Optimization of Speaker Extraction Neural Network with Magnitude and Temporal Spectrum Approximation Loss

Chenglin Xu[1,2], Wei Rao[3], Eng Siong Chng[1,2], Haizhou Li[2,3]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NUS National University of Singapore

1. Contributions

We propose a speaker extraction approach to extract target speaker's voice from a multi-talker mixture. Our contributions are:

- a magnitude and temporal spectrum approximation loss that calculates direct signal reconstruction error and considers the speech context;
- a concatenation framework that encodes speaker characteristics into the mask estimation network instead of context adaptive deep neural network (CADNN) in SpeakerBeam-FE (SBF) method [1];

2. Speaker Extraction

- **Problem Formulation**

The speech extraction aims to extract the target speaker's voice $x(n)$ from a multi-talker mixture $y(n)$ given a different speech segment $a(n)$ of the target speaker. The mixed signal is,

$$y[n] = x[n] + \sum_{i=1}^{I} z_i[n] \quad (1)$$

where $z_i[n]$ might be any number of interference speech or background noise.

- **Frequency Domain Solution**

a). With the spectra of mixed signal $|Y|$ and enroll signal $|A|$, the mask M for target speaker is always estimated by a network using either mask approximation loss or spectrum approximation loss.

$$M = G(|Y|, |A|) \quad (2)$$

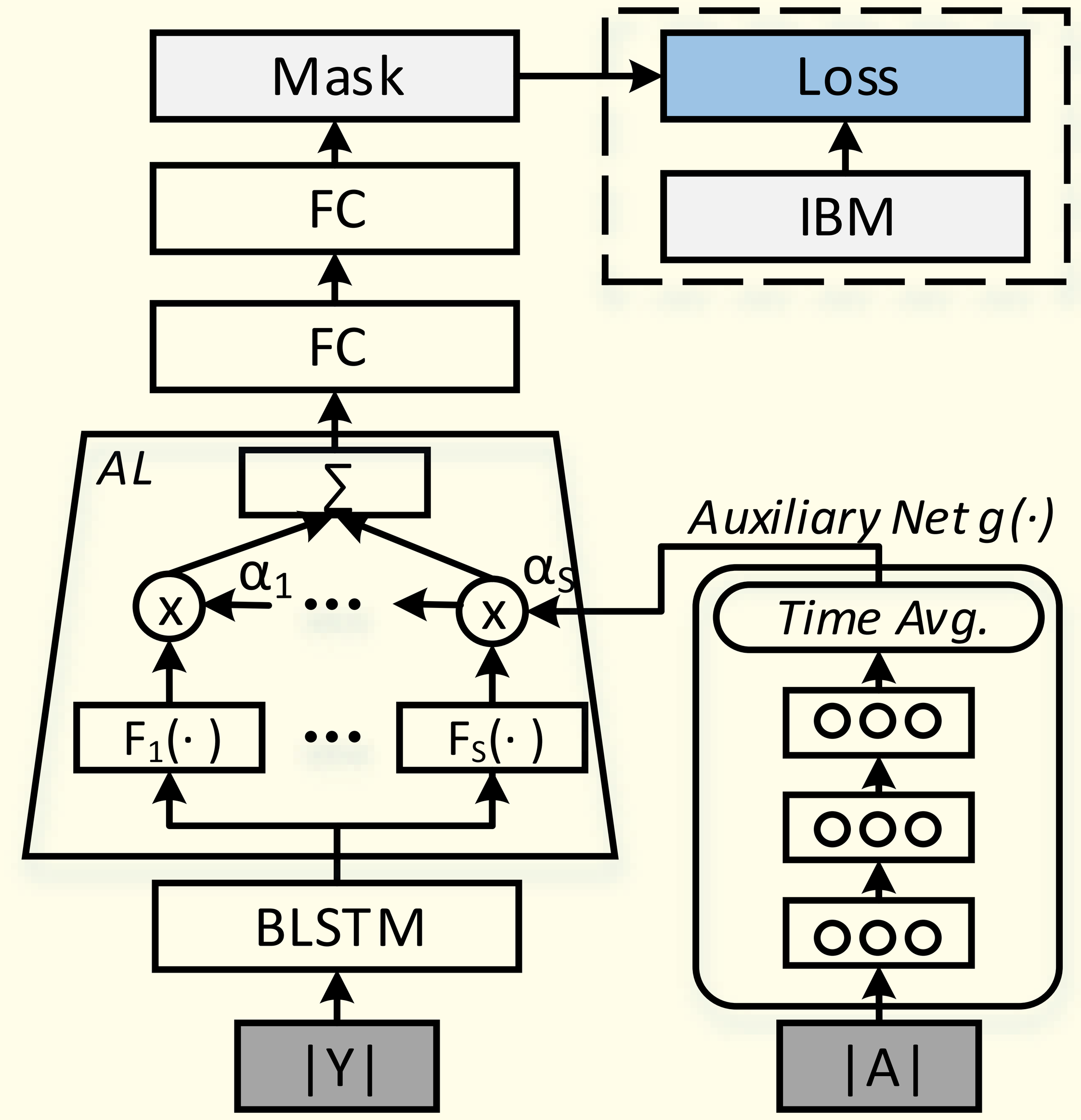
b). The magnitude $|\hat{X}|$ of the target speaker is obtained by,

