INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Robust Speaker Extraction Network based on Iterative Refined Adaptation

Chengyun Deng[1], *Shiqian Ma*[1], *Yongtao Sha*[1], *Yi Zhang*[1],
Hui Zhang[2], *Hui Song*[1], *Fei Wang*[1]

[1]Didi Chuxing, Beijing, China
[2]Baidu, Inc., Beijing, China

{dengchengyun, mashiqian, warrenwangfei}@didiglobal.com
zhanghui41@baidu.com

Abstract

Speaker extraction aims to extract target speech signal from a multi-talker environment with interference speakers and surrounding noise, given a reference speech from target speaker. Most speaker extraction systems achieve satisfactory performance in the closed condition. Such systems suffer from performance degradation given unseen target speakers and/or mismatched reference speech. In this paper we propose a novel strategy named Iterative Refined Adaptation (IRA) to improve the robustness and generalization capability of speaker extraction systems in the aforementioned scenarios. Given an initial speaker embedding encoded by an auxiliary network, the extraction network can obtain a latent representation of the target speaker as the feedback of the auxiliary network to refine the speaker embedding, which provides more accurate guidance for the extraction network. Experiments show that the network with IRA confirm the superior performance over comparison approaches in terms of SI-SDRi and PESQ on WSJ0-2mix-extr and WHAM! dataset.

Index Terms: speaker extraction, iterative refined adaptation, speaker embedding, robustness

1. Introduction

Auditory attention allows human to focus on a specific speaker in a crowd, which is also called the cocktail party effect [1]. Speaker extraction aims to extract the target speaker signal in a challenging acoustic environment according to the reference speech. Speaker extraction plays an important role in improving the intelligibility of speech.

Speaker extraction algorithms can be roughly divided into three categories, i.e., time-frequency (T-F) approaches, hybrid approaches and complete time-domain approaches. T-F methods aim to encode the target speaker information using T-F features. They estimate the target speaker's magnitude spectrum guided by the speaker embedding [2, 3, 4], and then rebuild the waveform via inverse short-time Fourier transform (ISTFT) by combining the extracted magnitude spectrum with potentially mismatched phase from the mixture, such as Voicefilter [2] and SBF-MTSAL-Concat [5]. Hybrid methods are then proposed integrating a T-F based speaker embedding with a time-domain speech separation network, such as SpEx [6]. Furthermore, to avoid the mismatch of latent feature space between speech encoder and speaker encoder, a complete time-domain method SpEx+ [7] is proposed, which using multi-scale weight-shared CNN speech encoders to projects mixture and reference speech into a common latent space. Time-domain methods have showed competitive performance on speech separation tasks in recent years.

Recent studies on time-domain audio separation network with encoder-separator-decoder architecture have been significantly progressed in speaker-independent speech separation tasks, such as TasNet[8], Conv-TasNet [9], DPRNN-TasNet [10] and DPTNet [11]. Compared with T-F approaches [12, 13], time-domain separation methods can bypass the difficulty of phase reconstruction and avoid the long latency effect caused by large window size of T-F analysis [9].

Numerous speaker extraction systems achieve good performance, in closed condition, with seen target speakers and matched reference speech. However, in realistic circumstances, i.e. open condition, unseen target speakers and/or mismatched reference speech have an adverse effect on speaker extraction systems. How to improve the robustness and generalization capability of speaker extraction networks in open condition is still an open question. Motivated by this phenomenon, we propose an IRA strategy to make the extraction procedure more reliable.

