CRIM's Speech Recognition System for the 4th CHiME Challenge

Md Jahangir Alam, Vishwa Gupta, Patrick Kenny

[1] Computer Research Institute of Montreal (CRIM), Montreal, Canada
{jahangir.alam, vishwa.gupta, patrick.kenny}@crim.ca

Abstract

This paper describes CRIM's contribution to the 4-th CHiME speech separation and recognition challenge. We took part in all the three tracks of the CHiME-4 challenge. Since the focus of this challenge was to address the more difficult 1 channel and 2 channel tasks, we focussed on algorithms that will have the largest impact on these two tasks. We focussed on increasing the training data and on using proven robust features from previous challenges so that they can favorably impact the word error rates (WER) for 1 channel and 2 channel tasks. We enhanced the training data by using the audio from all the microphones (i.e., microphones 1-6) instead of just microphone 5. We also added beamformed data from mic 1, 3-6. We band-limited the above training data to 4 kHz bandwidth and added these to the original training set, thereby doubling the training data. We tried many different robust feature parameters to see which ones actually gave lower WER than the Mel-frequency cepstral coefficients. In all our sub-systems we used the baseline language model and the backend provided by the organizers. Three different robust features actually gave lower WER for the 1 channel task. Combining the recognition outputs of 6 or 7 different features gave the optimal reduction in WER for the 1 channel, 2 channel and 6 channel tasks. Among all the features used in this task the **R**egularized **M**VDR **C**epstral **C**oefficients (RMCC) features performed the best.

Index Terms: 4th CHiME challenge, speech recognition, robust features, RMCC, ROVER, DNN.

1. Introduction

Automatic speech recognition is a key component in hands-free man-machine interaction. State-of-the-art speech recognition systems are based on statistical acoustic models which are trained in a clean and controlled environment. In recent years the use of deep neural network acoustic model and large amount of training data has helped to improve the performance of automatic speech recognition significantly. In many applications, speech recognition systems are deployed in real world scenarios (e.g. cafe, bus station, street, and pedestrian area) where the speech signal is severely distorted by background noise and reverberation. Consequently, the performance of speech recognition systems trained on clean data degrades severely in noisy and reverberant environments because of the mismatch between the training and the test conditions. Therefore, robust speech recognition in real world scenarios has attracted increasing attention in ASR research and development. This attention is due to the widespread use of mobile devices with speech enabled personal assistants.

The fourth edition of CHiME (CHiME-4) challenge, designed to be close to a real world application, provides a common framework for the evaluation and comparison of various approaches for the noise robustness of speech recognition system. Although CHiME-4 challenge revisits the corpora originally collected for CHiME-3, the level of difficulty has been increased by imposing constraint on the number of microphones available for testing. Depending on the number of microphones available for testing CHiME-4 offers three tracks: 1 channel, 2 channel and 6 channel tracks. CHiME-4 corpus is comprised of Wall Street Journal corpus sentences spoken by speakers situated in challenging noisy environments (such as bus, street junction, cafe, and pedestrian area) recorded using a 6-channel tablet based microphone array [1]. A Kaldi-based [2] baseline speech recognizer is provided by the organizers which uses sequence trained deep neural network (DNN) acoustic models and language model (LM) rescoring based on a linear combination of 5-gram LM and RNNLM [3].

In this work we present CRIM's system designed for CHiME-4 challenge tasks and report evaluation results. We took part in all the three tracks of the 4-th CHiME challenge: 1 channel (1ch), 2 channel (2ch), and 6 channel (6ch) tracks. In our contribution we mainly focussed on the robust features extraction and combination of systems based on different frontends using ROVER. In order to reduce the word error rate (WER), we tried many robust features that have performed better in other evaluations of noisy corpus such as the REVERB challenge [4] / AURORA-4 corpus [5], and also features that showed good performance in a speaker recognition task. In addition to the conventional Mel-frequency cepstral coefficients (MFCC) features, we tried the following robust features for speech recognition for CHiME-4 challenge tasks:

- ✓ The regularized MVDR spectrum-based cepstral coefficients (RMCC) [6, 7].
- ✓ Gabor filter-bank feature (GBFB) [8].
- ✓ The ETSI - advanced front end (ETSI-AFE) [9].
- ✓ Infinite impulse response – constant Q transform (IIR-CQT) [10] - based cepstral coefficients (ICQC).
- ✓ The IIR-CQT–based log filterbank (ICQF) features [11].

