TEA-PSE: TENCENT-ETHEREAL-AUDIO-LAB PERSONALIZED SPEECH ENHANCEMENT SYSTEM FOR ICASSP 2022 DNS CHALLENGE

Yukai Ju[1,2], *Wei Rao*[2], *Xiaopeng Yan*[1,2], *Yihui Fu*[1], *Shubo Lv*[1,2], *Luyao Cheng*[1], *Yannan Wang*[2], Lei Xie[1], Shidong Shang[2]

[1] Audio, Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University, Xi'an, China
[2] Tencent Ethereal Audio Lab, Tencent Corporation, Shenzhen, China

ABSTRACT

This paper describes Tencent Ethereal Audio Lab – Northwestern Polytechnical University personalized speech enhancement (TEA-PSE) system submitted to track 2 of the ICASSP 2022 Deep Noise Suppression (DNS) challenge. Our system specifically combines the dual-stage network which is a superior real-time speech enhancement framework with the ECAPA-TDNN speaker embedding network which achieves state-of-the-art performance in speaker verification. The dual-stage network aims to decouple the primal speech enhancement problem into multiple easier sub-problems. Specifically, in stage 1, only the magnitude of the target speech is estimated, which is incorporated with the noisy phase to obtain a coarse complex spectrum estimation. To facilitate the formal estimation, in stage 2, an auxiliary network serves as a post-processing module, where residual noise and interfering speech are further suppressed and the phase information is effectively modified. With the asymmetric loss function to penalize over-suppression, more target speech is preserved, which is helpful for both speech recognition performance and subjective sense of hearing. Our system reaches 3.97 in overall audio quality (OVRL) MOS and 0.69 in word accuracy (WAcc) on the blind test set of the challenge, which outperforms the DNS baseline by 0.57 OVRL and ranks 1st in track 2.

Index Terms— Personalized speech enhancement, two-stage network, ECAPA-TDNN, real-time.

1. INTRODUCTION

Personalized speech enhancement (PSE), also called speaker extraction, aims to extract the target speaker's speech from a complicated multi-talker noisy and reverberant observed signal by target speaker's enrollment speech. PSE is very useful when the system is expected to respond to specific target speaker. It can be widely applied to real-time communication (RTC), speaker diarization, automatic speech recognition, etc.

The latest ICASSP 2022 DNS challenge [1] aim to promote the full-band real-time speech enhancement task. Besides the perceptual speech quality requirement, DNS also adopts the word accuracy (WAcc) as an important evaluation metric for back-end ASR applications. Aiming at real-time full-band speech communication, there are two tracks in the challenge – non-personalized DNS (track 1) and personalized DNS (track 2). This paper mainly focuses on track 2. Different from non-real-time PSE methods [2, 3, 4, 5, 6, 7], real-time PSE methods need to specifically consider the requirement of model size, inference time, and limited future information. In recent years, many real-time PSE methods were proposed, such as Voicefilter-lite [8], pDCCRN [9], personalized PercepNet [10], and so on, leading to superior performance. However, the environment of real application is complicated and variable. The algorithms not only need to deal with noise and reverberation, but also have to consider interfering speakers. Although with the help of speaker information, the ability of a single real-time PSE model is still limited with imperfect noise and interference suppression. Recently, multi-stage approaches have been introduced with great success in speech enhancement [11, 12]. In a multi-stage approach, each stage model only focuses on a single task, which is usually guided by one explicit loss function. After the pre-processing from the former stage model, the latter stage model will have a more clear and simple enhanced input and allocate more accurate computing power on its own task. In [11], the first stage network is designed to estimate the magnitude with noisy phase, and the second stage network aims to estimate the residual real and imaginary part of the first stage output. In SpEx++ [13], the output of the first stage network is first sent to a speaker encoder network to get another new speaker embedding, and then the enhanced speech and the two speaker embeddings are fed to the second stage network to further suppress the residual noise and interfering speech.