Inspired by SpEx+ [7], we present a complete time-domain speaker extraction network named DPRNN-Spe, which integrate DPRNN-TasNet [10] with a time-domain auxiliary residual network. For the purpose of refining the mismatched reference speaker embedding, we furtherly apply Iterative Refined Adaptation (IRA) strategy, is called DPRNN-Spe-IRA. Firstly, we use the original target speaker embedding encoded by auxiliary network to extract target latent representations from the mixture. Secondly, auxiliary network re-encodes the latent representations to obtain a refined speaker embedding, which is then combined with the original one to form a new embedding through a linear layer. We extract the target speech using the new embedding correspondingly. Finally, we reconstruct the waveform of the target speaker by the decoder. Experiments on WSJ0-2mix-extr [14] show that DPRNN-Spe-IRA achieves 1.05 dB and 0.58 dB gain, in terms of SI-SDRi, over DPRNN-Spe and the state-of-the-art SpEx+ correspondingly. Moreover, DPRNN-Spe-IRA yields better extraction performance in noisy scenarios as well.

This paper is organized as follows. In Section 2, give an overview of DPRNN-Spe network. In Section3, go into IRA strategy. In Section 4, we report the experiments. Section 5 concludes the study.

2. DPRNN-Spe Architecture

Single-channel speaker extraction system can be formulated in terms of extracting speech of target speaker from the mixtures, where $\boldsymbol{x}(t)$, is the observation signal which is the mixture of the target source $\boldsymbol{s}_{target}(t)$ and $C-1$ interference sources $\boldsymbol{s}_i(t)$.

$$\boldsymbol{x}(t) = \boldsymbol{s}_{target}(t) + \sum_{i=1}^{C-1} \boldsymbol{s}_i(t) \tag{1}$$

According to [10], DPRNN uses global information and

http://dx.doi.org/10.21437/Interspeech.2021-2250

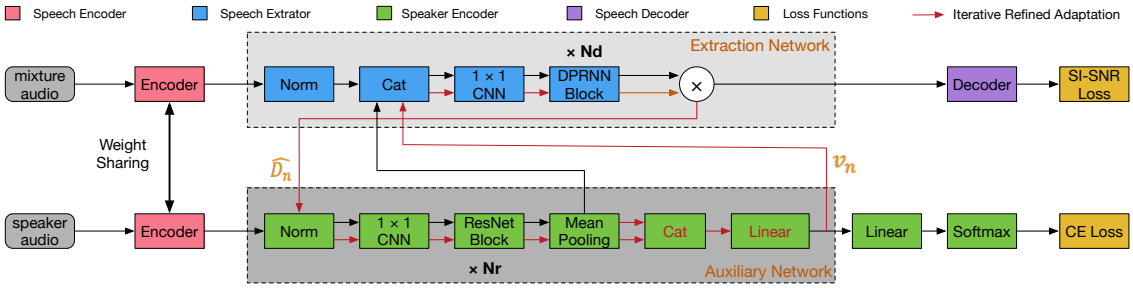


Figure 1: *The architecture of the proposed speech extraction system. For IRA, represent by red arrow line, the speaker embedding* $\boldsymbol{v}_n$ *encoded by the auxiliary network and the representation of target speaker* $\hat{\boldsymbol{D}}_n$ *extracted by the extraction network iteratively refine and adapt to each other.*

achieves superior performance with smaller model size. Thus we use it as our extraction network. As shown in Figure 1, the proposed DPRNN-Spe system consists of twin speech encoders, an extraction network, a decoder and an auxiliary network. The difference between DPRNN-Spe and SpEx+ lies in three aspects, (i) we use single-scale rather than multi-scale encoders to reduce complexity of the model, (ii) the speaker embedding is concatenated just once instead of repeatedly before each single DPRNN block, and (iii) we use DPRNN rather than TCN for the extraction network.

2.1. Twin speech encoders

The mixture speech $\boldsymbol{x} \in \mathbb{R}^{1\times T_1}$ and reference speech $\boldsymbol{r} \in \mathbb{R}^{1\times T_2}$ can be transformed to latent representations $\boldsymbol{MixE} \in \mathbb{R}^{K_1\times N}$ and $\boldsymbol{SpeE} \in \mathbb{R}^{K_2\times N}$ respectively. K_1 and K_2 are number of overlapping segments, and N is feature dimension.

$$\begin{aligned} \boldsymbol{MixE} &= \text{Enc}(\boldsymbol{x}) \\ \boldsymbol{SpeE} &= \text{Enc}(\boldsymbol{r}) \end{aligned} \tag{2}$$

where $\text{Enc}(\cdot)$ represents the single-scaled twin speech encoders, which are weight-shared 1-D CNNs [7] with window length L and stride length $L/2$.

