformer relies on the estimation of the spatial correlation matrices of clean and noise signals, which in turn are used to estimate the steering vector $\mathbf{d}$ and the beamformer coefficients $\mathbf{W}$. Complex GMM based spectral mask estimation is found to be superior for this purpose. (4) Replacing the fixed beamformer for GSC is not able to improve the performance of GSC. The block matrix and noise canceller may play a more important role than the fixed beamformer for GSC.

In summary, among those beamforming techniques show in Table 2, ME works well for its ability to reduce noise; WDAS (BeamformIt) performs well for its robustness; ME+MVDR-EV is found to be most effective, which combines MVDR's distortionless and reliable estimate of the steering vector by ME. Noise reduction, distortionless and robustness should be considered together when designing a beamformer.

The Table 2 results are obtained by training back-end GMMs and DNNs over the enhanced speech. Results in all later Tables (starting from Table 3) are obtained by 1) using cross-correlation based mic failure detection, 2) training back-end acoustic models over only channel 5 but testing over the enhanced speech from ME+MVDR-EV.

Table 3: Average WER (%) for the ME+MVDR-EV enhanced speech with the CHiME-4 baseline back-end.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	GMM	10.89	10.45	16.42	12.10
	DNN sMBR	7.20	6.44	11.10	8.02
	KN5+RNN	5.16	4.70	8.21	5.79

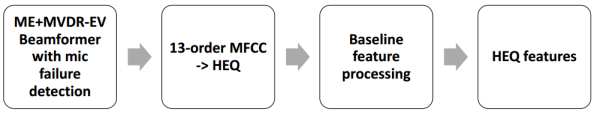


Figure 1: HEQ feature enhancement flow chart in testing.

3.3. Microphone failure detection

For the 6-ch speech recognition, there exists microphone failure, which hurts the recognition performance. Energy based microphone failure detection does not work well, so we propose to use segmental cross-correlation to detect microphone failure.

Microphone failure is mainly caused by microphones not working or touched by the speaker, thus there may have small or large energies. Considering that cross-correlation is influenced by speech magnitudes, we first normalize the 6-ch signals to have equal energies for each channel. Then we calculate the summed segmental maximum cross-correlation:

$$\mathrm{corr}[i,m] = \sum_{j,j\neq i} \max_{n} \mathrm{corr}[i,j,m,n] \tag{11}$$

where $\mathrm{corr}[i,j,m,n]$ denotes the cross-correlation between the m-th segment from ch-i and the m-th segment from ch-j with n-point shift. The $\mathrm{corr}[i,m]$ is further scaled by the median as follows:

$$\mathrm{scorr}[i,m] = \frac{\mathrm{corr}[i,m]}{\underset{i}{median}\ \mathrm{corr}[i,m]} \tag{12}$$

When $\mathrm{scorr}[i,m]$ is smaller than the threshold α, the ch-i's m-th segment is considered as a failure segment. If one channel contains more than β failure segments, this channel is thrown away. In our experiments, a segment is of 128ms duration, α is set to be 0.6 and β is set to be 2.

3.4. Histogram equalization (HEQ)

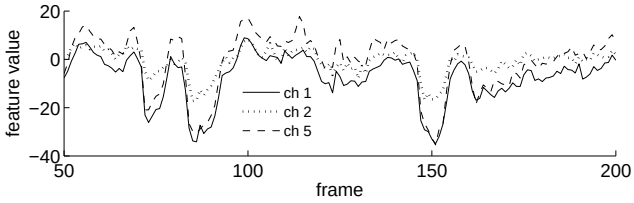
The baseline acoustic features are 13-order MFCCs. HEQ is to warp each component of the cepstral vector over a specified time interval to match the standard Gaussian. While HEQ is applied in sentence level in [14], HEQ over sliding 3-second windows performs better in our experiments. After HEQ, other feature transformations are applied as in the CHiME-4 baseline. In training, HEQ is applied to the MFCCs of channel 5[1]. In testing, HEQ is applied to the enhanced speech, as shown in the flow chart in Figure 1.

It is worthwhile to compare the well-known CMVN and the HEQ. While both are for feature normalization, HEQ is potentially more effective to compensate for additive noise due to the nonlinear nature of the distortion caused by additive noise in the cepstral domain. Comparing Table 3 and 4, it is clear to see the benefit of applying HEQ to compensate for the residual noise in the MVDR beamformed speech, espeically for the real data.