$$|\hat{X}| = M \odot |Y| \quad (3)$$

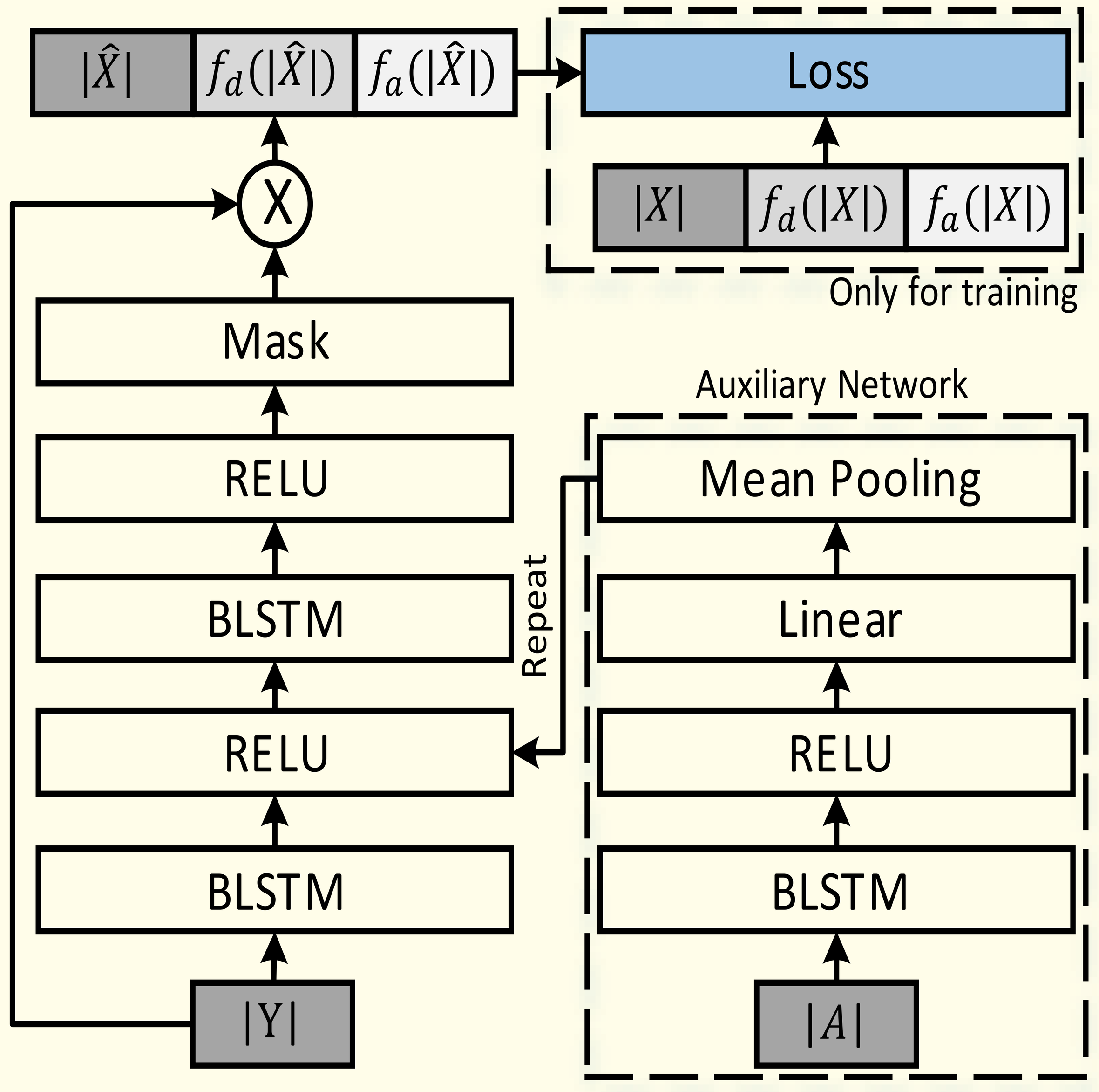
c). The time domain signal $\hat{x}$ of target speaker is reconstructed by overlap-and-add algorithm after doing iSTFT on estimated magnitude $|\hat{X}|$ and noisy phase $\angle Y$.