Inspired by the success of the multi-stage approach, this paper investigates its feasibility in the personalized speech enhancement task. Specifically, in our approach, the first stage aims to estimate the target speech's magnitude coarsely, and the second stage aims to further suppress the residual noise and interfering speech as well as modify the phase information of the target speech. We also explore the effect of power compression (PC) [14] for the PSE task. Besides, considering the ASR evaluation metric in the challenge, asymmetric (Asym) loss [9] is particularly adopted in our TEA-PSE submission system to alleviate the effect of over-suppression that may lead to more speech recognition errors. Our TEA-PSE submission system finally achieves 4.19 SIG, 4.55 BAK, 3.97 OVRL, and 0.69 WAcc on the blind test set of the ICASSP 2022 DNS Challenge, which came in first place in track 2.

2. PROPOSED TEA-PSE SYSTEM

Our proposed TEA-PSE system is mainly composed of two modules: speaker encoder and speech enhancement. Specifically, we adopt the ECAPA-TDNN network [15] as speaker encoder and investigate a dual-stage network for speech enhancement. We first train the speaker encoder. Once well-trained, the speaker encoder network will be frozen and used to extract speaker embeddings for the speech enhancement network to conduct the PSE task.

2.1. Speaker Encoder: ECAPA-TDNN

In a PSE system, speaker embedding is used to identify the target speaker in the observed signal to distinguish the target

Work done during an internship at Tencent Ethereal Audio Lab.

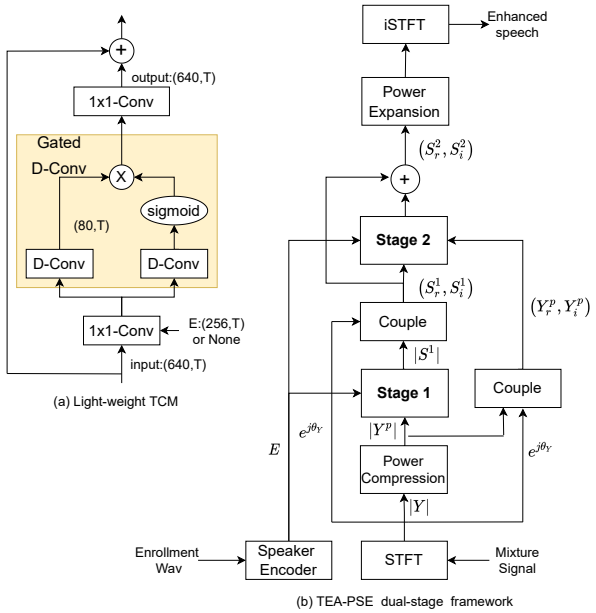

Fig. 1. (a): Light-weight TCM. (b): TEA-PSE dual-stage framework for personalized speech enhancement. The notation are in Section 2.2

speech from all interfering sounds and noises. We take ECAPA-TDNN as our speaker encoder network in the proposed TEA-PSE system. ECAPA-TDNN is known as one of the state-of-the-art speaker embedding networks, thanks to the outstanding design on 1D Res2Net [16] with a squeeze-excitation module [17]. In our ECAPA-TDNN, we adopt 2048 channels in the convolutional frame layers and 256 dimensions in the channel and context-dependent statistics pooling layer. The dimension of the bottleneck in the SE-block is set to 128. Finally, 256-dim speaker embedding is obtained. More details on ECAPA-TDNN can be found in [15].

