ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152www.ijmece .com
Vol 5, Issuse.2 May 2017

Multi-channel speech processing architectures for noise robust speech recognition: 3rd Chi ME challenge results

Mr.JADA LINGAIAH , Mr.M.SREENU, Mr.JEEDIMADLA VENKATESHAM, GANJI RAMYA

ABSTRACT

Recognizing speech under noisy condition is an ill-posed problem. The CHiME 3 challenge targets robust speech recognition in realistic environments such as street, bus, caf- fee and pedestrian areas. We study variants of beamformers used for pre-processing multi-channel speech recordings. In particular, we investigate three variants of generalized side- lobe canceller (GSC) beamformers, i.e. GSC with sparse blocking matrix (BM), GSC with adaptive BM (ABM), and GSC with minimum variance distortionless response (MVDR) and ABM. Furthermore, we apply several post- filters to further enhance the speech signal. We introduce MaxPower postfilters and deep neural postfilters (DPFs).DPFs outperformed our baseline systems significantly when measuring the overall perceptual score (OPS) and the per- ceptual evaluation of speech quality (PESQ). In particular DPFs achieved an average relative improvement of 17.54% OPS points and 18.28% in PESQ, when compared to the CHiME 3 baseline. DPFs also achieved the best WER when combined with an ASR engine on simulated development and evaluation data, i.e. 8.98% and 10.82% WER. The proposed MaxPower beamformer achieved the best overall WER on CHiME 3 real development and evaluation data, i.e. 14.23% and 22.12%, respectively.

Index Terms— multi-channel speech processing, deeppostfilter, automatic speech recognition

INTRODUCTION

Background noise is the primary source of performance degradation in speech recognition systems. While the capabilities of single-channel speech pre-processing are limited, multi-channel systems exploit the spatial information of the sound field and usually achieve better speech recognition results. Adaptive beamforming is a widely used technique for multi-channel pre-processing of speech as alternative to blind source separation approaches. For a sufficient amount of noise reduction, beamformers are generally used in con- junction with a postfilter.The aim of the 3rd CHiME challenge is to develop a multi- channel speech recognition system [1], where we encounter multi-channel recordings of a speaker located in the near- field, embedded in mostly far-field noise. The setup covers different speakers, noise environments, and real-world prob-lems like microphone failure, clipping, and other recording glitchesIn this paper, we present a multi-channel speech enhance-ment system which tries to cope with these conditions: First,we

ASSISTANT PROFESSOR[1,2,3], STUDENT [4]
Department of ECE
Arjun College Of Technology & Sciences
Approved by AICTE& Affiliated to JNTUH
SPONSORED BY BRILLIANT BELLS EDUCATIONAL SCOITEY

detect recording glitches using the prediction error of an auto-regressive model. Then, we estimate the position of thespeaker relative to the microphone array using our *direction-dependent signal-to-noise ratio* (DD-SNR) algorithm [2],which also provides a sufficiently accurate *voice activity detection* (VAD). The speaker position is used to obtain a steering vector for a *generalized sidelobe canceller* (GSC) beamformer, which we implemented in three different variants. We also present two novelties here: Firstly we introducea *MaxPower* postfilter (PF), leading to the best speech recognition result on CHiME 3 real data. Secondly we present deep neural PFs – deep neural networks attached to beamformers, improving the overall perceptual quality (OPS) of the target speech significantly and also outperforming baseline systems on simulated data. This front-end, i.e. the three beamformer variants and different PFs, are empirically evaluated using thePESQ and the OPS measures [3].

In the back-end, we use two speech recognition systems based on the Kaldi toolkit [4]. The first is a GMM system which makes extensive use of feature transformationsas this was shown to provide good results for distant talk speech recognition [5]. The second is a DNN system that employs pre-training with restricted Boltzmann machines, cross entropy training and state-level minimum Bayes risk training [1]. Our best model, i.e. MaxPower PF with a GMMbackend, reduces word error rate (WER) from 37.61% for the baseline enhancement system to 22.12% (41% relative improvement) on the real evaluation set.

The outline of the paper is as follows: In Section 2 we introduce the architecture of the proposed system. Section 3 details the multi-channel speech processing approaches including the proposed beamformers. PFs are introduced in Section 4 while the PESQ and PEASS scores of the front-end aresummarized in Section 6.1. The ASR system is presented in Section 5 and the results are discussed in Section 6.2. Section 7 concludes the paper.

