Investigation of Neural Networks Based Beamforming Approaches for Speech Recognition: The NTU Systems for CHiME-4 Evaluation

Xiong Xiao[1], *Chenglin Xu*[1], *Zhaofeng Zhang*[2], *Shengkui Zhao*[3], *Sining Sun*[4], *Shinji Watanabe*[5]
Longbiao Wang[6], *Lei Xie*[4], *Douglas L. Jones*[3], *Eng Siong Chng*[1], *Haizhou Li*[7,1]

[1]Nanyang Technological University (NTU), Singapore, [2]Nagaoka University of Technology, Japan,
[3]Advanced Digital Sciences Center, Singapore, [4]Northwestern Polytechnical University, China,
[5]Mitsubishi Electric Research Laboratories, USA, [6]Tianjin University, China,
[7]National University of Singapore, Singapore.

{xiaoxiong, xuchenglin}@ntu.edu.sg, s147002@stn.nagaokaut.ac.jp, shengkui.zhao@adsc.com.sg

Abstract

This paper studies neural networks based beamforming for robust speech recognition. We investigated two types of network based beamforming, 1) directly predicting beamforming weights; 2) predicting time frequency mask that is used to determine MVDR beamforming weights. The beamforming networks are trained using both mean square error (MSE) criterion and cross entropy (CE) of frame phone classification. Variations of these approaches are studied on the CHiME-4 Evaluation.

1. Background

We consider mainly the 2-channel (2ch) and 6-channel (6ch) tracks, but also provide baseline results on 1-channel to show the effect of using filterbank features in acoustic modeling.

2. Contributions

The main contribution is the comparison of two recently proposed beamforming networks in the same framework, and also extend them. The first network is proposed in [1] in which generalized cross correlation (GCC) features are used to predict BF weights in frequency domain. The extension we did here is to use one network to cover several array geometries. Specifically, in the 2ch track, we train one network to handle the different between-microphone distances of different microphone pairs. In [1], only one array geometry can be handled by the network.

The second network we study is the time-frequency (TF) mask predicting network for spatial covariance matrix (SCM) estimation [2, 3, 4]. We made several improvements over the previous works: 1) use CE cost function of ASR to refine the mask estimation network to optimize the network for ASR and also to avoid using heuristics in obtaining target mask; 2) cascaded mask prediction to significantly boost ASR performance.

Another contribution of this work is to build a Matlab based platform for deep learning based array speech processing. The toolkit is called SignalGraph and freely available in Github: https://github.com/singaxiong/SignalGraph. SignalGraph supports arbitrary directed acyclic graph (DAG) network topology and a lot of signal processing layer types. It also supports DNN, CNN, and LSTM modules. The toolkit has been used in more than 12 publications in the past 3 years. We will release the recipes of this work when they are ready.

3. Experimental evaluation

We use 40D log Mel filterbank features unless stated otherwise, followed by cepstral mean normalization (subtracting utterance mean). No pre-emphasis or DC removal is applied. Delta and acceleration features are appended and then 11 frames of feature vectors are cascaded to form the input for the DNN acoustic model. Two types of DNN acoustic model is used, one is trained from the channel 5 data (called ch5 model), while the other is trained from all the 6 channel's data (called chall model).

Notations: MLBF (maximum likelihood beamforming [5]), BFnet (network to predict beamforming weights [1]), Masknet (network to predict TF mask). The MLBF uses a GMM with 32 diagonal covariance Gaussians in MFCC domain for likelihood computation. All BFnets uses 3x1024 hidden layer/nodes DNN. All Masknet uses 1 layer LSTM with 1024 cells.

3.1. 1 channel

- System 1: MFCC + ch5 model (Official baseline)
- System 2: Fbank + ch5 model
- System 3: Fbank + chall model
- System 4: Masknet + chall model

In System 4, we use the mask as filter weights and multiply it with the complex Fourier coefficients of input elementwise. The Masknet is trained using CE criterion.

3.2. 2 channels

- System 1: BeamformIt + MFCC + ch5 model (Official)
- System 2: BeamformIt + Fbank + ch5 model
- System 3: BeamformIt + Fbank + chall model
- System 4: MLBF + chall model
- System 5: BFnet + MSE training + chall model
- System 6: BFnet + CE training + chall model
- System 7: Masknet 1ch + MSE + chall model
- System 8: Masknet 2ch + MSE + chall model
- System 9: Masknet 1ch + CE + chall model
- System 10: Masknet 1ch + CE + 3-pass + chall
- System 11: Split Masknet 1ch + CE + chall
- System 12: Split Masknet 1ch + CE + 3-pass + chall

- System 13: Split Masknet 1ch + CE + 3-pass + Mask filter + chall
- System 14: System 13 + 5gram
- System 15: System 14 + rnnlm