2.2. Speech Enhancement: Dual-stage Network

Inspired by the success of the multi-stage network [11, 12] in speech enhancement, we adopt a dual-stage network for the PSE task in the time-frequency domain. In a multi-stage network, each stage only focuses on a single task for reducing the learning complexity and promoting the convergence speed of the model. Furthermore, the input of the latter stage is pre-enhanced by the previous stage, which is beneficial to clarify the learning objectives of each stage. As shown in Fig. 1(b), in stage 1, we use the magnitude of the observed signal as the input and the target magnitude as the training target. This stage aims to suppress the unnatural noise components and interfering speech coarsely. After generating the enhanced magnitude from stage 1, we couple it together with the noisy phase and transform them into the real and imaginary spectrum as the input of stage 2. We also adopt the observed noisy complex spectrum as the input of stage 2 to further remove the remaining noise and interfering speech as well as fix the phase information of the target speech. Residual connection is applied between the input and output of stage 2 for avoiding vanishing gradients. In a nutshell, the process is as follows:

$$\mathbf{E} = \mathcal{F}_{spk}\left(\text{fbank}(|\mathbf{A}|); \mathbf{\Phi}_{spk}\right) \tag{1}$$

$$\left|\mathbf{S}^1\right| = \mathcal{F}_1\left(|\mathbf{Y}^p|, \mathbf{E}; \mathbf{\Phi}_1\right) \tag{2}$$

$$\left(\mathbf{S}_r^2, \mathbf{S}_i^2\right) = \left(\mathbf{S}_r^1, \mathbf{S}_i^1\right) + \mathcal{F}_2\left(\mathbf{S}_r^1, \mathbf{S}_i^1, \mathbf{Y}_r^p, \mathbf{Y}_i^p, \mathbf{E}; \mathbf{\Phi}_2\right) \tag{3}$$

Here $\mathbf{A}$ is the spectrum of time-domain enrollment signal and $\mathbf{E}$ is the extracted speaker embedding. Subscript r and i denote the real and imaginary parts of the complex spectrum, respectively. $\mathbf{Y}^p$ denotes the compressed observed spectrum and p means the compress factor. $\mathbf{S}^1$ and $\mathbf{S}^2$ denote the enhanced complex spectrum of stage 1 and 2, respectively. It should be noted that $\mathbf{S}_r^1 = \Re\left(\left|\mathbf{S}^1\right| e^{j\theta_Y}\right)$ and $\mathbf{S}_i^1 = \Im\left(\left|\mathbf{S}^1\right| e^{j\theta_Y}\right)$. The mapping functions of stage 1, stage 2, and speaker encoder are defined as $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_{spk}$ with the parameter sets being $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$, $\mathbf{\Phi}_{spk}$, respectively.

Both stage 1 and stage 2 take the similar network topology as [11, 18], which includes the gated convolutional encoder, decoder, and stacked temporal convolution modules (dubbed TCMs) [18]. Especially, Stage 1 network only has one decoder to estimate magnitude while stage 2 network has two decoders to estimate real and imaginary parts, respectively. About the fusion of speaker embedding into the enhancement network, we only concatenate the hidden feature and speaker embedding along the channel axis at the first TCM in every TCM groups in Fig. 1(a) according to [5]. It can combine speaker information with hidden feature efficiently and progressively.

2.2.1. *Loss Function*

As for the objective function of the network, we first apply scale-invariant signal-to-noise ratio (SI-SNR) [19] loss, which is a time-domain loss function:

$$\begin{cases} \mathbf{s}_{\text{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}} \\ \mathcal{L}_{\text{si-snr}} = 10 \log_{10} \frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \end{cases} \tag{4}$$

where $\hat{\mathbf{s}} \in \mathbb{R}^{1\times T}$ and $\mathrm{s} \in \mathbb{R}^{1\times T}$ refer to the estimated and original clean sources, respectively, and $\|\mathbf{s}\|^2 = \langle \mathbf{s}, \mathbf{s} \rangle$ denotes the signal power.