2.2. Auxiliary network

We extract a highly discriminative target speaker embedding $\boldsymbol{v} \in R^E$, E is the speaker embedding dimension,

$$\boldsymbol{v} = \text{Aux}(\boldsymbol{SpeE}) \tag{3}$$

where $\text{Aux}(\cdot)$ represents the auxiliary network. ResNet is commonly used in speaker recognition tasks [15, 16], we use it as the core of the auxiliary network. The architecture of ResNet block is similar with SpEx+, but has smaller size since we don't use multi-scale encoders.

The training of auxiliary network can be seen as a sub-task in a multi-task learning of DPRNN-Spe, which is a speaker classification task. A cross-entropy loss $\mathcal{L}_{CE}$ [7] is used,

$$\begin{aligned} \mathcal{L}_{CE} &= -\sum_{i=1}^{N_s} I_i log(\hat{p}_i) \\ \hat{p}_i &= \sigma(\boldsymbol{W_l}^\top \boldsymbol{v} + \boldsymbol{b})_i \end{aligned} \tag{4}$$

where N_s is the total number of speakers in training data, I_i is the true class label for the i^{th} speaker, $\hat{p}_i$ is the corresponding predicted probability, $\sigma(\cdot)$ is a soft-max activation function. $\boldsymbol{W_l}$ is a weight matrix of FC and $\boldsymbol{b}$ is the bias .

2.3. Extraction network

The speaker embedding is repeatedly concatenated along the feature dimension of the normalized $\boldsymbol{MixE}$ before the first DPRNN block. The speaker extraction network $\text{Ext}(\cdot)$ predicts a latent representation $\hat{\boldsymbol{D}} \in \mathbb{R}^{K_1\times N}$ of the target speaker by learning a mask $\boldsymbol{M} \in \mathbb{R}^{K_1\times N}$ in this latent space,

$$\begin{aligned} \boldsymbol{M} &= \text{Ext}([\boldsymbol{v} : norm(\boldsymbol{MixE})]) \\ \hat{\boldsymbol{D}} &= \boldsymbol{MixE} \odot \boldsymbol{M} \end{aligned} \tag{5}$$

where $[:]$ is the concatenate operation, $\odot$ denotes element-wise multiplication and $norm$ is for layer normalization [17].

2.4. Speech decoder

Finally, the overlapping-segment waveform of the target source $\hat{\boldsymbol{S}} \in \mathbb{R}^{K_1\times L}$ is reconstructed by the decoder which consists of 1-D transposed convolution. The overlapping reconstructed matrix generates the final waveform $\hat{\boldsymbol{s}}_{target} \in \mathbb{R}^{1\times T_1}$ by overlap and sum operation.

$$\begin{aligned} \hat{\boldsymbol{S}} &= \hat{\boldsymbol{D}}\boldsymbol{V} \\ \hat{\boldsymbol{s}}_{target} &= \text{OverlapAndSum}(\hat{\boldsymbol{S}}) \end{aligned} \tag{6}$$

where $\boldsymbol{V} \in \mathbb{R}^{N\times L}$ is the decoder basis functions matrix.