In System 4, MLBF is similar to LIMABEAM [5] in concept, but predicting BF weights in frequency domain instead of time domain. Results shown is from a preliminary study and not fine tuned. In System 5, BFnet is trained to mimic delay-and-sum (DS) beamforming on the simulated training data (10x amount generated using the provided tool). In System 6, BFnet from System 5 is refined to optimize framewise phone classification CE cost obtained with ch5 acoustic model (AM) on the 8738 training set. The AM is not updated. All other CE trained nets in this study is similarly configured. In System 7, Masknet is trained to predict ideal binary mask using a threshold of 0dB on 10x simulated data. Input is log spectra (mean subtracted, dynamic features appended) of first given channel. In System 8, log spectra of both channels are appended and used as the input of the Masknet. In System 9, Masknet is refined with CE cost. In System 10, the Masknet is applied 3 times, where the last two passes use the enhanced speech as input. In System 11, the Masknet predicts speech and noise masks separately. In System 13, the mask filter used in System 4 of 1ch track is used to postprocess beamformed speech.

From the results, Masknet outperforms BFnet, possibly due to its use of noise information. It is good to 1) refine networks with CE training; 2) estimate noise and speech masks independently; 3) use multiple mask estimation passes to iteratively refine masks; 4) use both channels as input to estimate masks.

3.3. 6 channels

- System 1: Masknet 1ch + MSE + chall
- System 2: Masknet 1ch + CE + 3-pass + chall
- System 3: Split Masknet 1ch + CE + chall
- System 4: Split Masknet 1ch + CE + 3-pass + chall
- System 5: System 4 + 5gram
- System 6: System 5 + rnnlm
- System 7: Split Masknet 2ch + CE + Split Masknet 1ch + CE + 2-pass + Mask filter + chall
- System 8: System 7 + 5gram
- System 9: System 8 + rnnlm

Compared to System 4, System 7 uses 2ch input (channels 4&5) for first pass mask estimator and also uses a mask filter for post-processing. System 7 significantly improves performance on et05-simu (because System 4 uses channel 1 for mask estimation which has very low SNR in et05-simu) but degrades et05-real (reason unknown yet). The results show the importance of choosing channels for mask estimation.

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

Track	System	Dev		Test	
		real	simu	real	simu
1ch	System 1	14.86	15.74	27.27	24.09
	System 2	15.51	16.86	26.44	25.40
	System 3	12.39	14.80	21.62	21.96
	System 4	11.63	14.25	21.18	21.37
2ch	System 1	10.90	12.36	20.44	19.03
	System 2	11.92	13.05	20.77	20.22
	System 3	10.10	11.68	17.20	18.24
	System 4	11.47	14.66	18.81	19.81
	System 5	9.94	11.45	17.15	18.51
	System 6	9.65	11.52	16.52	16.77
	System 7	9.80	10.40	16.57	17.03
	System 8	9.24	10.19	15.48	14.88
	System 9	9.37	10.13	15.67	16.24
	System 10	9.07	10.02	15.00	14.96
	System 11	8.86	9.97	15.17	15.31
	System 12	8.75	9.92	14.46	14.34
	System 13	8.35	9.50	14.35	14.19
	System 14	6.99	8.10	12.26	12.11
	System 15	6.07	7.06	10.77	10.72
6ch	System 1	8.26	7.09	12.82	19.47
	System 2	6.41	6.10	9.41	11.12
	System 3	6.52	6.08	10.06	11.92
	System 4	6.08	5.98	9.02	9.90
	System 5	4.88	4.91	7.44	8.12
	System 6	4.13	4.29	6.37	7.11
	System 7	6.41	5.87	9.63	7.61
	System 8	5.26	4.75	7.79	5.98
	System 9	4.49	4.12	6.75	4.99

Table 2: WER (%) per environment for the best system. System 4 for 1ch, System 15 for 2ch, System 9 for 6ch.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	14.99	12.40	28.35	16.08
	CAF	12.15	18.55	24.17	25.70
	PED	8.37	11.36	18.04	21.35
	STR	11.02	14.69	14.16	22.34
2ch	BUS	7.22	5.86	14.27	7.77
	CAF	6.11	9.99	12.01	13.04
	PED	4.81	5.77	8.70	10.76
	STR	6.12	6.62	8.09	11.30
6ch	BUS	6.58	3.61	8.37	4.15
	CAF	3.90	5.34	6.16	5.38
	PED	3.73	3.78	7.12	5.02
	STR	3.75	3.73	5.34	5.42

4. Acknowledgments

This work is supported by DSO funded project MAISON DSOCL14045.

5. References

[1] X. Xiao and et al, "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*. IEEE, 2016.

[2] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *ASRU*. IEEE, 2015, pp. 444–451.

[3] J. Heymann and et al, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016.

[4] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *INTERSPEECH*, 2016.

[5] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, sep 2004.