Speech over-suppression is a general problem in neural denoiser [20]. To solve the speech over-suppression issue, here we use the asymmetric loss [9] in the magnitude part of the loss function, with the purpose to penalize the T-F bins where the target speaker's voice is casually removed.

$$\begin{aligned} h(x) &= \begin{cases} 0, & \text{if } x \le 0, \\ x, & \text{if } x > 0, \end{cases} \\ \mathcal{L}_{\text{asym}} &= \frac{1}{T}\sum_t^T \sum_f^F \left| h\left(|S(t,f)|^p - |\hat{S}(t,f)|^p\right)\right|^2 \end{aligned} \tag{5}$$

The following strategy is applied to train the two-stage network. First, we only train stage 1 network with the hybrid loss:

$$\mathcal{L}_{\text{mag}} = \frac{1}{T}\sum_t^T \sum_f^F \left| |S(t,f)|^p - |\hat{S}(t,f)|^p \right|^2 \tag{6}$$

$$\mathcal{L}_1 = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{si-snr}} + \mathcal{L}_{\text{asym}} \tag{7}$$

Afterwards, the pre-trained model of stage 1 is loaded and the parameters are frozen to optimize stage 2 network by:

$$\mathcal{L}_{\text{RI}} = \frac{1}{T}\sum_t^T \sum_f^F \left| |S(t,f)|^p e^{j\theta_{S(t,f)}} - |\hat{S}(t,f)|^p e^{j\theta_{\hat{S}(t,f)}} \right|^2 \tag{8}$$

$$\mathcal{L}_2 = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{si-snr}} + \mathcal{L}_{\text{asym}} \tag{9}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ denote the loss function of stage 1 and stage 2, p is a spectral compression factor which is set to 0.5, and the operator θ calculates the argument of a complex number, which is used in [21].

3. EXPERIMENT SETUP

3.1. Datasets

The DNS challenge has released 48kHz training and development sets to participants. The training set consists of a wide range of approximately 750 hours of clean speech and 181 hours of noise clips. Specifically, the clean speech dataset includes six languages, namely English, French, German, Italian, Russian, and Spanish. The English part consists of reading speech and singing voice while the rest of the languages only have reading speech. The speech data comes from 3,230 speakers in total. Noise data mainly comes from Audioset [22], Freesound, and DEMAND [23].

As introduced in Section 2, our proposed TEA-PSE system includes two modules: speaker encoder and speech enhancement. The training data of these two modules are different. For the speaker encoder, we employ the development set of VoxCeleb2 [24] with over one million utterances, coming from 5,994 different speakers. For training the speech enhancement network, we use 675 hours of clean speech data, together with 144 hours of noise data, both extracted from the DNS dataset. As for the development set, 75 hours of clean speech data is selected from the DNS dataset, while 20 hours of noise data is selected from the DNS dataset. We also generate 100,000 single-channel room impulse responses (RIRs) based on the image method. The RT60 of RIRs ranges from 0.1s to 0.6s. The room size ranges from $3 \times 3 \times 3\ \mathrm{m}^3$ to $8 \times 8 \times 4\ \mathrm{m}^3$. The distance between the sound source and microphone ranges from 0.3m to 6.0m. The training and development sets contain 85,000 and 10,000 RIRs, respectively. It should be mentioned that there is no overlap between the training and development sets.

The test data of the proposed TEA-PSE system can be divided into three parts. The *simulation set* aims to measure the performance of the model on out-of-set speakers. We use the KING-ASR-215 dataset as the source speech, 17 hours of data from the DNS noise set as source noise, together with 5,000 RIRs, to generate 2,000 noisy-clean pairs as the simulation set, containing 200 speakers. For each noisy-clean pair, a random interfering speaker is added with SIR range of [-5, 20]dB and a random noise is added with SNR range of [-5, 20]dB. The second part is the *official development set* provided by the challenge organizer. It has 1443 clips which were all collected using desktop and laptop computers. The third part is the *official blind test set* provided by the challenge organizer which consists of 859 clips that were collected using desktop/laptop computers and mobile devices.

