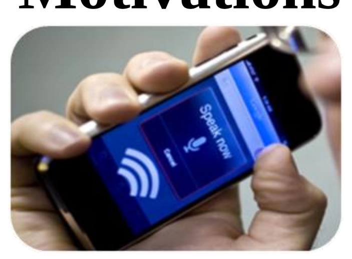
Geometric Information Based Monaural Speech Separation Using Deep Neural Network

Yang Xian, Yang Sun, Jonathon A. Chambers, Syed Mohsen Naqvi
y.xian2@ncl.ac.uk
Newcastle University

Introduction

The performance of deep neural network (DNN) based monaural speech separation methods is limited in reverberant and noisy room environments. Therefore, we propose a new DNN training target which incorporates geometric information describing the target speaker and microphone to improve the performance in reverberant and noisy room environments.

Motivations


Speech recognition


Teleconferencing


Robotic


Hearing aid

Related methods

- Statistical signal processing and computational auditory scene analysis (CASA) based methods [1].
- Deep neural network (DNN) based methods [2][3].

Challenges

- The reverberant and noisy room environments are complex, which increases the difficulty of speech separation.
- The new training target that can better reflect the relation between the clean speech and noisy speech mixture should be developed.

Algorithm

Direct path impulse response

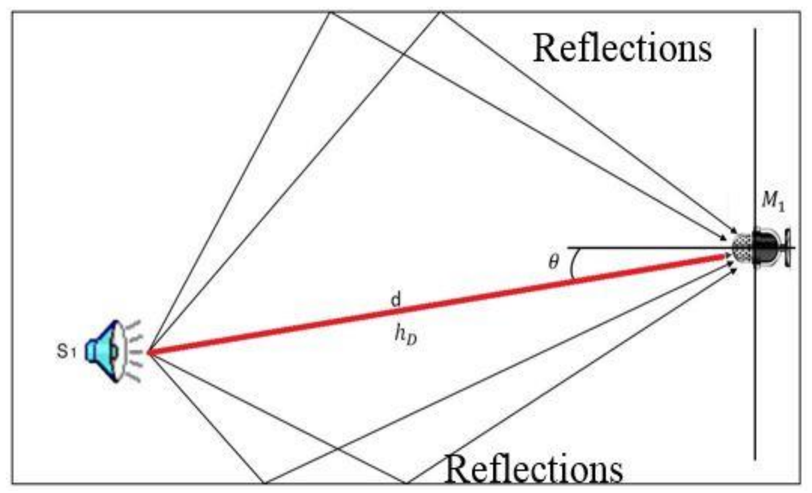
- The reverberant speech mixture can be modelled as:

$$y(t) = s(t) * h(t)$$

- The impulse response can be divided into the direct path and reflections as:

$$h(t) = h_D(t) + h_R(t)$$

- The direct path means the speech is transmitted from speaker to sensors without any reflections.
- The geometric information provides the distance and bearing between the speech source and the microphone, which helps to estimate direct path impulse response.



Monaural speech separation setup within a reverberant room environment

- The attenuation of sound: $\beta = \frac{\kappa}{d^2}$
- The propagation time: $\tau = \frac{f_s}{c} d$
- The direct path impulse response:

$$h_D(t) = \beta\delta(t-\tau) = \frac{\kappa}{d^2}\cos\left(\frac{\theta}{\gamma}\right)\delta\left(t - \frac{f_s}{C} d\right)$$

- The reverberant speech mixture can be represented as:

$$\begin{aligned} y(t) &= s(t) * [h_D(t) + h_R(t)] \\ &= s(t) * h_D(t) + s(t) * h_R(t) \\ &= s_D(t) + s_R(t) \end{aligned}$$

Training target

- The direct path ratio mask (DRM) is defined as:

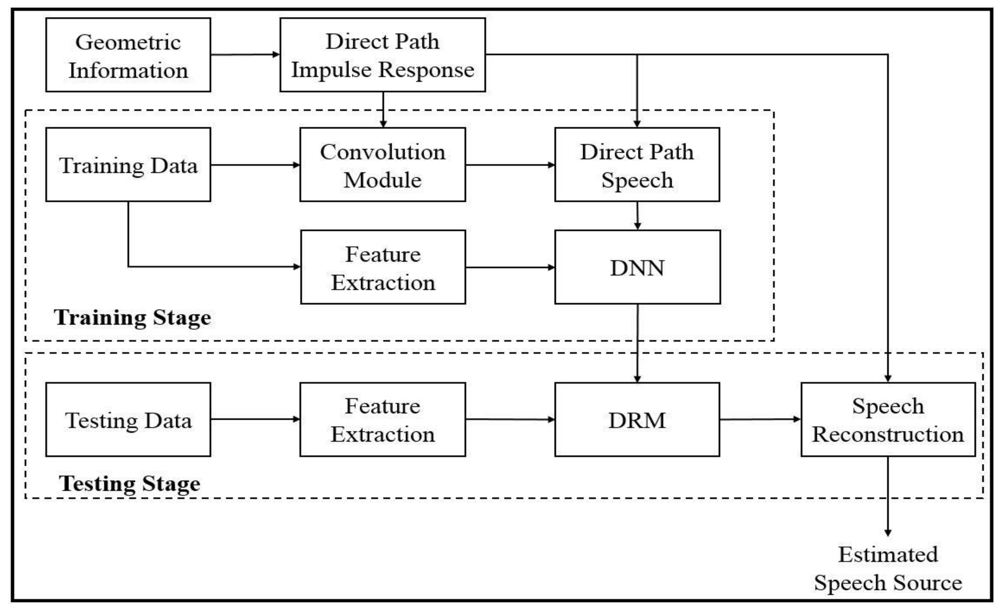
$$DRM(t,f) = \left(\frac{S_D^2(t,f)}{S_D^2(t,f) + N^2(t,f)}\right)^{\eta}$$