For the 2ch and 6ch tasks, all our systems employ beamformed speech signals supplied by a weighted delay-and-sum beamforming technique. In two systems we apply beamforming after enhancing the signals using weighted prediction error (WPE)-based dereverberation [12] and Consistent Wiener filtering (CWF)-based audio source separation [13] techniques. We denote those two systems as the WPE-MFCC, CWF-MFCC, respectively. The only difference between the CWF-MFCC and CWF2-MFCC systems is in the noise spectrum estimation while performing audio source separation using CWF. CWF-MFCC uses a MMSE-based noise spectrum estimator whereas CWF2-

MFCC utilizes regional statistics-based noise spectrum estimator. The motivation behind using the ICQC and ICQF features is that these features provide good performance in speaker verification and spoofing detection tasks [11]. As mentioned in the abstract, using all the training data (channels 1-6) gave significantly lower WER than using just the 5 channels (1, 3-6). Also, band-limiting the training data and adding it to the training data [14] had only a small effect on the WER of the development set. Among all the frontends considered for the CHiME-4 tasks, the **R**egularized **M**VDR **C**epstral **C**oefficients (RMCC) features yielded the lowest WER. Combining results of 6 or 7 different feature-based systems with ROVER (Recognizer Output Voting Error Reduction) [15] gave the lowest WER for all the tasks.

2. CHiME-4 Tasks

The CHiME-4 challenge revisits the CHiME-3 corpora with increased level of difficulty by imposing a constraint on the number of microphones available for testing. CHiME-4 tasks consist of three tracks: 1 channel (1ch), 2 channel (2ch) and 6 channel (6ch) tracks. The 6ch track is based on a subset of the channels of CHiME-3 data. CHiME-4 challenge is designed to be close to the real world applications having real acoustic mix, i.e., speakers speaking in challenging noisy environments such as bus, street junction, cafe, and pedestrian area.

3. Overview of CRIM System

In this section we provide an overview of the CRIM system as presented in fig. 1, for the 1ch, 2ch and 6ch tasks of CHiME-4 challenge. Our main contributions include:

i. We band-limit the training data to 4 kHz bandwidth and include these to the original training set, thereby doubling the training data.
ii. For multi-channel tasks, as a pre-processing step, we apply beamforming to enhance the target speech. This step is same as the baseline system provided by the organizer. In two of our systems we additionally enhance the signals using weighted prediction error (WPE)-based dereverberation [12] and Consistent Wiener filtering (CWF)-based audio source separation [13] techniques and then apply beamforming.
iii. We extract robust features by employing RMCC feature extractor.
iv. We combine different robust-feature-based systems using ROVER.

3.1. Pre-processing

As a pre-processing for 2ch and 6ch tasks we enhance the target speech by using a weighted delay and sum beamforming technique. After selecting a reference signal based on the pair-wise cross-correlation, the time delay between a microphone and the reference is estimated using the GCC-PHAT algorithm. Weights for the *m*-th microphone are estimated from the cross-correlations of the *m*-th microphone with other microphones. Finally beamformed signal $\hat{y}(t)$ is obtained using the estimated delays and microphone weights as

$$\hat{y}(t) = \sum_{m=1}^{M} w_m y_m (t - \tau_m), \quad (1)$$

where *m* is the microphone index, *M* is the total number of microphones, w_m and τ_m are the estimated weights and time delays, respectively and $y_m(t)$ is the *m*-th microphone signal.

Among our systems, one system utilizes weighted prediction error (**WPE**)-based dereverberation to enhance the 1ch, 2ch and 6ch signals. The WPE does dereverberation using a linear time invariant filter and produces *M*-channel outputs from *M*-channel inputs. From the M-channel dereverberated signals ($M > 1$) beamformed signal is obtained using a weighted delay and sum beamforming technique.
Another one of our systems employs a consistent Wiener filtering (CWF)-based audio source separation to enhance the signals. The CWF refers to a time-frequency masking which takes into account the consistency of spectrograms for the computation of true optimal solution to the Wiener filtering problem. In this framework, to estimate noise spectrum we used either a MMSE-based noise spectrum estimator or a regional statistics-based noise spectrum estimator.