3.2. Training setup for the speaker encoder network

During the training of the ECAPA-TDNN speaker encoder network, we adopt the following data augmentation strategies:

- Waveform dropout: some random chunks of original waveform were replaced with zero.
- SpecAug: time and frequency masking were applied to the input spectrum [25].
- Additive noise: three types of noises from MUSAN [26] were added to the original speech of VoxCeleb 2.
- Reverberation: reverberation was convoluted with the original speech in VoxCeleb 2 and the RIRs are from [27].
- Speed perturbation: speed perturbation (0.9,1.1) without change of pitch was performed.
- Additive noise & reverberation: additive noise and reverberation were added at the same time.

A group of 80-dim FBank features with 25ms window size and 10ms window shift is extracted as the input. During the training procedure, the batch size is set to 256. The learning rate of model training varies between 1e-8 and 1e-3 using the triangular2 policy [28]. The optimizer during training is Adam. The hyper-parameter scale and margin of AAM-softmax are set to 30 and 0.3, respectively. To prevent over-fitting, we apply a weight decay of 2e-4 on all weights in the model.

3.3. Training setup for the speech enhancement network

Our training data are generated on-the-fly with 48kHz sampling rate and segmented into 4s chunks in each batch, with SNR range of [-5, 20]dB and SIR range of [-5, 20]dB. Specifically, we adopt the following strategies for data augmentation:

- Reverberation: to simulate far-field scenarios, 50% of the clean speech data is randomly selected to convolve with RIRs.
- Down-sampling & up-sampling: we select 20% training data to randomly down-sample to [12, 44]kHz and then up-sample to 48kHz.
- Different interference scenario: during adding the interfering speaker, 20% of the training data only contains one interfering speaker, 30% of the training data contains one interfering speaker and one type of noise, 30% of the training data only contains one type of noise and the rest 20% of the training data contains two types of noises.

We use 20ms frame length and 10ms frame shift, together with Hanning window, for front-end processing. The number of the STFT points is 1024, leading to 513-dim spectral features. The model is optimized by Adam. The two stages are trained independently. Stage 1 model is trained firstly, and then it is frozen and considered as the pre-trained model for the second stage model. The initial learning rate is 0.001, which will be halved if the validation loss of two consecutive epochs no longer decreases. The batch size is set to 20. Similar to [11], for both stage networks, the kernel size and stride of the convolution layers in encoders and decoders are (3, 2) and (2, 1) along with frequency and time axis, respectively. The number of channel remains 80 for all convolution layers. There are 6 encoders and decoders in each network. In the TCMs, the compressed feature size is set to 80 after 1x1-Conv. The total number of the TCMs is 4 with kernel size of 3 and dilation rate of $\{1,2,5,9\}$, respectively. Recently, the power compressed spectrum has been proved effective in the speech enhancement task [12, 14]. Hence we conducted the compression to the spectral magnitude before feeding into the network. Specifically, the compression parameter β is set to 0.5.

4. EXPERIMENT RESULTS AND ANALYSIS

4.1. Performance comparison on the simulation set

We conducted ablation experiments to prove the effectiveness of each proposed module, including a) single-stage, b) single-stage with PC, c) single-stage with PC and Asym, d) dual-stage, e) dual-stage with PC, and f) dual-stage with PC and Asym. We also compared the proposed model with Voicefilter [3], pDCCRN [9] and GateCRN [29]. We implemented and trained those models with the same data described above. Voicefilter and GateCRN are in the real domain, which aims to model magnitude only, while pDCCRN, the personalized version of DCCRN [30], works in the complex domain, modeling magnitude and phase simultaneously. All real domain models are optimized with $\mathcal{L}_1$ and all complex domain models are optimized with $\mathcal{L}_2$. For the development set and blind test set, the result of the DNS baseline is provided by the organizer.

According to the results in Table 1, the following conclusions can be drawn.

Table 1. Performance comparison on the simulation set. "PESQ-WB" represents PESQ for 16kHz audio files. "Noisy" represents the noisy speech without speech enhancement. "Single-stage" represents the single first-stage model. "Dual-stage" represents the dual-stage networks. "PC" represents power compression. "Asym" represents asymmetric loss.