3. System Architectures

- **SBF Method [1]**



- **SBF-MTSAL-Concat Method**



4. SBF-MTSAL-Concat Method

- **Magnitude and Temporal Spectrum Approximation Loss**

$$\begin{aligned} J = \frac{1}{T}\sum(&||M \odot |Y| - |X| \odot cos(\theta_y - \theta_x)||_F^2 \\ &+ w_d||f_d(M \odot |Y|) - f_d(|X| \odot cos(\theta_y - \theta_x))||_F^2 \\ &+ w_a||f_a(M \odot |Y|) - f_a(|X| \odot cos(\theta_y - \theta_x))||_F^2) \end{aligned} \quad (4)$$

- **The Concatenation Framework**

The extracted magnitude $|\hat{X}|$ and time domain signal $\hat{x}$ of target speaker are,

$$\begin{aligned} |\hat{X}| &= M \odot |Y| \\ &= G(\sigma([BLSTM(|Y|); g(|A|)])) \odot |Y| \end{aligned} \quad (5)$$

$$\hat{x} = OLA(iSTFT(|\hat{X}| \cdot e^{\angle Y})) \quad (6)$$

5. Discussion

- Unlike speech separation techniques, the number of speakers is not necessary in the speaker extraction.
- Although the target speaker characteristic is needed, this speaker extraction technique is practical to the applications where only registered speakers need to be responded.
- For example, speaker verification application [2].

6. Experiments

- **Closed Condition vs. Open Condition**

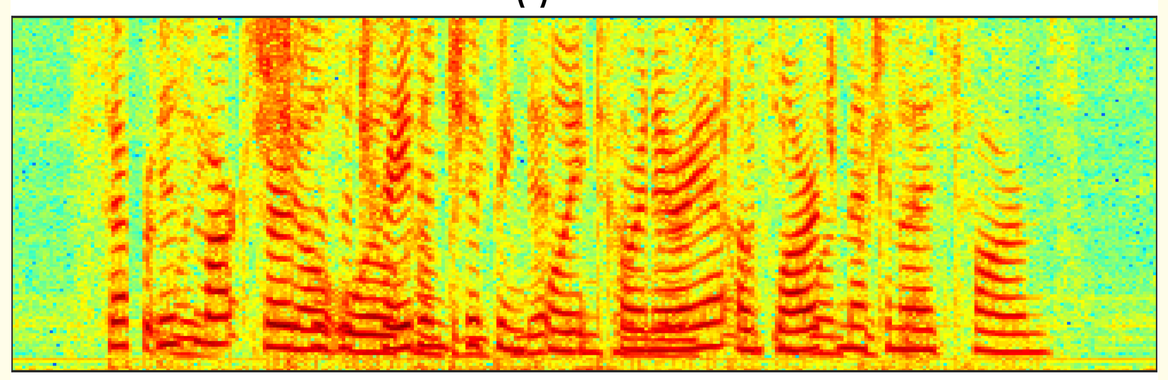
Method	Paras	CC		OC	
		SDR	PESQ	SDR	PESQ
Mixture	-	2.60	2.32	2.60	2.31
SBF [1]	19.3M	6.48	2.30	6.45	2.32
SBF-MTSAL	19.3M	10.36	2.69	9.90	2.66
SBF-MTSAL-Concat	8.9M	**11.39**	**2.77**	**10.99**	**2.73**