2.5. Multi-Task Learning

The main objective of DPRNN-Spe training is to maximize the scale-invariant source-to-distortion ratio (SI-SDR) [18] of the target speaker, with the assistance of high-quality speaker embedding. SI-SDR is defined as,

$$\begin{cases} \boldsymbol{s}_{target} := \frac{<\hat{\boldsymbol{s}}, \boldsymbol{s}>\boldsymbol{s}}{\|\boldsymbol{s}\|^2} \\ \boldsymbol{e}_{noise} := \hat{\boldsymbol{s}} - \boldsymbol{s}_{target} \\ SI\text{-}SDR := 10log_{10}\frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \end{cases} \tag{7}$$

where $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ are the predicted and original clean sources respectively, and $\|\boldsymbol{s}\|^2$ denotes the signal power.

Meanwhile, $\mathcal{L}_{CE}$ is used as a regularization to guide the auxiliary network training. Hence the multi-task learning loss of DPRNN-Spe is defined as:

$$\mathcal{L}_{DPRNN\_Spe} = \mathcal{L}_{SI\text{-}SDR} + \lambda\mathcal{L}_{CE} \tag{8}$$

where λ is a scaling parameter.

3. Iterative Refined Adaptation

3.1. Robustness problems

Speaker extraction systems often yield inferior results in open conditions. Apart from unseen target speakers, mismatch reference problems also exist, e.g. individual speaker characteristics may change due to factors of age, physical health, mood, speaking rate, etc. Besides, the mixture and reference speech may be recorded in different acoustic environments or different channels, making the reference voiceprint information misleading to some extent. That is to say, speaker embedding would vary over time and environments, it cast negative effects on speaker extraction tasks.

3.2. Formulation of iterative refined adaptation

We propose a training strategy called IRA to eliminate such adverse effects. Suppose we have mixture input $\boldsymbol{x}$, an initial mismatched reference $\boldsymbol{r}$, a main speaker extraction function $\mathbf{F}$ and an auxiliary function $\mathbf{A}$. Through $\mathbf{F}$ function, we obtain a rough extraction result $\hat{\boldsymbol{s}}_{target_0}$ from $\boldsymbol{x}$ on the basis of condition $\boldsymbol{a}_0 = \mathbf{A}(\boldsymbol{r})$. Then we feed $\hat{\boldsymbol{s}}_{target_0}$ back to the auxiliary function to produce a refined reference information $\boldsymbol{a}_1$, which is then fed back into $\mathbf{F}$ to produce a more accurate result $\hat{\boldsymbol{s}}_{target_1}$. After n times of feedback and modification, the network can deliver a more matched condition $\boldsymbol{a}_n$ and a more accurate result $\hat{\boldsymbol{s}}_{target_n}$

$$\begin{aligned} \boldsymbol{a}_n &= \boldsymbol{a}_{n-1} + \mu\mathbf{A}(\hat{\boldsymbol{s}}_{target_{n-1}}) \\ \hat{\boldsymbol{s}}_{target_n} &= \mathbf{F}(\boldsymbol{x} \mid \boldsymbol{a}_n) \end{aligned} \tag{9}$$

where μ is a scaling parameter.

3.3. Speaker extraction system with IRA

IRA strategy can be applied to any speaker extraction systems easily. In this paper, we validate its effectiveness by DPRNN-Spe and DPRNN-Spe-IRA, illustrated in Figure 1.

First of all, we use the original reference speech to obtain initial embedding $\boldsymbol{v}_0$. Then, we get the estimated representation $\hat{\boldsymbol{D}}_0$ of the target speaker from the extraction network condition on the embedding $\boldsymbol{v}_0$ and $\boldsymbol{MixE}$. Next, we feed $\hat{\boldsymbol{D}}_\mathbf{0}$ back to the auxiliary network to produce a new embedding $\boldsymbol{v}_1 = \text{Aux}(\hat{\boldsymbol{D}}_0)$, and concatenate it together with $\boldsymbol{v}_0$. The additional FC layer is used to transform the $\boldsymbol{v}_1$ feature dimension back to before. And then this refined embedding $\boldsymbol{v}_1$ is fed back into the extraction network to produce a better extracted representation $\hat{\boldsymbol{D}}_1$. Repeating n times above steps, we can get more matched embedding $\boldsymbol{v}_n$, more accurate mask $\boldsymbol{M}_n$ and leads to a better representation $\hat{\boldsymbol{D}}_n$,

