Jing-jing CHEN, Qi-rong MAO, You-cai QIN, Shuang-qing QIAN, Zhi-shen ZHENG, 2020. Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder. *Frontiers of Information Technology & Electronic Engineering*, 21(11):1639-1650. https://doi.org/10.1631/FITEE.2000019

Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder

Key words: Speech separation; Generative factors; Autoencoder; Deep learning

Corresponding author: Qi-rong MAO
E-mail: mao_qr@ujs.edu.cn
ORCID: https://orcid.org/0000-0002-0616-4431

Motivation

1. Existing approaches based on autoencoder directly use a decoder to construct a specific audio source of interest, ignoring the process of reconstructing the original mixed signal.

2. Existing approaches based on autoencoder cannot learn the generative factors of the original input, leading to negative effect on separation performance.

3. Existing approaches based on autoencoder cannot construct each audio source in mixed speech.

Main idea

1. A novel weighted-factor autoencoder (WFAE) model is proposed for monaural speech separation, which can learn source-specific generative factors for each acoustic source.

2. Discriminative features for each source are important, and have significant effects on the monaural speech separation performance.

3. It is important to isolate one source without containing other sources.

Method

1. We propose the WFAE model for monaural speech separation, leading to performance improvement.

2. A latent attention mechanism is incorporated to learn source-specific generative factors and a set of discriminative features for each source.