- **Different Gender vs. Same Gender**

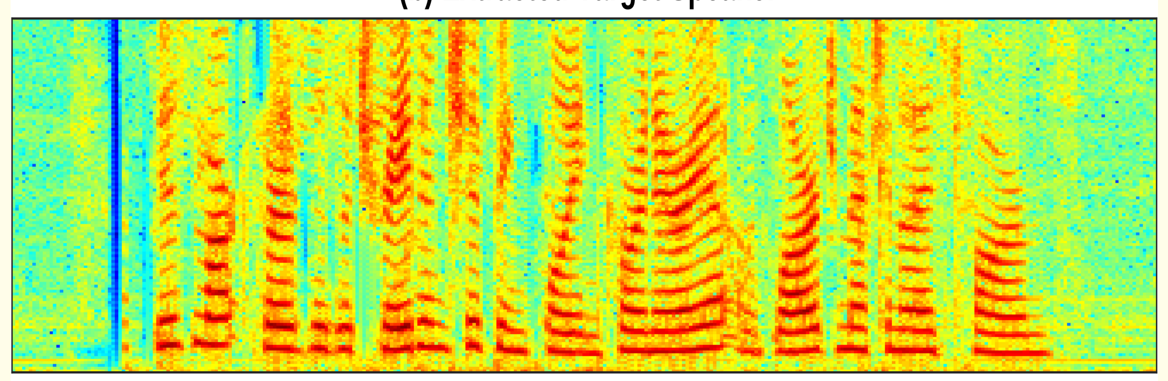
Method	SDR		PESQ	
	Diff.	Same	Diff.	Same
Mixture	2.51	2.69	2.29	2.34
SBF [1]	7.61	5.13	2.42	2.19
SBF-MTSAL	12.27	7.17	2.85	2.44
SBF-MTSAL-Concat	**12.87**	**8.84**	**2.90**	**2.54**

- **Visualization: Female-Female Example**

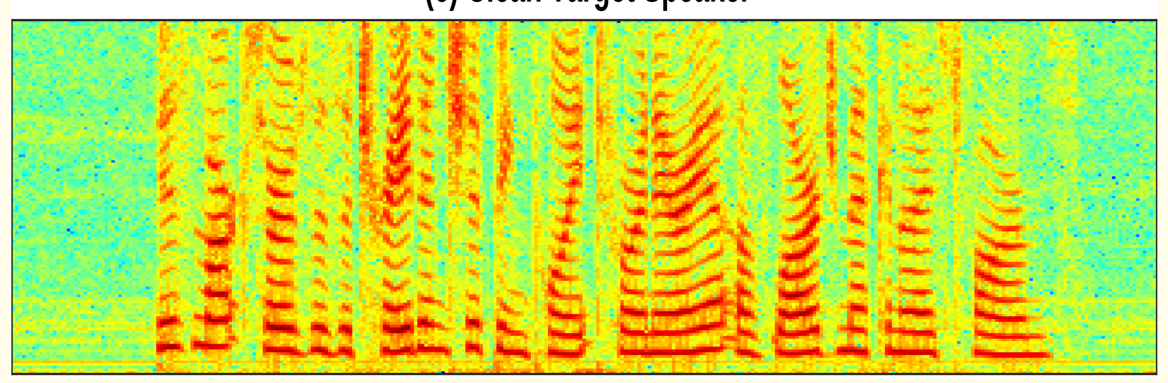
(a) Mixture



(b) Extracted Target Speaker



(c) Clean Target Speaker



7. Acknowledgements

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (and other co-funders, where applicable). [AISG-100E-2018-006].

8. Key References

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa and T. Nakatani, Single Channel Target Speaker Extraction and Recognition with Speaker Beam. In *ICASSP2018*

[2] W. Rao, C. Xu, E. S. Chng and H. Li, Target Speaker Extraction for Multi-Talker Speaker Verification. Submitted to *Interspeech 2019*, arXiv preprint arXiv:1902.02546