$$\begin{aligned} \boldsymbol{v}_n &= \boldsymbol{W}^\top([\boldsymbol{v}_{n-1} : \text{Aux}(\hat{\boldsymbol{D}}_{\boldsymbol{n-1}})]) + \boldsymbol{b} \\ \boldsymbol{M}_\boldsymbol{n} &= \text{Ext}([\boldsymbol{v}_n : norm(\boldsymbol{MixE})])) \\ \hat{\boldsymbol{D}}_\boldsymbol{n} &= \boldsymbol{MixE} \odot \boldsymbol{M}_\boldsymbol{n} \end{aligned} \tag{10}$$

where $\boldsymbol{W} \in \mathbb{R}^{2E\times E}$ and $\boldsymbol{b} \in \mathbb{R}^E$ are the weights and bias of the FC layer respectively.

4. Experimental Evaluation

4.1. Dataset

We evaluate the speech extraction performance using WSJ0-2mix-extr[14], WSJ0-2mix[19] and WHAM![20] dataset with 8kHz sample rate. WSJ0-2mix-extr and WSJ0-2mix are noise-free, and WHAM! contains numerous real ambient noise samples. In WSJ0-2mix-extr, speaker $s1$ is the target speaker, the reference speech of the target speaker is randomly selected in clean speech except for the one in the mixture. We simulate a clean and a noisy speaker extraction dataset based on WSJ0-2mix and WHAM! respectively. We choose each speaker as target speaker in turn, and the selection of reference speech is the same as the WSJ0-2mix-extr.

These three datasets contain 30 hours of training data with 101 speakers, 10 hours of validation data with the same 101 speakers and 5 hours of evaluation data with 18 different speakers.

Table 1: *SDRi (dB), SI-SDRi (dB) and PESQ of extracted speech for different systems on the WSJ0-2Mix-Extr and WHAM!. Best scores are highlighted in* ***bold****. L is the filter length of the encoder. DPRNN-Spe-2IRA for applying IRA strategy twice.*

Algorithms	Params	SI-SDRi	SDRi	PESQ
WSJ0-2Mix-Extr				
Mixture	-	0.00	0.00	2.31
$L = (20, 80, 160)$				
SpEx [6]	10.80M	14.18	14.55	3.36
SpEx+ [7]	13.30M	15.70	15.94	3.49
$L = 16$				
DPRNN-Spe	2.91M	15.23	15.48	3.42
DPRNN-Spe-IRA	2.94M	**16.28**	**16.53**	**3.53**
DPRNN-Spe-2IRA	2.94M	**16.45**	**16.70**	**3.54**
$L = 8$				
DPRNN-Spe	2.90M	15.94	16.27	3.50
DPRNN-Spe-IRA	2.94M	**17.50**	**17.73**	**3.62**
WHAM!				
Mixture	-	0.00	0.00	1.66
$L = 16$				
SpEx+ [7]	13.30M	13.12	13.66	2.46
DPRNN-Spe	2.91M	13.17	13.78	2.48
DPRNN-Spe-IRA	2.94M	**14.15**	**14.61**	**2.57**

4.2. Training and evaluation setup

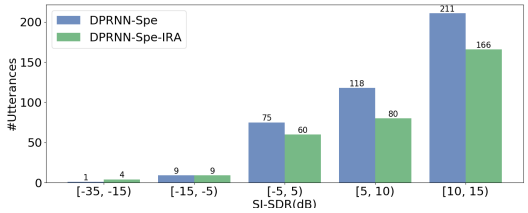
For DPRNN-Spe, we use the same encoder and decoder design as in [10]. The number of ResNet blocks in auxiliary network is set to 3 ($N_r = 3$), and dimension of the speaker embedding is set to 128. Similar to DPRNN-TasNet [10], our extraction network includes 6 DPRNN blocks, where BLSTM [21] is used as the intra- and inter-chunk RNNs, and each direction has 128 hidden units. For DPRNN-Spe-IRA, the additional FC layer has 128 linear node. The hyper-parameter that control the CE loss of the speaker classifier is set as $\lambda = 0.5$.