3. A regularization loss is introduced in the objective function to isolate one source without containing other sources.

Major results

Weighted-factor autoencoder

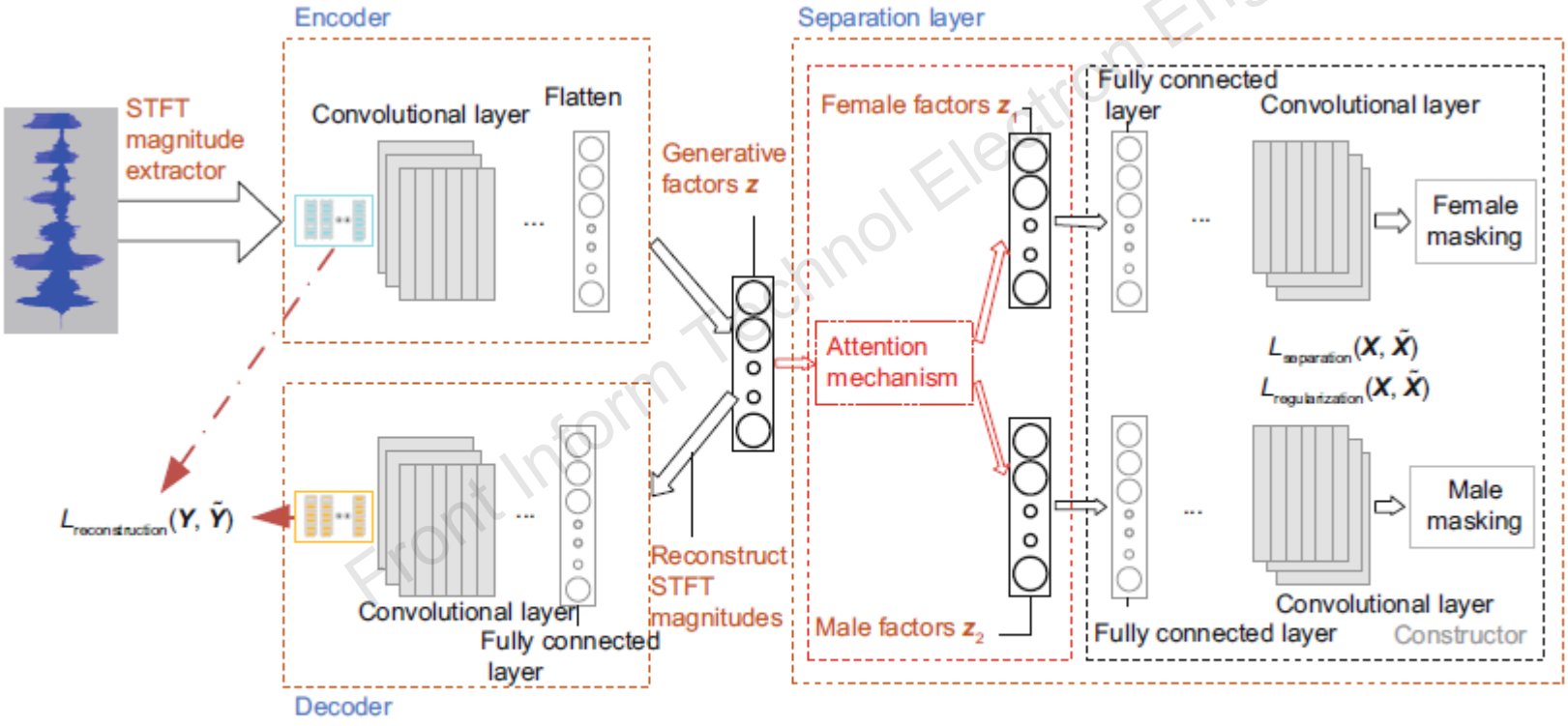

Fig. 1 Architecture of our weighted-factor autoencoder

References to color refer to the online version of this figure

Major results (Cont'd)

1. Test results of our model and related methods

Table 1 Performance comparison on TIMIT

Setting	Model	SDR (dB)	SIR (dB)	SAR (dB)
Speaker-dependent	No-mask	4.25	8.84	6.91
	Binary-mask	3.51	12.50	4.33
	Soft-mask	4.25	8.44	6.91
	Autoencoder	3.01	6.77	6.31
	VAE	6.03	8.80	7.02
	Deep-VAE	6.13	8.85	7.16
	VAE	5.84	10.32	8.21
	Deep-VAE	6.19	10.56	8.61
	*z*AE-MLP	8.40	12.19	11.33
	*z*AE-CNN	9.16	13.42	11.63
	WFAE-no_reg	9.43	13.79	**11.82**
	WFAE	**9.70**	**14.77**	11.69
Speaker-independent	*z*AE-MLP	7.94	11.67	11.00
	*z*AE-CNN	8.34	12.44	11.05
	WFAE-no_reg	8.67	12.88	**11.26**
	WFAE	**8.81**	**13.75**	10.99

The first half of the speaker-dependent results comes from Pandey et al. (2018), and the second half of the speaker-dependent results comes from implemented baselines and our models with our created mixtures (seen speakers). The speaker-independent results come from our models with created unseen speakers. The best results are in bold

Major results (Cont'd)

2. Test results of our model and related methods

Table 2 SDR, SIR, and SAR results on Mini LibriSpeech

Model	SDR (dB)	SIR (dB)	SAR (dB)
Deep-VAE	5.86	12.35	7.29
WFAE	**7.76**	**13.29**	**9.66**

The best results are in bold

Table 3 SDR, SIR, and SAR results on Common Voice CN

Model	SDR (dB)	SIR (dB)	SAR (dB)
Deep-VAE	8.25	14.92	9.44
WFAE	**11.02**	**16.90**	**12.54**

The best results are in bold

Conclusions

1. Generative factor is an important research issue for monaural speech separation.

2. The proposed weighted-factor autoencoder can learn source-specific generative factors and a set of discriminative features for each source, leading to performance improvement.

3. In terms of three important metrics, the weighted-factor autoencoder has achieved great success on a relatively challenging MSS case, i.e., speaker-independent monaural speech separation.

Qi-rong MAO received her MS and PhD degrees from Jiangsu University, Zhenjiang, China in 2002 and 2009, respectively, both in Computer Application Technology. She is currently a professor at the School of Computer Science and Communication Engineering, Jiangsu University. Her research interests include affective computing, pattern recognition, and multimedia analysis. She has published over 40 technical articles. She is a member of the IEEE.