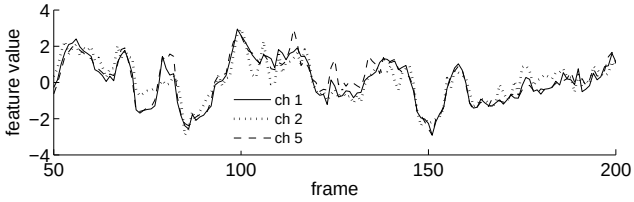
[1] In multi-channel training, HEQ is applied to all six channels.

Table 4: Average WER (%) for the stack-HEQ features with the CHiME-4 baseline back-end.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	GMM	10.39	10.37	13.53	12.01
	DNN sMBR	6.73	6.12	9.95	8.19
	KN5+RNN	4.64	4.22	7.15	5.51



(a) MFCC (tr05 real noisy)



(b) HEQ (tr05 real noisy)

Figure 2: Effect of HEQ over the second component of MFCC feature vectors.

For illustration purpose, Figure 2 plots the second component of the MFCC feature vectors and the corresponding HEQ features for utterance 011_011C0201_PED in real training set. Three channels (ch 1, ch 5 and ch 6) are plot separately.

HEQ reduces variations in noisy signals but may lose details. We stack two types of fMLLR features with and without HEQ as the input of the DNN for information fusion (called stack-HEQ).

3.5. Trans-dimensional random field (TRF) LM

In addition to the 5-gram LM and RNN LM provided in the baseline, a TRF LM is trained on the official training corpus with 200 word classes and the features "w+c+ws+cs+wsh+csh" [15]. "w"/"c" denotes the word/class n-gram up to order 4 and "ws"/"cs" denotes the word/class skipping n-gram up to order 4. "wsh"/"csh" denotes the higher-order long-skipping features. The definition of feature types is shown in Table 1 of [15].

Here is a brief introduction to TRF LMs. Denote by $x^l = (x_1, \ldots, x_l)$ a sentence (i.e., word sequence) of length l ranging from 1 to m. Each element of x^l corresponds to a single word. D denotes the whole training corpus and D_l denotes the collection of length l in the training corpus. n_l denotes the size of D_l and $n = \sum_{l=1}^{m} n_l$.

As defined in [15], a trans-dimensional random field model represents the joint probability of the pair (l, x^l) as

$$p(l, x^l; \lambda) = \frac{n_l/n}{Z_l(\lambda)} e^{\lambda^T f(x^l)}, \tag{13}$$

where n_l/n is the empirical probability of length l. $f(x^l) = (f_1(x^l), \ldots f_d(x^l))^T$ is the feature vector, which is usually defined to be position-independent and length-independent, e.g. the n-grams. d is the dimension of the feature vector $f(x)$. λ is the corresponding parameter vector of $f(x^l)$. $Z_l(\lambda) = \sum_{x^l} e^{\lambda^T f(x^l)}$ is the normalization constant of length l. By making explicit the role of length in model definition, it is clear that the model is a mixture of random fields on sentences of different lengths (namely on subspaces of different dimensions), and hence will be called a trans-dimensional random field (TRF).

In the joint SA training algorithm [15], another form of mixture distribution is defined as follows:

$$p(l, x^l; \lambda, \zeta) = \frac{n_l/n}{Z_1(\lambda) e^{\zeta_l}} e^{\lambda^T f(x^l)} \tag{14}$$

where $\zeta = \{\zeta_1, \ldots, \zeta_m\}$ with $\zeta_1 = 0$ and ζ_l is the hypothesized value of the log ratio of $Z_l(\lambda)$ with respect to $Z_1(\lambda)$, namely $\log \frac{Z_l(\lambda)}{Z_1(\lambda)}$. $Z_1(\lambda)$ is chosen as the reference value and can be calculated exactly. An important observation is that if and only if ζ were equal to the true log ratios, then the marginal probability of length l under distribution equals to n_l/n. This property is then used to construct the augmented SA algorithm, which jointly estimates the model parameters λ and normalization constants ζ.

TRF LMs have the potential to integrate a richer set of features, and as shown in [15], outperform the traditional 4-gram LM significantly with the relative WER reduction 9.1%. Moreover TRF LMs also achieve slightly better WER results than RNN LMs, but with much faster speed in computing sentence probabilities.

In this experiment, the RNN LM is trained using the CHiME-4 baseline script with 300 hidden units. The LSTM LM is trained using the open source toolkit provided by [19] with 2 hidden layers and 500 hidden units of each layer. 10 epoch iterations are performed before early stop and no dropout is used. Following the challenge instructions, we tune the LM weight and interpolation weight over the whole development set including all noisy environments and data types. The experiment scripts can be found in [17]. As shown in Table 5, TRF alone performs as good as RNN; TRF+RNN further reduces the WER from KN5+RNN; TRF+LSTM performs even better.

