PBF-GSC BEAMFORMING FOR ASR AND SPEECH ENHANCEMENT IN REVERBERANT ENVIRONMENTS

Yi Ren Leng, Jonathan William Dennis, Ng Wen Zheng Terrence, and Huy Dat Tran
Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore

Optimizing performance in realistic noisy environments:

- Using multi-condition models instead of clean models for better performance in noise
- Concentrate on RealData test set as it is closer to real situations and offers more room for improvement
- Beamform 8-channel audio into a single enhanced channel using proposed PBF-GSC system combining Generalized Sidelobe Canceller (GSC) and Phase-error based filtering (PBF)

Generalized Sidelobe Canceller (GSC)

- Reference signal for GSC obtained by combining the audio channel STFT weighted by the PBF masking coefficients transformed back into the temporal domain
- Adaptive LMS filter using SGD iteratively filters the reference signal using a simple blocking matrix

Phase-error based filtering (PBF)

- Maximize SNR between channels using phase angle

$$\theta_{ij}(t) = \angle X_i(t) - \angle X_j(t), \qquad i \neq j$$

where X_i is the STFT of the audio channel

- The phase errors are combined into a masking coefficient for each channel *i*

$$M_i(t) = \left(\prod_{j=1, j\neq i}^{n} \frac{1}{1+\gamma\theta_{ij}^2(t)} \right)^{\frac{1}{k}}$$

Speech Recognition Results:

	Beamform	SimData Near1	SimData Far1	SimData Near2	SimData Far2	SimData Near3	SimData Far3	SimData Avg.	RealData Near	RealData Far	RealData Avg.
HTK	Baseline	16.76	18.35	20.78	32.75	24.79	39.96	25.55	49.92	47.54	48.73
HTK	FBF only	16.66	17.35	16.88	23.35	18.94	27.95	20.18	37.11	38.15	37.63
HTK	GSC only	17.50	18.65	17.14	21.93	18.84	27.71	20.29	37.66	37.44	37.55
HTK	PBF only	16.59	17.64	16.93	23.01	19.01	27.69	20.14	36.92	37.85	37.38
HTK	Proposed	17.43	18.67	17.14	22.03	18.89	27.52	20.27	35.68	36.66	36.17
Kaldi DNN	Proposed	13.17	12.52	11.00	13.33	12.63	17.54	13.36	32.32	33.56	32.94
Kaldi SGMM	Proposed	11.52	11.44	9.66	12.56	11.09	15.25	11.92	28.68	30.89	29.79

Speech Enhancement

- PBF-GSC output is enhanced further by online estimation of gamma modeling of speech power: where σ_S^2 denotes the variance of the speech spectrum and (*a;b;L*) are distribution parameters, while the Gaussian distribution is used to model the noise complex spectrum distributions.

$$p(|S|) = \frac{b^a}{\Gamma(a)\sigma_S}\left(\frac{|S|}{\sigma_S}\right)^{La-1} \exp\left[-b\left(\frac{|S|}{\sigma_S}\right)^L\right]$$

- MAP estimator gives a closed form solution for the spectral magnitude estimate: where ξ and γ denote the prior and posterior SNR estimates at each time-frequency index

$$G = \frac{1}{2\left(1+\frac{b}{\xi}\right)} + \sqrt{\frac{1}{4\left(1+\frac{b}{\xi}\right)^2} + \frac{4a-3}{4\gamma\left(1+\frac{b}{\xi}\right)}}$$

Speech Enhancement Results:

	Measure	SimData							RealData		
		Near1	Far1	Near2	Far2	Near3	Far3	**Avg.**	Near	Far	**Avg.**
Single-channel	Cepst Distance	3.30	3.71	5.01	5.58	4.76	5.35	4.62	-	-	-
	Loglike Ratio	0.58	0.64	0.79	1.00	1.00	1.12	0.86	-	-	-
	Seg SNR	6.61	6.16	4.23	3.25	4.19	4.19	4.58	-	-	-
	SRMR	5.3	5.72	6.26	5.37	5.15	4.42	5.37	7.97	7.65	7.81
8-channel	Cepst Distance	2.15	2.48	2.75	3.88	2.88	3.94	3.01	-	-	-
	Loglike Ratio	0.26	0.31	0.39	0.57	0.48	0.67	0.45	-	-	-
	Seg SNR	11.08	10.1	6.84	5.76	7.35	4.53	7.61	-	-	-
	SRMR	4.09	4.15	3.5	3.67	3.90	3.21	3.75	3.67	3.79	3.73

Selected References:

- Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura, “Gamma modeling of speech power and its on-line estimation for statistical speech enhancement,” IEICE - Trans. Inf. Syst., vol. E89-D, no. 3, pp. 1040–1049, Mar. 2006.
- C. Knapp and G. Clifford Carter, “The generalized correlation method for estimation of time delay,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 24, no. 4, pp. 320–327, 1976.
- Lloyd J. Griffiths and Charles W. Jim, “An alternative approach to linearly constrained adaptive beamforming,”IEEE Trans. Antennas and Propagation, vol. 30, no. 1, pp. 27–34, Jan. 1982.
- C. Y. Lai and P. Aarabi, “Multiple-microphone timevarying filters for robust speech recognition,” in Proc. ICASSP, 2004, vol. 1, pp. 233–236.

A*STAR Institute for Infocomm Research