All models are trained using Adam optimizer [22] for 100 epochs on 4-second long segments with an initial learning rate of 0.0005 and using a batchsize of 12 for $L = 16$ and batchsize of 8 for $L = 8$. The learning rate is divided by 2 if the validation loss does not improve for 2 epochs.

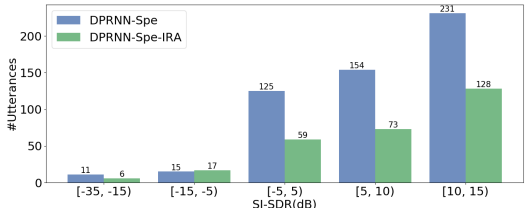
We use the SI-SDR improvement (SI-SDRi), signal-to-distortion ratio improvement (SDRi) [23] and perceptual evaluation of speech quality (PESQ) [24] as objective measures of extraction accuracy.

4.3. Comparative study on WSJ0-2mix-extr

Table 1 shows the results for the different algorithms. SpEx+ is the state-of-the-art method of speaker extraction. In $L = 16$ case, the proposed DPRNN-Spe algorithm with single-scale encoders and a smaller auxiliary network has matching performance with SpEx+. Comparing with DPRNN-Spe in terms of SI-SDRi and PESQ, DPRNN-Spe-IRA results in 6.91% and



(a) validation data



(b) test data

Figure 2: *Distributions of the number of utterances with SI-SDR lower than 15dB, when L=16. The smaller number of utterances with low SI-SDR suggests better performance of extraction. Validation data for closed condition, test data for open condition.*

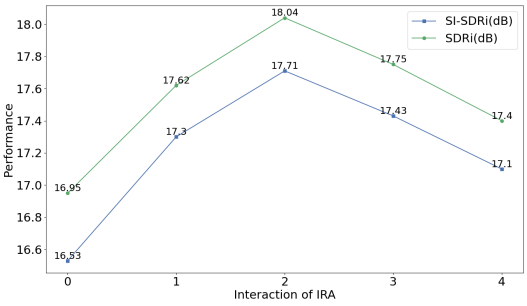


Figure 3: *IRA iteration study on WSJ0-2mix. $L = 16$ in all systems.*

3.32% relative improvement, DPRNN-Spe-2IRA which using IRA twice (correspond to $\hat{D}_2$), results in 8.03% and 3.51% relative improvement. DPRNN-Spe-IRA significantly outperforms SpEx and SpEx+ with relative improvements of 14.81% and 3.69% in terms of SI-SDRi respectively. That is, IRA can further improve the robustness of speaker extraction model and more iteration of IRA can yield more robust model. In $L = 8$ case, DPRNN-Spe-IRA leads to 9.78% improvement of SI-SDRi comparing with DPRNN-Spe, and 11.48% comparing with SpEx+.

In Figure 2, we further compare the robustness of DPRNN-Spe and DPRNN-Spe-IRA in closed and open condition under $L = 16$ and SI-SDR lower than 15dB separately. In closed condition, average SI-SDR is 19.41 dB and 19.51 dB for DPRNN-Spe and DPRNN-Spe-IRA respectively, and the bad cases number of DPRNN-Spe-IRA has 22.94% relative reduction compared with DPRNN-Spe. In open condition, average SI-SDR is 17.65 dB and 18.79 dB for DPRNN-Spe and DPRNN-Spe-IRA respectively, and the bad cases number of DPRNN-Spe-IRA has 47.20% relative reduction compared with DPRNN-Spe.