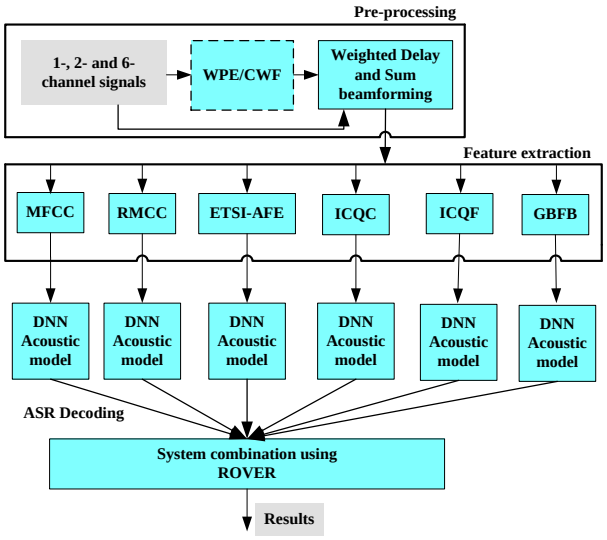

Fig. 1. Schematic diagram of CRIM's system for the 4-th CHiME challenge. Beamforming is applied to the multi-channel signals only. Only two of our systems use weighted prediction error (WPE)-based dereverberation and consistent Wiener filtering (CWF)-based audio source separation (shown with dotted rectangle).

3.2. Extraction of robust features

In this section we describe the robust features used for CHiME-4 challenge tasks.

3.2.1. The ETSI-advanced front-end (ETSI-AFE)

The ETSI-advanced frontend (**ETSI-AFE**) [9] employs a two-stage Wiener filter and blind equalization technique, which is based on the comparison to a flat spectrum and the application of the LMS (Least Mean Squares) algorithm, for improving robustness of ASR systems against additive noise distortions and channel effects.

3.2.2. *Gabor filterbank features (GBFB)*

The Gabor filterbank (**GBFB)** features [8] are extracted from the log Mel-filterbank spectrum using auditory motivated spectral-temporal 2D filters. These filters were tuned to specific spectro-temporal modulation patterns that occur in speech signals and motivated by the fact that some neurons in the primary auditory cortex of mammals were found to be tuned to very similar spectro-temporal modulation patterns.

3.2.3. *IIR-Constant Q transform-based features*

The ICQC and ICQF feature representations are derived from the infinite impulse response - constant Q transform by recursively filtering the multi-resolution fast Fourier transform of the signal. We refer to these features by the acronym ICQC for Infinite impulse response Constant Q transform Cepstrum and ICQF for Infinite impulse response Constant Q transform log filterbank features. In order to compute ICQC features we first estimate the IIR-CQT spectra by designing an infinite impulse response (IIR) filterbank that has constant Q behavior. The location of the poles of the IIR filterbank vary for each frequency bin along the real axis in order to make wider window width for lower frequency and narrower for higher frequency. Then a linear time variant (LTV) IIR filter is devised based on the poles of the filterbank. The filter is applied in the forward direction followed by reverse filtering to obtain the IIR-CQT spectrum [10]. The ICQC features, as shown in fig. 3, are obtained by applying discrete cosine transform to the estimated spectrum following logarithmic compression [11].

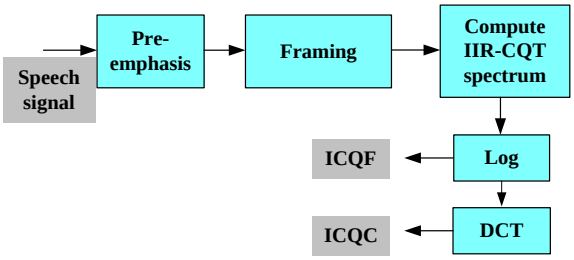

Fig. 2. The ICQC and ICQF feature extraction from the IIR-CQT spectra. Here Q = 13 was chosen.

3.2.4. *Regularized MVDR cepstral coefficients*

The conventional Mel-frequency cepstral coefficients (MFCC) are usually computed from a DFT-based spectral estimate. When regularized MVDR (RMVDR) spectrum estimator is used to compute the cepstral features instead of the DFT-based spectrum estimator we denote the features as the regularized MVDR cepstral coefficients (RMCC). RMCC was introduced in [6, 7] and evaluated on the AURORA-4 corpus under both clean and multistyle training modes. Here we use RMCC to extract robust features for the CHiMe-4 challenge tasks.