Table 5: Average WER (%) for different language models.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	KN5	5.57	5.11	8.25	6.42
	RNN	5.24	4.82	7.92	5.91
	TRF	5.09	4.56	7.92	6.04
	LSTM	5.35	4.20	7.08	5.28
	KN5+RNN	4.64	4.22	7.15	5.51
	KN5+LSTM	4.68	3.74	6.79	5.15
	TRF+RNN	4.48	4.06	6.96	5.26
	TRF+LSTM	4.58	3.78	6.55	4.95

Table 6: WER (%) comparison w/o multi-channel training (enhanced speech with HEQ and TRF+LSTM back-end).

Track	System	Dev		Test	
		real	simu	real	simu
6ch	trained on only ch 5	4.58	3.78	6.55	4.95
	multi-channel	4.32	3.47	5.81	4.41

3.6. Multi-channel training

After the CHiME-4 submission, we perform multi-channel training as a straightforward way to expose the acoustic model to larger training data, as did in [20], and obtain further significant improvement, as shown in Table 6.

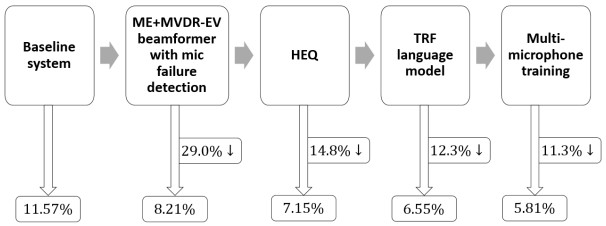


Figure 3: WERs on the real evaluation data, showing the relative contribution of each technique.

Table 7: Average WER (%) for the CHiME-4 baseline front-end (BeamformIt) with our submitted back-end (stack-HEQ).

Track	System	Dev		Test	
		real	simu	real	simu
6ch	GMM	12.56	14.37	18.82	19.84
	DNN sMBR	7.75	8.74	13.84	13.57
	KN5+RNN	5.46	6.29	10.35	9.97
	TRF+LSTM	5.23	5.66	9.36	9.16

Table 8: WER (%) per environment for the submitted system w/o TRF LM.

Track	Envir.	Dev		Test	
		real	simu	real	simu
6ch without TRF (KN5+RNN)	BUS	5.80	4.13	10.36	4.09
	CAF	3.78	4.90	6.13	5.12
	PED	3.85	3.63	5.08	5.60
	STR	5.12	4.22	7.04	7.23
6ch with TRF (TRF+LSTM)	BUS	5.68	4.56	9.67	4.74
	CAF	3.98	4.06	5.60	4.22
	PED	3.78	3.14	4.17	4.65
	STR	4.90	3.36	6.76	6.18

4. Summary

In this paper, we build a lightweight advanced system for CHiME-4 far-field multi-channel speech recognition challenge, with three key techniques. After experiments with a bundle of beamforming methods, the MVDR beamformer with the steering vector being estimated by time-frequency masks is found to be most effective. HEQ is successfully applied to compensate for the residual noise in the MVDR beamformed speech. Interpolated TRF+LSTM LMs perform significantly better than the baseline KN5+RNN LMs and are also superior to the state-of-the-art interpolated KN5+LSTM LMs in language model rescoring. In combination these techniques are surprisingly effective, achieving 6.55% WER for the real evaluation data while keeping low system complexity. Applying multi-channel training further reduces the WER to 5.81%.

Figure 3 shows how the system performance is incrementally improved over the CHiME-4 baseline with the introduced techniques from front-end to back-end. Following the challenge instructions, Table 7 shows the results of the CHiME-4 baseline front-end (BeamformIt) with our submitted back-end (stack-HEQ, TRF LM, without multi-channel training); Table 8 shows the WER per environment for our submitted system.

5. References

[1] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[2] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[3] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[4] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, 2010.

[9] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling." in *Interspeech*, 2012, pp. 194–197.

[10] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, to appear.

[11] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

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

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

[14] A. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[15] B. Wang, Z. Ou, and Z. Tan, "Trans-dimensional random fields for language modeling," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[16] B. Wang, Z. Ou, Y. He, and A. Kawamura, "Model interpolation with trans-dimensional random field language models for speech recognition," *arXiv preprint arXiv:1603.09170*, 2016.

[17] "https://github.com/wbengine/spmilm."

[18] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997.

[19] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[20] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.