THE I2R SYSTEM FOR CHIME-4 CHALLENGE

Tran Huy Dat, Ng Wen Zheng Terence, Sunil Sivadas, Luong Trung Tuan, Tran Anh Dung,

Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632

ABSTRACT

This paper reports developments and evaluation results of I2R system for CHiME-4 challenge which addresses distant speech recognition on tablet device in challenging noisy environments. It features three tracks of 6-channel; 2-channel; and 1-channel data, respectively. Our developments are more focused on the algorithms with potentials in real-time implementation. In front-end processing, time-domain weighted delay-and-sum beamforming (WDAS) was implemented with following specific processing compared to the provided baseline processing [1]: (1) channel SNR and coherence measurements were used to calculate the beamforming weighted coefficients; (2) slow updating of the beamforming weights with 2-second windows; (3) a modified single channel speech enhancement was applied on top of output beamforming enabling further reduction of the background noise while keep controlling the introduced distortion. In the back-end processing, two new components were applied compared to the provided baseline: (1) LSTM language model for re-scoring; and (2) Semi-supervised DNN adaptation for each individual speaker in test. In evaluations, we stay with unique acoustic models for all the task and apply the processing on test data only. Consistent improvements were obtained across all three tasks. The submitted results for the real test set were 5.00%, 8.32%, and 11.19% for the 6-channel, 2-channel, and 1-channel tasks, respectively.

1. BACKGROUND

The industrial applications of speech recognition has been moving from closed talk microphones to daily real life scenarios thanks to booming developments in robotic and artificial intelligence (AI) areas. The task, however, is remained challenging due to the problems of attenuation, noise, distortion, and reverberation. Following the success of the CHiME-3 challenge which attracted many international teams to participate, CHiME-4 revisits the CHiME-3 data, i.e., utterances recorded via a 6-microphone tablet device in challenging noisy environments. The difficulty is increased by reducing the number of microphones. CHiME-4 features three tracks depending on the number of microphones available for testing: 6-channel track; 2-channel track; and 1-channel track. Excepting the 6-channel task, the channels are randomly chosen from the pool so that no specific geometrical prior information is given to the samples. The audio was recorded under real acoustic mixing conditions, i.e. talkers speaking in challenging noisy environments, including four varied noise settings: caf, street junction, public transport and pedestrian area. We participated in both three tasks and our focus is the approaches which are suitable for real-time implementations. In the front-end, the weighted delay-and-sum beamforming (WDAS) was implemented with a specific way to determine the weighted coefficients, using both coherence [1] and SNR estimations [2]. A post-processing filter is applied on top of WDAS output and that was modified from a previous speech enhancement development [3]. The modification is made to reduce the distortion level from speech enhancement and was found useful for ASR task. The same enhancement filter is applied on noisy speech in the 1-channel task. The back-end acoustic modelling follows a typical Kaldi recipe [4] and unique DNN acoustic model is applied for all the tasks [5]. In the decoding stages, LSTM LM [6] for rescoring is applied and semisupervied DNN adaptation [7] is applied on individual speaker data. Consistent improvements from baseline were obtained cross all three tasks. The major contributions come from beamforming, LSTM LM re-scoring and semisupervied DNN adaptation and additional improvements were provided by post-processing enhancement and its two-stage implementation. The submitted results for the real test set were 5.00%, 8.32%, and 11.19% for the 6-channel, 2-channel, and 1-channel tasks, respectively. These results significantly outperformed the baseline results of 11.51%, 16.58%, and 23.70% on the same datasets. The advantages of our system is that it is applicable for universal situations of environments and can be translated into real-time. We also evaluated the data-driven BLSTM trained masking GEV beamforming [8], proposed by Paderborn University (Germany), with our back-end processing on the 6-channel data. Although the masking GEV outperformed our front-end it requires extra matching data to train putting a question on its performance in an totally unknown and mismatch conditions. Further studies are necessary to prove its practical value.

2. SYSTEM DESCRIPTIONS

The block diagram of our system is illustrated in Figure 1. The highlighted yellow are the important modules which are different from the baseline method.