The first step in computing RMCC is to estimate RMVDR spectra. Similar to the MVDR spectrum estimator, the *p*-th order regularized MVDR spectral estimate can be parametrically written as

$$Y_{rmvdr}(f)=\frac{1}{\sum_{k=-p}^{k=p}\mu_r(k)e^{-i2\pi fk}}, \qquad (2)$$

where the parameter $\mu_r(k)$ of the regularized MVDR method can be obtained from a non-iterative computation using the regularized LP (RLP) coefficients a_q^r and the prediction error variance σ_e^r as:

$$\mu_r(k)=\begin{cases}\dfrac{1}{\sigma_e^r}\sum_{q=0}^{p-k}(p+1-k-2q)\,a_q^r a_{q+k}^{r*}, & \text{for } k\geq 0\\ \mu_r^*(-k), & \text{for } k<0.\end{cases} \qquad (3)$$

The regularized predictor coefficients a_q^r are computed by adding a penalty measure $\psi(a^u)$, which is a function of the unknown predictor coefficients a^u, to the objective function of the LP method and therefore, minimizing the modified objective function of the following form [1, 2]

$$\sum_n\left(y(n)+\sum_{q=1}^{p}a_q y(n-q)\right)^2+\lambda\psi(a^u), \qquad (4)$$

Where $s(n)$ is the current speech sample, regularization constant $\lambda>0$ controls the smoothness of the all-pole spectral envelope. RLP method helps to penalize the rapid changes in all-pole spectral envelope and therefore, produces a smooth spectral estimate keeping the formant positions unaffected [6]. The optimal values chosen for the model order p and regularization constant λ are 100 & 10^{-7}, respectively [6, 7].

After estimating RMVDR spectrum, RMCC features are obtained by integrating Mel-scale filterbank and taking discrete cosine transform following logarithmic compression. Mean and variance normalization is used for feature normalization.

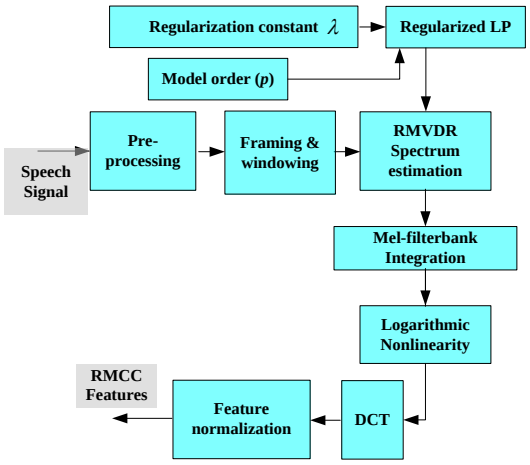

Fig. 3. Regularized MVDR cepstral coefficients (RMCC) feature extraction.

3.3. Backend

The backend of our system is very similar to the default system provided by the challenge organizers. The language models (LM) are the same: the search language model, the 5-gram rescoring LM and the RNNLM are the same. The training process is the same for the features with small dimension. For features with large dimension (like GBFB

and ICQF features), the output states are the same as for the MFCC features, but the input to the DNN corresponds to the feature dimension (with +/- 5 frames context). For features with smaller dimension, the initial alignment of the training set with MFCC features is used to train the GMM-HMM sat models for the new features. As mentioned before, the training data consists of all the training data from channels 1-6 and also includes the beamformed training data from channels 1, 3-6. The data is doubled by band-limiting each training audio file to 4 kHz. The training process is the same as provided by the organizers. We discriminatively train one DNN for each feature. For each track, we generate one ctm file for each feature and each set (i.e., development and evaluation). These ctm files are generated after rescoring with 5-gram LM followed by RNNLM rescoring.

3.4. Combining systems using ROVER

In this step we combine the ctm files of 6 or 7 systems, obtained in the previous step, using ROVER. As mentioned before, some of the features gave significantly lower WER for the evaluation set for some of the tracks. Combining the results from six or seven different features-based systems reduced the WER even further.
ROVER [15] reduces word error rates for automatic speech recognition systems by exploiting differences in the nature of the errors made by multiple speech recognition systems. It works in two steps:

- ✓ The outputs of several speech recognition systems are first aligned and a single word transcription network (WTN) is built.
- ✓ The best scoring word (with the highest number of votes) at each node is selected. The decision can also incorporate word confidence scores if these are available for all systems [15].