OPS scores

train
valid
test

1. **SYSTEM OVERVIEWs recording glitches, amplitude variations, time shifts or total signal loss must be detected before multi-channel speech enhancement such as beamforming. In particular, we noticed that especially channel 4 and 5 exhibit rather complex recording glitches in about 15% of all isolated recordings. To address these problems, a mere energy threshold may not suffice. We therefore employed an auto-regressive linear predictive coding (LPC) on each channel c in time-domain [6, 7], and used the predictor error $e(t)$ as criterion whether a channel is considered as failed, i.e.**

MULTI-CHANNEL SPEECH PROCESSING

The input signal vector $\boldsymbol{X}$ of the 6 microphone channels is written as

$$\boldsymbol{X}(k, l) = \boldsymbol{A}(k, l)S(k, l) + \boldsymbol{N}(k, l), \qquad (1)$$

where S is the speech signal, $\boldsymbol{N}$ is the noise part of the 6-channel input signal in frequency-domain, k and l denote the frequency bin and time frame, respectively, and $\boldsymbol{A}(k, l)$ where $a(m)$ are LPC coefficients and M = 100. A channel$x_c(t)$ is considered as failed if the power of its predictor error$e(t)$ lies outside the $10dB$ corridor around the median ± of the energy of the predictor errors of all channels. If a failed channel is detected this channel is not used for further processing.

Direction Of Arrival Estimation

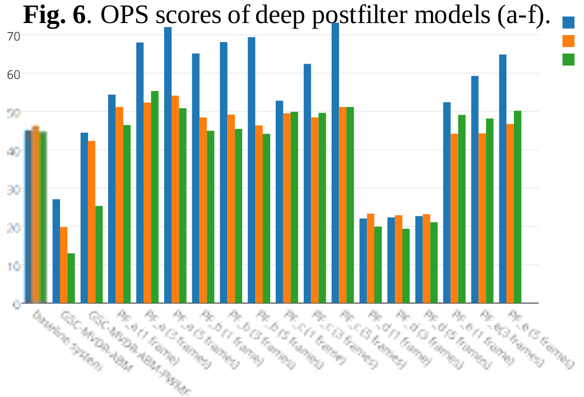
For successful beamforming an accurate *direction of arrival* (DOA) estimation is required. Therefore, the *steered response power phase transform* (SRP-PHAT) [8] algorithm has been already provided for this purpose. But it lacks a proper VAD estimate, which might also be useful for estimating the spatial noise correlation matrix $\boldsymbol{\Phi}_{NN}$ during speech pauses. For this purpose, we used our DD-SNR algorithm [2], which provides a direction-dependent a-priori SNR ξ_τ (k, l) under the assumption of an ideal, spherical noise sound field,i.e.

$$\xi_\tau(k, l) = \mathrm{Tr}([\boldsymbol{\Gamma}_{\boldsymbol{XX}}(k, l) - \boldsymbol{A}_\tau(k, l)\boldsymbol{A}^H(k, l)]^{-1}$$

denotes the *acoustic transfer function* (ATF) from the true speaker position to each microphone. In this challenge, additional information is supplied by the *noise context*, a short

are clean recordings mixed with noise that has been recordedin the same noisy environments. The real recordings were made using 6 microphones custom-fitted to a tablet hand-held device. The recordings with this device were conductedin four different environments: on a bus (BUS), in a cafe´ (CAF), in a pedestriean area (PED), and at a street junction (STR). For real data, there is an additional channel recorded with a head-mounted close-talking microphone. This channel, however, may not be used directly for obtaining ASR results but is only to be used in training.

Fig. 6. OPS scores of deep postfilter models (a-f).



the CHiME 2 challenge [5]. The DNN system employs 40-dimensional filterbank features and is pre-trained using restricted Boltzmann machines with 6 hidden layers. The ac-tual training stage of the DNN uses 4 hidden layers and also does cross entropy training. Finally, sequence discriminative training is performed using a state-level minimum Bayes riskcriterion.

In the following sections, we describe the changes we made to the baseline system. These are to be found in the frontend and in the postprocessing stage.

Feature extraction

In contrast to the baseline which uses MFCC features, we additionally employ power-normalised cepstral coefficients (PNCC) [26]. For these features, we use a Hamming window with a window duration of 25 ms and a step size of 10 ms. Parallel to MFCCs, we extract 13 features and collect deltas and delta-deltas of these.

Rescoring

The postprocessing step features n-best list language model rescoring. For this, we collect the 36 best hypotheses for each utterance and reweight them with a class-based recurrent neu-ral network language model (RNN-LM) [27]. The RNN-LM is trained on the official training data only and is configured to use a class size of 50.

2. RESULTS AND DISCUSSION

The data of the challenge and the recording setup is de-scribed in detail in [1]. The data is a collection of two sets of recordings: real data and simulated data. The first are speech recordings made in noisy environments. The secondABM and deep postfilter (PF$_a$) outperforms the other beam- formers in terms of OPS and PESQ scores. In particular the proposed system achieved an average relative improvement of 17.54% in OPS and 18.28% in PESQ compared to the baseline enhancement system.