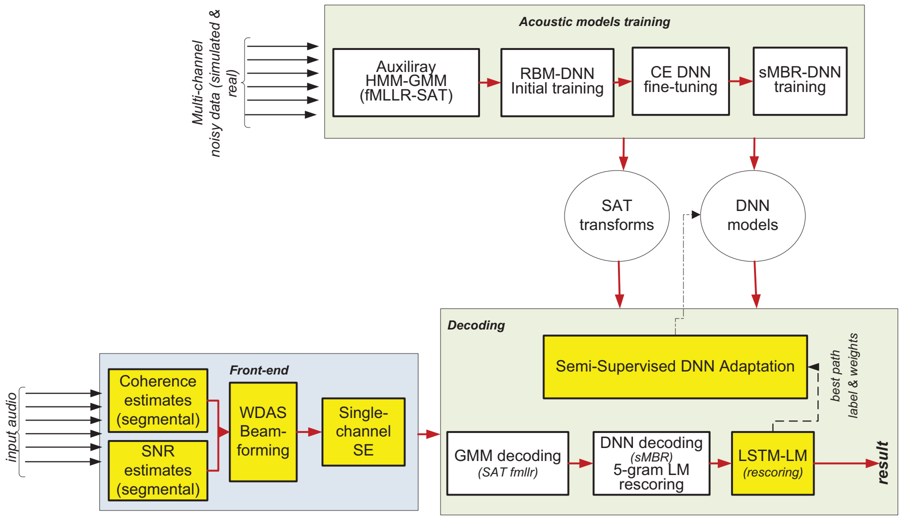
Multi-channel noisy data (simulated & real)
Acoustic models training
Auxiliray HMM-GMM (fMLLR-SAT)
RBM-DNN Initial training
CE DNN fine-tuning
sMBR-DNN training
SAT transforms
DNN models
Decoding
Semi-Supervised DNN Adaptation
best path label & weights
input audio
Front-end
Coherence estimates (segmental)
SNR estimates (segmental)
WDAS Beam-forming
Single-channel SE
GMM decoding (SAT fmllr)
DNN decoding (sMBR) 5-gram LM rescoring
LSTM-LM (rescoring)
result

Fig. 1. Overview of our CHiME-4 ASR system.

2.1. Front-end processing

Our front-end processing includes two stages of weighted-delayed-and-sum beamforming (WDAS) and a parametrized single channel speech enhancement enabling optimization of performances on the test data.

2.1.1. *Beamforming*

Time-domain weighted delay-and-sum (WDAS) method is applied in the beamforming step.

1. The microphone signals are first alighted using time difference of arrival (TDOA) which are estimated through GCC-PHAT.

2. The reference channel is initialized as the channel with the highest estimated SNR from channels and then iteratively tracking to the lowest negative TDOAs until they turn positive. Note that since the SNR estimated from channel number 2 is consistently bad, we have excluded this channel from our beamforming processing.

3. The weighted coefficients are calculated in two different ways before getting averaged: (1) using channel coherence measurements; (2) using SNR estimation.

$$w_i = \alpha_C \frac{CHR_i}{\sum\limits_{j=1}^{N} CHR_j} + \alpha_S \frac{SNR_i}{\sum\limits_{j=1}^{N} SNR_j}, \qquad (1)$$

   where the coherence measurements are calculated from pair-wise cross-correlation coefficients [1], noted as

$$CHR_i = \sum_{j\neq i}^{N} c_{ij}. \qquad (2)$$

   The SNR in each channel is estimated and updated by 2 second segments using the algorithm in [2]. α_C and α_S denote the weighting regularization coefficients between coherence and SNR measurements. Particularly, we set both of them equal to 0.5.