4. Experiments and Evaluation Results

Word error rates (WER) for each feature parameter and for each task are shown in Table 1. As mentioned before, for each feature parameter, we discriminatively train one DNN as provided by the default scripts. The same DNN is used to compute WER for all the tasks. For 1 channel task, there is no beamforming. For 2 channel and 6 channel tasks, the dev and eval sets go through appropriate beamforming using the beamforming software supplied by the organizers. In Table 1, the first row in each task corresponds to the default setup provided by the organizers. We ran the provided scripts and the results correspond to those scripts. The first row only uses channel 5 training data. The 2nd row for each task uses training data from channels 1 through 6 (channel 0 is not used). We also use the training data after beamforming using channels 1, 3, 4, 5, 6. Channel 2 was not used in this beamforming.

From Table 1 we can see that for 1ch task, the RMCC, GBFB and ETS-AFE features (rows 3-5) gave lower WER for the real test set than using the MFCC features (row 2). For 2 channel and 6 channel cases, only RMCC feature gave better results than the MFCC features. We combined results from different features using ROVER. We combined them in the WER order.

Table 1 : Average WER for the tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	MFCC (5ch)	11.46	13.10	23.08	20.88
	MFCC	9.46	10.65	18.87	16.43
	RMCC	8.46	11.24	15.16	15.83
	GBFB	9.33	12.74	17.61	18.03
	ETSI-AFE	10.02	12.54	17.65	17.01
	ICQF	11.03	15.93	22.12	22.28
	WPE-MFCC	14.02	15.78	28.44	22.87
	ICQC	13.62	19.03	26.06	27.62
	CWF-MFCC	16.39	18.39	31.09	23.70
	CWF2-MFCC	17.40	19.65	32.47	25.46
	ROVER	**6.79**	**9.27**	**12.70**	**13.72**
2ch	MFCC (5ch)	8.39	9.44	16.70	15.16
	MFCC	6.72	7.75	13.77	12.00
	RMCC	6.22	8.29	11.54	11.74
	GBFB	7.29	9.63	13.91	14.52
	ETSI-AFE	8.96	10.95	16.14	14.52
	ICQF	8.48	12.28	18.10	18.13
	WPE-MFCC	10.11	11.11	20.18	17.47
	ICQC	10.39	14.16	21.17	22.39
	CWF-MFCC	13.40	13.82	23.67	19.89
	CWF2-MFCC	12.79	14.31	25.81	21.23
	ROVER	**5.13**	**6.69**	**9.97**	**10.34**
6ch	MFCC (5ch)	6.08	6.82	11.50	10.73
	MFCC	4.86	5.49	9.97	8.75
	RMCC	4.86	5.98	8.65	8.71
	GBFB	5.96	7.40	10.40	10.70
	ETSI-AFE	7.09	8.67	12.42	11.30
	ICQF	6.74	9.41	13.31	13.72
	WPE-MFCC	6.75	7.82	13.54	13.27
	ICQC	8.19	10.16	14.16	16.00
	CWF-MFCC	8.13	9.94	17.09	15.56
	CWF2-MFCC	9.20	11.78	18.71	16.47
	ROVER	**4.00**	**5.07**	**7.23**	**7.53**

Table 2 : WER per environment for the best system.

Track	Envir.	Dev		Test	
		Real	simu	real	simu
1ch	BUS	8.54	7.95	18.75	9.73
	CAF	7.51	12.37	13.80	16.59
	PED	4.68	7.04	9.55	13.73
	STR	6.43	9.71	8.70	14.85
2ch	BUS	6.40	5.66	14.21	7.28
	CAF	5.24	8.63	9.90	12.05
	PED	3.78	5.03	8.20	10.80
	STR	5.10	7.42	7.58	11.23
6ch	BUS	5.24	4.48	9.44	4.97
	CAF	3.95	6.28	6.50	8.11
	PED	2.74	3.86	6.02	7.47
	STR	4.07	5.65	6.95	9.58

For 1ch task, we achieved the best results when we combine following 6 features: RMCC, GBFB, ETSI-AFE, MFCC, ICQF, and ICQC as shown in the last row for 1 channel results. For the 2 channel task, we achieved the best results when we combine 7 different features, namely, RMCC, MFCC, GBFB, ETSI-AFE, ICQF, WPE-MFCC and ICQC. For 6 channel task

also, we achieved the best results when we combine the outputs from these 7 different feature parameters in the same order.
These results are shown in the last row of each track. Results for each environment after ROVER are shown in Table 2. For 1 channel task, for real test set, we have reduced the WER by 45% (from 23.08% to 12.7%). For 2 channel task, WER has been reduced by 40% (from 16.7% to 9.97%), and for the 6 channel task the WER has been reduced by 37% (from 11.5% to 7.23%).