#	System	PESQ-WB	STOI	ESTOI	SI-SNR
1	Noisy	1.395	69.90	55.36	1.559
2	Voicefilter	1.767	77.23	63.02	6.497
3	pDCCRN	1.690	74.45	60.24	5.629
4	GateCRN	1.870	78.32	64.49	7.070
5	Single-stage	1.910	79.33	65.85	7.672
6	Single-stage+PC	1.980	79.80	66.43	7.862
7	Single-stage+PC+Asym	1.925	79.44	65.94	7.536
8	Dual-stage	2.058	80.56	68.07	8.428
9	Dual-stage+PC	2.068	80.70	68.44	8.573
10	Dual-stage+PC+Asym	**2.139**	**81.39**	**69.33**	**8.587**

Table 2. DNSMOS P.835 performance comparison on the official development set. DNS challenge used DNSMOS P.835 to measure speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL).

#	System	SIG	BAK	OVRL
1	Noisy	3.89	2.99	3.13
2	Voicefilter	3.85	3.62	3.28
3	pDCCRN	3.84	3.87	3.35
4	GateCRN	3.92	3.73	3.38
5	DNS Baseline	3.54	**4.02**	3.22
6	Single-stage	3.92	3.85	3.44
7	Single-stage+PC	3.95	3.88	3.45
8	Single-stage+PC+Asym	**3.99**	3.84	3.48
9	Dual-stage	3.86	3.99	3.40
10	Dual-stage+PC	3.91	4.00	3.42
11	Dual-stage+PC+Asym	3.97	4.00	**3.49**

1. **Dual-stage** Firstly, our dual-stage network with PC and Asym surpassed other comparison methods in all metrics. The dual-stage network with PC and Asym outperforms Voicefilter by 0.372, 4.16%, 6.31%, and 2.09 in terms of PESQ-WB, STOI, ESTOI, and SI-SNR on average. Secondly, to the first stage, performance is improved when the second stage is introduced. Some unpleasant residual noise components and unwanted interfering speaker voices can be further suppressed. Moreover, stage 2 network can also repair target speech's phase information while 0.214, 1.95%, 3.39%, 1.051 gains are achieved in terms of PESQ-WB, STOI, ESTOI, and SI-SNR on average.
2. **Power compression** When power compression is adopted, the performance can be slightly improved. For example, in stage 1, when the range of the spectrum is compressed, it gets 0.07, 0.47%, 0.58%, 0.19 improvements for PESQ-WB, STOI, ESTOI, and SI-SNR. This is because power compression can decrease the dynamic range of the spectrum, which improves the significance of low-energy regions with more informative speech components.
3. **Asymmetric loss** For the single-stage model, the asymmetric loss will bring about declines of 0.055, 0.36%, 0.49%, and 0.326 about PESQ-WB, STOI, ESTOI, and SI-SNR, respectively. However, for the dual-stage model, if the asymmetric loss is used, 0.071, 0.69%, 0.89%, and 0.014 of improvements are obtained for PESQ-WB, STOI, ESTOI, and SI-SNR, respectively. The possible reason is that asymmetric loss is proposed to keep more speech and may cause slight degradation of noise suppression ability. So the performance of the single-stage model by asymmetric loss slightly decreases. However, speech distortion after dual-stage processing is more serious than that after single-stage and the asymmetric loss can help the dual-stage model to improve speech quality, which can be confirmed from System 6-11 in Table 2.

Table 3. MOS with ITU-T P.835 framework and WAcc results on the official blind test set.