4. Slower updating of WDAS weights, compared to provided baseline BeamformIt front-end [1] is implemented using longer segments of 2 seconds

2.1.2. *Post-processing filter*

The advanatge of time-domain WDAS beamforming is that it produces very low distortions in the output signal. However, the method is less effective in removing background noise, particularly under low SNR conditions. Hence, post-processing filter is introduced to partially solve the problem. In this work, we applied the spectral estimation speech enhancement method introduced in a previous work [3]. This method estimates the speech spectral amplitude using Maximum A Posterior (MAP) criteria using generalized gamma distribution modelling of speech. While the method is effective in removing the background noise, it introduces distortions which is harmful to ASR systems. To control the distortion level, a simple modification has been applied and found to be effective in applying this method for ASR under severe noise conditions. It is done by introducing a rational power order to the original gain filter

$$\mathbf{G} \rightarrow \mathbf{G}^{\alpha}, \tag{3}$$

where the original gain filter is

$$\mathbf{G} = \frac{\hat{\mathbf{S}}}{\mathbf{X}}, \tag{4}$$

with the MAP spectral amplitude estimation noted by [3]

$$\hat{\mathbf{S}}{=}\mathbf{argmax_S}\left[\mathbf{p}\left(\mathbf{S},\mathbf{X}\right)\right] \tag{5}$$

The distortion controlling parameter α is chosen between $0 < \alpha < 1$. As closer to original $\alpha = 1$, the post-processing filter provides more noise removal but also more distortions. A trade-off in middle way near to $\alpha = 0.5$ seems always able to boost the ASR performances. Particularly, $\alpha = 0.5$ is used in our experiments for CHiME-4 data.

2.2. Back-end processing

2.2.1. *Data augmentation*

The 6-channel official training data, including both simulated and real noisy recordings provided by the challenge organizers, was used in the training [9].

2.2.2. *Acoustic modelling*

The acoustic modelling is carried out using standard Kaldi recipe [4]. The processing includes MFCC feature extraction followed by auxiliary HMM-GMM which provides speaker adaptive transforms (SAT) and the initial alignments. The DNN training is started with RBM initialization followed by two rounds of 4-iterations cross-entropy fine-tuning runs. The DNN training is finally carried out to deliver the acoustic models using the sMBR optimization [5].

2.2.3. *Language modelling*

Default 3-grams LM was used in the decoding followed by a re-scoring by provided 5-grams. Additional LSTM LM [6] was trained with provided text extracted from WSJ corpus and being used in the final re-scoring stage.

2.2.4. *Decoding with semi-supervised DNN adaptation*

In the decoding stages, the enhanced signals from front-end processing were used to input to the ASR system. It first passes to the HMM-GMM decoder to get the SAT-fMLLR transforms. Then the transformed features are used in the first pass of speaker independent DNN decoding using the default 3-gram LM followed by a 5-gram LM rescoring. From here, two important modifications were made, compared to the baseline method. First, instead of using RNN-LM re-scoring, we adopt more advanced LSTM LM described above. Secondly, semi-supervised adaptation is utilised, on each individual speaker data [8] using the best path state sequence and confidence measures, decoded from testing data, as the label and weightings, respectively for additional iterations of DNN fine-tuning. Five rounds of adaptations has been applied to maximize the WER reduction though it normally converges after just two rounds of adaptations.

Table 1. Average WER (%) for the tested single systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	I2R-fb-2	6.08	7.33	**11.19**	10.87
	I2R-fb	6.14	7.42	11.25	11.34
	Noisy-I2Rb	6.15	7.60	13.05	12.89
	Baseline	11.57	12.98	23.70	20.84
2ch	I2R-fb-2	4.32	5.10	**8.32**	7.57
	I2R-fb	4.35	5.33	8.43	7.70
	BeamformIt-I2Rb	4.76	6.62	9.37	8.48
	Baseline	8.23	9.50	16.58	15.33
6ch	MaskBF-I2Rb	2.70	2.16	3.94	2.90
	I2R-fb-2	3.18	3.39	**5.00**	4.97
	I2R-fb	3.25	3.48	5.08	5.00
	BeamformIt-I2Rb	6.35	6.14	6.44	6.06
	Baseline	5.76	6.77	11.51	10.90

3. EXPERIMENTAL EVALUATIONS

This section reports the results achieved by your system. Following methods have been evaluated and compared for both 1-channel, 2-channel and 6-channel tasks, respectively.

1. **Baseline** refers to the use of provided BeamformIt front-end and also provided decoding script.

2. **Noisy-I2Rb** refers to the use of original noisy audio and our developed decoding script. This is applied for single channel task only.