- $S_D^2(t,f)$ denotes the energy of the direct path speech at time t and frequency frame f, and $N^2(t,f)$ is the energy of noise. And η is the tunable parameter to scale the mask.
- Advantage: the proposed DRM requires less accuracy in the separation of noisy reverberant speech mixture, because the DRM mitigates reflections and noise.
- The direct path impulse response based speech is estimated as:

$$\hat{S}_D(t,f) = Y(t,f)DRM(t,f)$$

- Then the speech reconstruction is applied to generate the desired speech source.

$$\hat{s}(t) = IFFT\left[\hat{S}_D(t,f)H_D(t,f)^{-1}\right]$$



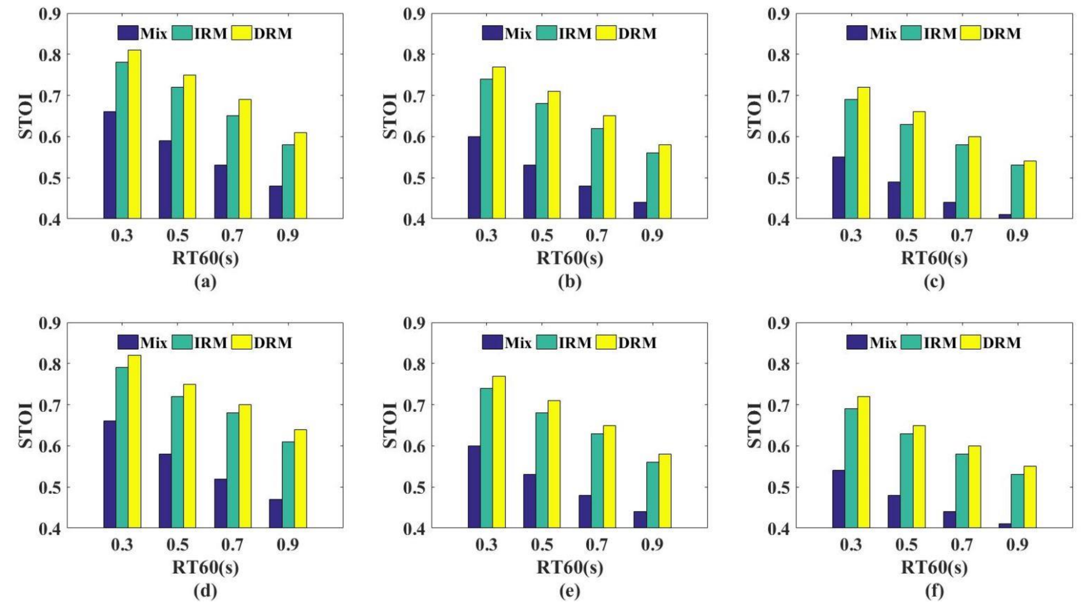
The block diagram of the propose reverberant and noisy speech separation system

Experiments

Settings

- Speech database: IEEE corpus.
- Noise database: NOISEX (Factory noise and Babble noise).
- The direct path impulse response is obtained by using the geometric information.
- Impulse responses: the synthetic and real room impulse responses (RIRs).
- Performance measures: short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ).
- The number of layers: 6 (4 hidden layers).
- The number of units: 1024.
- The activation function of hidden unit: rectified linear unit (ReLU) function.
- Dropout rate:0.2.