In table 3 we compared the WER of CRIM's system with the USTC-iFlytek system for CHiME-4 challenge with the lowest WER on the real portion of evaluation set [16]. Since we only used the default LMs, this comparison is with the default LMs for both the systems. Note that in [16], DNN-based single channel speech enhancement was used to enhance the signals, and, besides DNN-based acoustic model, deep convolutional neural networks (DCNN)-based upgraded acoustic models were also used. As we can see from table 3, CRIM's WER for 1ch system is close to the WER for the best CHiME-4 system. The primary reason for this is the noise robust RMCC features.

Table 3: WER comparison of CRIM's system with the best CHiMe-4 system [16] using the baseline (or default) language models on the evaluation set (real only).

Track	Real	
	CRIM	Best system [16]
1ch	12.7	11.15
2ch	9.97	5.41
6ch	7.23	3.24

5. Conclusion and Future Works

We presented automatic speech recognition systems developed at CRIM for the all three tracks (1ch, 2ch and 6ch) of CHiME-4 challenge. We used the same backend and baseline language models provided by the organizer. Therefore, to reduce word error rates (WER) we mainly focussed on the extraction of robust features and on system combination of various robust features-based sub-systems. Compared to the other features the RMCC features provided lowest WERs in all three tracks. By combining multiple hypotheses from different robust features-based systems we were able to reduce WER significantly from the baseline system. For 1ch track, for real test set, the WER was reduced by 45% (from 23.08% to 12.7%). For 2ch track, WER was reduced by 40% (from 16.7% to 9.97%), and for the 6 channel task the WER was reduced by 37% (from 11.5% to 7.23%).

In our future works we intend to keep RMCC features extractor fixed and focus on modifying the acoustic model and language models.

6. Acknowledgements

This work has been made possible by investments from the Ministere de l'economic et exportation (MEIE) of Government du Quebec.

7. References

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

[2] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K. "The Kaldi Speech Recognition Toolkit" in proc. of ASRU, pp. 4. Hawaïï, USA, December 2011.

[3] The 4th CHiME speech separation and recognition challenge: http://spandh.dcs.shef.ac.uk/chime_challenge/index.html.

[4] The REVERB challenge: http://reverb2014.dereverberation.com.

[5] N. Parihar, J. Picone, D. Pearce, H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," Proceedings of the European Signal Processing Conference, Vienna, Austria, 2004.

[6] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Regularized Minimum Variance Distortionless Response-Based Cepstral Features for Robust Continuous Speech Recognition", Speech Communication (2015), vol. 73, pp. 28-46.

[7] M. J. Alam, P. Kenny, P, Dumouchel, D. O'Shaughnessy, "Robust Feature Extractors for Continuous Speech Recognition", Proc. EUSIPCO (2014), Lisbon, Portugal.

[8] Schädler, M. R., Meyer, B. T., and Kollmeier, B., "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", Journal of the Acoustical Society of America (2012), Volume 131 (5), pp. 4134-4151.

[9] ETSI ES 202 050, Speech Processing, Transmission and Quality aspects (STQ), "Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms;" (2003).

[10] P. Cancela, M. Rocamora, E. Lopez, "An efficient multi-resolution spectral transform for music analysis," in proc. of the ISMIR, 2009.

[11] M. J. Alam, P. Kenny, "Low level and high level features for spoofing detection," submitted to IEEE journal of selected topics on Signal Processing (2016), August.

[12] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Ito Nobutaka, Keisuke Kinoshita, Miquel Espi, Takaaki Hori, Tomohiro Nakatani, and Atsushi Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in Proc. of the REVERB Workshop (2014).

[13] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," IEEE Signal Processing Letters (2013), vol. 20, no. 3, pp. 217–220.

[14] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances of deep learning for speech research at Microsoft," ICASSP, 2013.

[15] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", Proc. ASRU, pp. 347-354, 1997.

[16] Jun Du, Yan-Hui, Lei Sun, Feng Ma, Hai-Kun Wang, Jia Pan, Cong Liu, Chin-Hui Lee, "The USTC – iFlytek System for CHiME-4 Challenge", http://spandh.dcs.shef.ac.uk/chime_workshop/papers/CHiME_2016_paper_21.pdf