3. **I2R-fb** refers to single system using our proposed front-end and back-end processing, illustrated in Fig. 1.

4. **I2R-fb-2** refers to our improved version combined two different enhancement setting ($\alpha = 0.5$ and $\alpha = 0.25$).

5. **MaskBF-I2Rb** refers to the BLSTM trained masking GEV beamforming front-end provided by Paderborn University (Germany) [9] with our back-end processing

3.1. Overall results

Table 1 reports the experimental evaluation results on both four data sets from development and testing phases. We can see that consistent and significant improvements were obtained across all the datasets and tracks, from both back-end and front-end components. Our best system (I2R-fb-2)

Table 2. WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	8.26	5.56	17.20	7.51
	CAF	6.46	9.99	11.82	13.69
	PED	3.64	5.58	7.70	10.27
	STR	5.94	8.22	8.05	12.03
2ch	BUS	5.65	4.14	12.60	5.64
	CAF	4.59	6.71	8.21	9.02
	PED	2.73	3.91	4.07	4.89
	STR	2.85	3.82	4.78	6.24
6ch	BUS	4.82	2.74	6.56	3.46
	CAF	3.01	4.16	4.58	5.30
	PED	2.04	2.85	4.07	4.89
	STR	2.85	3.82	4.78	6.24

achieved approximately 12%, 8%, and 7% absolute WER reductions for the real test sets in 1-channel, 2-channel and 6-channel tracks, respectively. The improvements were seen consistently over datasets. The real test set is the most challenging set but the results are closing up on the 6-channel data.

3.2. Back-end contributions

It can be seen that, our system achieved consistent improvements cross all the datasets. Most significant improvements come from our back-end processing which approximately 10%, 7% and 5% absolute accuracy gains when moving from baseline to BeamformIt-I2Rb system. Among the back-end processing components, LSTM LM re-scoring and Semi-supervised DNN adaptation contributes the most.

3.2.1. Data augmentation

The multi-condition training using data augmentation has proven to be very effective for the noisy ASR tasks. In our experiments, we noticed nearly 2% additional improvement compared to baseline training script just by using both 6-channel noisy data instead of single noisy in original script. While it seems redundant in speech content, it may add some more noise variation into the training which helps in delivering better models. Another explanation is adding more data may help in DNN convergence which naturally requires sufficient training data. This may have happened in this case because the size of data is significantly enlarged using 6-channel data. But our effort to further improve the training by adding more simulated data to the training was not successfully.

3.2.2. LSTM language model re-scoring

LSTM seems exclusively suitable for language modelling, as it could extract temporal dependency from text data while overcome fundamental vanish gradient problem in RNN training hence deliver better prediction of text contents. Consistent improvements of $2-3\%$ WER reductions compared to 5-grams LM and $1-2\%$ of the same compared to RNN LM were seen in our experiments, respectively.

3.2.3. Semi-supervised DNN adaptations

Semi-supervised DNN adaptation has repeated its great contributions in our experiments with consistent improvements from $2-4\%$ absolute WER reductions in both 1-channel, 2-channel and 6-channel tracks, respectively. Although the default 5-round adaptation was applied, in most of cases, the best results were converged after 1-2 steps.

3.3. Front-end contributions

Compared to provided BeamformIt baseline which stands as a very good baseline method, our front-end processing achieved consistent $1-2\%$ absolute WER reductions for both tracks of 1-channel, 2-channel, and 6-channel, respectively.

3.3.1. Speech enhancement

For the 1-channel tracks, the contribution of improvements was fully made by the introduced speech enhancement. Nearly 2% gain in WER reduction was obtained. Note that as the original speech enhancement did not improve the WER, the idea of gain modification to control the distortion has shown to be a practical solution enabling applications of speech enhancement methods in ASR. Although, a simplest way of introducing a rational power order is applied in this work, more sophisticated algorithms to address the introduced idea could be more useful.

For the 2-channel and 6-channel tracks, as the beamforming already enhances the input signals, effect is post-processing speech enhancement is less significant. Nevertheless, consistent improvements of $0.3-0.4\%$ were seen on top of beamforming method.