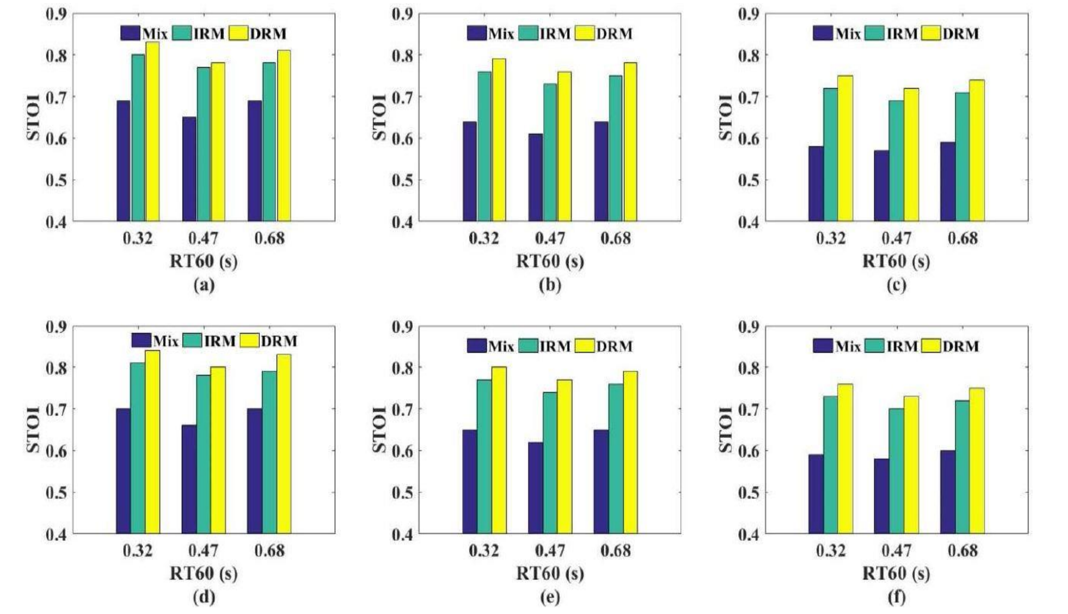
Evaluations with synthetic RIRs



SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	Factory	Babble	Factory	Babble	Factory	Babble
0.3	Unprocessed	0.92	1.06	0.65	0.87	0.48	0.52
	IRM	2.40	2.45	1.95	2.25	1.72	2.03
	DRM	**2.49**	**2.50**	**2.05**	**2.35**	**1.83**	**2.19**
0.5	Unprocessed	0.64	0.83	0.51	0.68	0.45	0.55
	IRM	1.89	2.18	1.69	2.00	1.48	1.83
	DRM	**2.05**	**2.25**	**1.79**	**2.12**	**1.60**	**1.95**
0.7	Unprocessed	0.50	0.64	0.47	0.55	0.44	0.52
	IRM	1.74	1.92	1.55	1.74	1.31	1.62
	DRM	**1.85**	**2.11**	**1.61**	**1.94**	**1.44**	**1.78**
0.9	Unprocessed	0.40	0.60	0.35	0.47	0.31	0.41
	IRM	1.51	1.75	1.32	1.61	1.23	1.46
	DRM	**1.59**	**1.90**	**1.43**	**1.74**	**1.34**	**1.60**

- In terms of PESQ and STOI, the proposed DRM outperforms the ideal ratio mask (IRM) at all RT60s.

Evaluations with real RIRs



SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	Factory	Babble	Factory	Babble	Factory	Babble
0.32	Unprocessed	1.02	1.25	0.74	0.99	0.56	0.78
	IRM	2.31	2.65	2.24	2.51	1.99	2.31
	DRM	**2.42**	**2.70**	**2.37**	**2.57**	**2.11**	**2.39**
0.47	Unprocessed	0.64	0.85	0.49	0.67	0.41	0.57
	IRM	2.17	2.43	1.99	2.31	1.80	2.14
	DRM	**2.28**	**2.53**	**2.11**	**2.40**	**1.89**	**2.21**
0.68	Unprocessed	0.74	0.91	0.69	0.80	0.52	0.61
	IRM	2.21	2.49	2.00	2.24	1.79	2.13
	DRM	**2.33**	**2.51**	**2.11**	**2.42**	**1.92**	**2.22**

- The direct to reverberant ratio (DDR) has positive effect on separation performance.
- The proposed method can separate the target speech from the noisy reverberant mixture in both simulated and real room environments, effectively.

Conclusion and Future Work

- We exploited the geometric information to provide the position of the target speaker and microphone to estimate the direct path impulse response.
- Based on the direct path speech, we calculated the DRM that is a new training target. The experimental results confirmed the DRM outperforms the state-of-the-art IRM based method.
- More effort will be dedicated to improve the proposed method for moving sources.

Selected References

[1] S. M. Naqvi, Yu M and J. A. Chambers , " A multimodal approach to blind source separation of moving sources." IEEE Journal of Selected Topics in Signal Processing 2010, 4(5), 895-910.
[2] Y. Sun, Y. Xian, P. Feng, J. A. Chambers, and S. M. Naqvi, "Estimation of the number of sources in measured speech mixtures with collapsed Gibbs sampling," *Proc. of SSPD*, 2017.
[3] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.