Table 2: *IRA strategy study on WSJ0-2mix. $L = 16$ in all systems.*

Algorithms	Params	SI-SDRi	SDRi	PESQ
Mixture	-	0.00	0.00	2.01
DPRNN-Spe(s)	2.91M	**17.57**	**17.82**	**3.46**
DPRNN-Spe	2.91M	16.53	16.95	3.37
DPRNN-Spe-IRA	2.94M	**17.30**	**17.62**	**3.43**
DPRNN-Unfold-Block	5.83 M	16.43	16.89	3.37
DPRNN-Spe-2IRA	2.94M	**17.71**	**18.04**	**3.48**

The IRA strategy iteratively fuse the target reference's speaker characters with the extracted target latent representation from mixture speech, and refine on target speaker representation, decrease the distortion. IRA can improve robustness of DPRNN-Spe regardless of environment, and get greater improvement under open condition. IRA has great application significance for real scenes.

4.4. Comparative study on WHAM!

We further verify whether DPRNN-Spe and DPRNN-Spe-IRA are profitable in a noisy environment. Table 1 show the DPRNN-Spe-IRA achieves 7.44% and 7.85% relative gain in terms of SI-SDRi over DPRNN-Spe and SpEx+ [1] respectively. In summary, the proposed IRA can further improve the robustness in noisy environment.

4.5. IRA iteration study on WSJ0-2mix

Moreover, we test the iteration steps of IRA on WSJ0-2mix. As showed in Figure 3, performances increase first and then decreases. Two iteration of IRA yields best performance with 1.18 dB gain over zero iteration in terms of SI-SDRi. Considering the computational complexity, we prefer to use IRA strategy once.

4.6. IRA strategy study on WSJ0-2mix

As shown in Table 2, we take the DPRNN-Spe as the baseline. DPRNN-Spe(s) using target speech as reference speech, observe 1.04 dB gain on SI-SDRi compared to DPRNN-Spe, which verifies that more matching embedding better performance. DPRNN-Unfold-Block is that the blocks of the extraction and auxiliary networks are stacked instead of IRA. DPRNN-Unfold-Block yields worse performance compared to DPRNN-Spe, indicating that simply increasing model size does not guarantee better performance and may lead to over-fit problem on small dataset. DPRNN-Spe-IRA evidently improves the SI-SDR by 0.77 dB compared to DPRNN-Spe. These results verify the effectiveness of the proposed IRA. DPRNN-Spe-2IRA get better performance than DPRNN-Spe(s) which is the upper bound exactly, indicate that IRA not only reduces mismatch but also improve target speech extraction ability.

5. Conclusions

This paper introduces a novel IRA strategy for robust speaker extraction tasks. We propose to use a smaller extraction network DPRNN-Spe. Then IRA is involved to improve the extraction performance. IRA can be applied to any extraction networks easily. Experimental results confirm that IRA makes the extraction network more robust for both noise-free and noisy environments. As future works, we will explore IRA on more networks and reduce the complexity of models with IRA.

[1] Our implement is based on https://github.com/gemengtju/SpExPlus.

6. References

[1] S. Getzmann, J. Jasny, and M. Falkenstein, "Switching of auditory attention in "cocktail-party" listening: Erp evidence of cueing effects in younger and older adults," *Brain and cognition*, vol. 111, pp. 1–12, 2017.

[2] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[3] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5554–5558.

[4] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 86–90.

[5] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6990–6994.

[6] ——, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 327–334.

[7] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *arXiv*, pp. arXiv–2005, 2020.

[8] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," *CoRR*, vol. abs/1711.00541, 2017. [Online]. Available: http://arxiv.org/abs/1711.00541

[9] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[11] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[12] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 711–715.

[13] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[14] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[15] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[16] N. Li, D. Tuo, D. Su, Z. Li, D. Yu, and A. Tencent, "Deep discriminative embeddings for duration robust speaker verification." in *Interspeech*, 2018, pp. 2262–2266.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[19] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[20] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.