The post-processing speech enhancement module also provides possibility for system combination in front-end level while keeping acoustic models unchanged. That is more practical than fusion of totally different front-end and back-end systems, often seen in the literature. In our experiments, simply combining two enhancement in lattice improved the performances of the ASR system. Further studies in this direction are suggested.

3.3.2. Beamforming

Our beamforming method which had been developed and applied in our previous works [5] is similar to the BeamformIt as the time-domain WDAS is applied in both cases. However, the way to calculate beamforming weights are different: BeamfromIt uses only cross-correlation coefficients while we use estimated SNR measurements on top of coherence measurements. The SNR estimation is also used in our approach for the channel selection. Our method uses slower updating windows. Finally, our algorithm is totally real-time while the BeamformIt requires batch processing. In CHiME-4 datasets, our beamforming achieved about $0.6 - 1\%$ improvement in absolute WER reduction for 2-channel, and 6-channel tracks, respectively.

We also compared our front-end method to the BLSTM trained masking GEV beamforming provided by Paderborn University (Germany)[8]. This method uses a parallel noisy/clean training data to train a BLSTM network to get the time-frequency mask before applying it into GEV beamforming which is a spatial filter in frequency domain. The masking-GEV BF achieved great results by other participants and also got the best result in our experiments when combining with our back-end processing. It achieved amazing 3.75% WER on real test set with our back-end and is superior to our front-end. However, this method requires training data which is matching to testing in CHiME-4 and this is unknown how it would perform in totally unknown environments. Further investigations are required to confirm its practical value.

3.4. Performances over noise conditions

Breakdown of the best performed system on real test set, per each environment condition is shown in Table 2. We can see that, excepting the bus conditions, the results from each track are quite clustered over four datasets. That means that the simulation could be used to predict and improve the developments for the real conditions. That is a very good finding for the industrial developments of far-field noisy ASR applications. For the bus condition, our system underperformed in the real test set compared to the rest of conditions. Note that the same things were not observed on the masking GEV method which deliver similar results for all the conditions. Further analyses should be carried out to find out the reasons of that.

4. CONCLUSIONS

This paper reports developments and evaluation results of I2R system for CHiME-4 challenge. We achieved consistent improvements compared to provided baseline across both tracks and datasets, in both front-end and back-end processing. More significant improvement achieved in back-end processing with LSTM language modelling for re-scoring and semi-supervised DNN adaptation. Consistent improvements were also obtained in front-end processing with coherence and SNR joint analytic based WDAS beamforming and distortion-controlled speech enhancement as a post-processing filter. The proposed front-end is a real-time processing method.

5. REFERENCES

[1] Xavier Anguera, Chuck Wooters, and Javier Hernando, Acoustic beamforming for speaker diarization of meetings, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 7, pp. 20112023, 2007.

[2] Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura, On-line gaussian mixture modeling in the log-power domain for signal-tonoise ratio estimation and speech enhancement, Speech Communication, vol. 48-1, pp. 15151527, 2006.

[3] Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura, Gamma modeling of speech power and its on-line estimation for statistical speech enhancement, IEICE Transactions on Information and Systems, vol. E89D(3), pp. 10401049, 2006.

[4] Daniel Povey at el., The kaldi speech recognition toolkit, in Proceedings of IEEE workshop on automatic speech recognition and understanding (ASRU) 2011, IEEE.

[5] Jonathan W.D. and H.D. Tran, Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: i2rs system description for the aspire challenge, in Proceedings of IEEE workshop on automatic speech recognition and understanding (ASRU) 2015, IEEE, 2015.

[6] Wojciech Zaremba, Ilya Sutskever,and Oriol Vinyals Recurrent neural network regularization, CoRR, vol. abs/1409.2329, 2014.

[7] Mirko Hannemann Karel Vesely and Lukas Burget, Semisupervised training of deep neural networks, in Proceedings of IEEE workshop on automatic speech recognition and understanding (ASRU). 2011, IEEE, 2013.

[8] Jahn Heymann, Lukas Drude, Reinhold Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming", Proceedings of ICASSP 2016, IEEE, 2016.

[9] Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer, An analysis of environment, microphone and data simulation mismatches in robust speech recognition, Computer Speech and Language, 2016, 2016.