Preprocessing results

To evaluate our three beamformers, we used PESQ and OPS scores. Evaluation is performed against the close-talking mi- crophone channel for the real data set, and against the WSJ corpus for the simulated data set. Tables 1 and 2 show the scores for our four beamformers, and the baseline enhance- ment system for comparison. Again the GSC-MVDR with

	set	train	dev	eval
Baseline enhancement	simu	2.00	1.64	1.72
system	real	1.59	1.42	1.50
GSC with sparse BM,	simu	2.15	1.73	1.81
and PMWF	real	1.51	1.37	1.35
GSC with ABM,	simu	1.53	1.49	1.52
and PMWF	real	1.36	1.30	1.36
GSC with MVDR	simu	2.05	1.60	1.73
and ABM	real	1.60	1.45	**1.73**
GSC with MVDR	simu	**2.55**	**2.17**	**2.28**
and ABM, and PF$_a$	real	**1.73**	**1.56**	1.56
GSC with MVDR	simu	1.98	1.69	1.63
and ABM, and MaxPower PF	real	1.51	1.39	1.44

Table 1. PESQ scores for our beamformers with PFs and thebaseline.

	set	train	dev	eval
Baseline enhancement	simu	54.80	44.22	47.31
system	real	44.66	40.98	31.48
GSC with sparse BM,	simu	59.64	46.99	46.77
and PMWF	real	38.69	33.05	29.04
GSC with ABM,	simu	48.61	43.84	43.71
and PMWF	real	43.16	42.81	**38.02**
GSC with MVDR	simu	52.4	45.87	47.18
and ABM	real	48.26	45.87	37.93
GSC with MVDR	simu	**63.94**	**53.83**	**54.53**
and ABM, and PF$_a$	real	**48.69**	**46.54**	37.72
GSC with MVDR	simu	56.08	44.82	44.48
and ABM, and MaxPower PF	real	47.18	44.90	36.96

Table 2. OPS scores for our beamformers with PFs and the baseline.

.2. ASR results

Table 3 shows ASR results for the preprocessing methods pre-sented in this paper. MaxPower outperforms all other pro-posed methods on the real development data and the real eval- uation data (14.53% WER and 22.14% WER, respectively), whereas PF$_a$ achieved the best ASR scores on simulated data,

i.e. 8.98% and 10.82% on development and evaluation, re-

spectively. When comparing MFCCs and PNCCs , on aver- age, PNCCs lead to an improvement of 6.04% WER on the real evaluation set. Improvements vary, however, depend- ing on noise environment and preprocessing. After language model rescoring, the scores for the real development set and the real evaluation set descrease slightly to 14.23% WER and 22.12% WER, respectively (see Table 4).

Due to time constraints, our results for the DNN-based ASR system are limited to MaxPower which achieves best results among GMM-based systems. While considerable improvements are gained for the system using MFCCs (3.02% WER on real evaluation set), DNNs lead to in- creased WER for the system using PNCCs (+2.03% WERon real evaluation set).

		development		evaluation	
	features	real	simu	real	simu
Baseline	MFCC	20.38	9.72	37.61	11.10
GSC sparse BM	MFCC	26.14	10.39	44.01	12.75
GSC ABM	MFCC	15.66	20.15	36.39	79.05

+ MVDR	MFCC	16.78	10.16	27.45	11.47
+ PF$_a$	MFCC	17.93	**8.98**	27.72	**10.82**
+ MaxPower	MFCC	15.70	10.77	25.22	14.86
+ DNN	FBANK	14.54	9.52	22.20	15.67
Baseline	PNCC	18.99	11.14	31.57	12.15
GSC sparse BM	PNCC	22.32	11.17	36.98	13.87
GSC ABM	PNCC	15.60	21.96	34.02	77.47
+ MVDR	PNCC	16.34	11.01	24.55	12.69
+ PF$_a$	PNCC	16.77	10.64	25.58	12.37
+ MaxPower	PNCC	**14.53**	12.05	**22.14**	15.08
+ DNN	FBANK	15.79	10.42	24.17	16.72

Table 3. ASR results for our beamformers and the baseline enhancement system.

	development		evaluation	
environment	real	simulated	real	simulated
BUS	16.17	10.52	29.00	12.46
CAF	13.78	13.97	24.04	15.61
PED	11.73	9.53	19.75	14.81
STR	15.26	13.38	15.69	16.64
AVG	14.23	11.85	22.12	14.88

Table 4. Detailed results for single best system, MaxPower using PNCC features and RNN language model rescoring.

CONCLUSION

We presented a comparison of different beamformers and postfilters applied to the CHiME 3 speech database. We studied three variants of GSC beamformers, i.e. GSC with sparse blocking matrix (BM), GSC with adaptive BM (ABM), and GSC with minimum variance distortionless response (MVDR) and ABM. In addition we investigated three postfil- ters (PF), a MaxPower PF, a parametric multi-channel Wiener filter, and a deep neural PF. The proposed ASR systens use either MFCC or PNCC features calculated from the the pre- processed signals which are fed into GMM or DNN-based systems. Finally n-best list re-scoring, using a recurrent neu-ral network (RNN) language model, was applied.