#	System	SIG	BAK	OVRL	WAcc
1	Noisy	**4.25**	2.14	2.56	**0.72**
2	DNS Baseline	3.64	4.24	3.40	0.64
3	TEA-PSE	4.19	**4.55**	**3.97**	0.69

4.2. Performance comparison on official dev. & blind test sets

Table 2 presents the subjective results of DNSMOS P.835 [31] on the official development set. It can be observed that the proposed dual-stage network with PC and Asym surpasses other models on OVRL. When stage 2 network is applied, SIG will decrease because more distortion on target speech is introduced and BAK will increase since more noise and interfering speech are removed. Furthermore, if we do not use the asymmetric loss to penalize over-suppression, OVRL of the dual-stage model is slightly lower than the single-stage model.

Above all, Table 3 compares the MOS with ITU-T P.835 framework [32] and WAcc results on the official blind test set. The performance of the proposed TEA-PSE system (Dual-stage+PC+Asym) is obviously better than that of the DNS baseline system, which is consistent with the conclusion in Table 2. Besides, compared with noisy speech, the WAcc of the submission system is slightly decreased. It is reasonable since the model introduces slight distortion to the extracted speech, which may results in a mismatch with the training data of the ASR engine.

4.3. Algorithmic delay and inference time

In the submission system, the window size T is 20ms, and stride time T_s is 10ms. The algorithmic delay T_d is equal to $T + T_s = 30ms$, which meets the latency requirement of the challenge. No future information is utilized in the submission system. In addition, the number of trainable parameters of our two-stage framework is 7.81 million. The average processing time per frame of the submission system is 9.66ms on an Intel(R) Xeon(R) CPU E5-2678 v3 clocked at 2.4G Hz.

5. CONCLUSIONS

This paper introduces our submission to the ICASSP 2022 DNS Challenge Personalized SE track. Our system combines an ECAPA-TDNN speaker encoder network and a two-stage speech enhancement network to perform personalized speech enhancement. We specifically investigate the effectiveness of the dual-stage network, comparing it with the single-stage network, and explore the effect of power compression and asymmetric loss on the system. The proposed system achieved 1st rank in the track.

6. REFERENCES

[1] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, et al., "ICASSP 2022 deep noise suppression challenge," .

[2] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[3] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Interspeech*, 2019, pp. 2728–2732.

[4] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[5] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Interspeech*, 2020, pp. 1406–1410.

[6] C. Deng, S. Ma, Y. Sha, Y. Zhang, H. Zhang, H. Song, and Wang F, "Robust Speaker Extraction Network Based on Iterative Refined Adaptation," in *Interspeech*, 2021, pp. 3530–3534.

[7] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and Accurate Time-Domain Speaker Extraction Network," in *Interspeech*, 2020, pp. 1421–1425.

[8] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," in *Interspeech*, 2020, pp. 2677–2681.

[9] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," *arXiv preprint arXiv:2110.09625*, 2021.

[10] R. Giri, S. Venkataramani, J. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement," in *Interspeech*, 2021, pp. 1124–1128.

[11] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *ICASSP*. IEEE, 2021, pp. 6628–6632.

[12] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *Interspeech*, 2021, pp. 2801–2805.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *ICASSP*, 2021, pp. 6109–6113.

[14] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech*, 2020, pp. 3830–3834.

[16] S. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[18] S. Bai, J Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[19] Y. Luo and N. Mesgarani, "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.

[20] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Interspeech*, 2021, pp. 2816–2820.

[21] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human Listening and Live Captioning: Multi-Task Training for Speech Enhancement," in *Interspeech*, 2021, pp. 2686–2690.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Meetings on Acoustics ICA*. Acoustical Society of America, 2013, vol. 19, p. 035081.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[25] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019, pp. 2613–2617.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[27] E. AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[28] L. N Smith, "Cyclical learning rates for training neural networks," in *IEEE Winter conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.

[29] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[30] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Interspeech*, 2020, pp. 2472–2476.

[31] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv preprint arXiv:2110.01763*, 2021.

[32] B. Naderi and R. Cutler, "Subjective evaluation of noise suppression algorithms in crowdsourcing," *arXiv preprint arXiv:2010.13200*, 2020.