We evaluated the overall perceptual score (OPS), and per- ceptual evaluation of speech quality (PESQ) of the proposed beamformers and postfilters. Deep neural postfilters using an GSC-MVDR-ABM beamformer outperformed other BF systems significantly, achieving an average relative improve-ment of 17.54% in OPS and 18.28% in PESQ compared to the baseline system. However, improvements in OPS were not reflected in the ASR performance on the real data set, al- though the best scores were achieved on the simulated data. The GSC-MVDR-ABM beamformer followed by the Max- Power postfilter and GMM ASR achieved the best WER on real data. This configuration obtained a 22.14% WER and a 22.12% WER on the real evaluation set, with or without re- scoring, respectively.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF) under the project number P27803-N15, and the K- Project ASD. The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), the Province of Styria – Government of Styria and the Technology Agency of the City of Vienna (ZIT). The pro- gram COMET is conducted by Austrian Research Promotion Agency (FFG). Furthermore, we acknowledge NVIDIA for providing GPU computing resources.

REFERENCES

J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ’CHiME’ speech separation and recognition chal- lenge: Dataset, task and baselines,” in *IEEE 2015 Auto- matic Speech Recognition and Understanding Workshop (ASRU)*, 2015, submitted.

L. Pfeifenberger and F. Pernkopf, “Blind source extrac- tion based on a direction-dependent a-priori SNR,” in *Interspeech*, 2014.

V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing So- ciety.

Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Her- shey, “Discriminative methods for noise robust speech recognition: A chime challenge benchmark,” in *Pro- ceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME*, 2013,pp. 19–24.

T. D. Rossing, *Springer Handbook of Acoustics*, Springer, Berlin–Heidelberg–New York, 2007.

P. Vary and R. Martin, *Digital Speech Transmission*, Wiley, West Sussex, 2006.

J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin–Heidelberg–New York, 2008.

E. Warsitz and R. Haeb-Umbach, “Blind acoustic beam- forming based on generalized eigenvalue decomposi- tion,” *IEEE Transactions on audio, speech, and lan- guage processing*, vol. 15, no. 5, 2007.

R. Talmon, I. Cohen, and S. Gannot, “Relative transfer function identification using convolutive transfer func- tion approximation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, 2009.

W. Herbordt and W. Kellermann, “Analysis of block- ing matrices for generalized sidelobe cancellers for non- stationary broadband signals,” *IEEE International Con- ference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002.

E. Warsitz, A. Krueger, and R. Haeb-Umbach, “Speech enhancement with a new generalized eigenvector block- ing matrix for application in a generalized sidelobe can- celler,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 73–76, 2008.

M. Souden, J. Chen, J. Benesty, and S. Affes, “An in- tegrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

O. Hoshuyama, A. Sugiyama, and A. Hirano, "A ro- bust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, 1999.

L. Pfeifenberger and F. Pernkopf, "A multi-channel postfilter based on the diffuse noise sound field," in *Eu- ropean Association for Signal Processing Conference*, 2014.

M. G. Shmulik, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints GSC beam- former," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

J. Li, Q. Fu, and Y. Yan, "An approach of adaptive blocking matrix based on frequency domain indepen- dent component analysis in generalized sidelobe can- celler," *IEEE 10th International Conference on Signal Processing*, pp. 231–234, 2010.

K. Lae-Hoon, M. Hasegawa-Johnson, and S. Koeng- Mo, "Generalized optimal multi-microphone speech enhancement using sequential minimum variance dis- tortionless response(MVDR) beamforming and postfil- tering," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2006.

J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin– Heidelberg–New York, 2008.

Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude es- timator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, 1985.

M. Zöhrer and F. Pernkopf, "Representation models in single channel source separation," in *IEEE InternationalConference on Acoustics, Speech, and Signal Process- ing*, 2015.

M. Zöhrer and F. Pernkopf, "Single channel source separation with general stochastic networks," in *Inter- speech*, 2014.

M. Zöhrer, R. Peharz, and F Pernkopf, "Representation learning for single-channel source separation and band-width extension," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015, accepted.

Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM Transactions on Audio, Speech and Language Process- ing, vol. 22, no. 12, pp. 1849–1858, 2014.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propaga- tion," in Neurocomputing: Foundations of Research, James A. Anderson and Edward Rosenfeld, Eds., pp. 673–695. MIT Press, Cambridge, MA, USA, 1988.

C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2012, pp. 4101–4104.

T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based lan- guage model," in INTERSPEECH, 2010.