ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448315

TARGET SPEECH EXTRACTION WITH PRE-TRAINED SELF-SUPERVISED LEARNING MODELS

Junyi Peng[1]*, Marc Delcroix*[2]*, Tsubasa Ochiai*[2]*, Oldřich Plchot*[1]*, Shoko Araki*[2]*, Jan Černocký*[1]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]NTT Corporation, Japan

ABSTRACT

Pre-trained self-supervised learning (SSL) models have achieved remarkable success in various speech tasks. However, their potential in target speech extraction (TSE) has not been fully exploited. TSE aims to extract the speech of a target speaker in a mixture guided by enrollment utterances. We exploit pre-trained SSL models for two purposes within a TSE framework, i.e., to process the input mixture and to derive speaker embeddings from the enrollment. In this paper, we focus on how to effectively use SSL models for TSE. We first introduce a novel TSE downstream task following the SUPERB principles. This simple experiment shows the potential of SSL models for TSE, but extraction performance remains far behind the state-of-the-art. We then extend a powerful TSE architecture by incorporating two SSL-based modules: an Adaptive Input Enhancer (AIE) and a speaker encoder. Specifically, the proposed AIE utilizes intermediate representations from the CNN encoder by adjusting the time resolution of CNN encoder and transformer blocks through progressive upsampling, capturing both fine-grained and hierarchical features. Our method outperforms current TSE systems achieving a SI-SDR improvement of 14.0 dB on LibriMix. Moreover, we can further improve performance by 0.7 dB by fine-tuning the whole model including the SSL model parameters.

***Index Terms*—** Target speech extraction, pre-trained models, self-supervised learning, feature aggregation

1. INTRODUCTION

Over the past several years, transformer models trained with self-supervised learning (SSL) [1, 2] have shown great success in various speech tasks, such as automatic speech recognition (ASR) [3], speaker verification (SV) [4, 5], and speech enhancement (SE) [6, 7]. This effectiveness is attributed to the models' ability to learn overcomplete and general-purpose features when pre-trained on large-scale datasets, thereby ensuring robust performance and generalization, even under data-limited conditions [8].

Despite their strong performance in several downstream tasks, there are only a few studies investigating the use of SSL representations for target speech extraction (TSE) [9], a task aiming at estimating the speech of a target speaker from a multi-talker mixture[10].

TSE approaches usually use an extractor module consisting of a neural network, which inputs a speech mixture and estimates the target speech by exploiting a speaker embedding derived from an enrollment of the target speaker to identify him/her in the mixture. Consequently, the TSE approaches are related to separation and SV, making it an interesting use case for pre-trained SSL models. In this paper, we explore using pre-trained SSL models for TSE.

Leveraging the principles of Speech processing Universal PERformance Benchmark (SUPERB) [11, 12], we propose an SSL-based TSE system with a straightforward downstream model, such as a stack of bidirectional long short-term memory (BLSTM) for the extractor. The SSL model plays a dual role: first, extracting features from the input mixture and, second, obtaining speaker embeddings from the enrollment utterance. We show experimentally that such a system can be used to design a TSE system. However, as for other SE tasks, the performance is far behind the state-of-the-art.

One potential reason for this performance gap could be the large strides used in SSL models, which typically operate with a stride of about 20ms, yielding only 50 frames per second. This is in stark contrast to widely-used SE models, such TasNet [13], which utilizes smaller strides of 1.25ms (i.e. 800 frames per second), thus have better temporal resolution that might be crucial for optimal performance in SE tasks. Moreover, it is worth noting that SSL models like WavLM [1, 6] typically consist of two main components: a CNN Encoder and a series of Transformer blocks. Recent studies [7, 14, 15] have indicated that lower layers of the SSL models, especially the outputs of the CNN encoder that serve as the input to the Transformer blocks, are more relevant for SE tasks. Despite this, most SE models [16] typically focus on the representations obtained from the Transformer layers, neglecting the outputs of the intermediate CNN Encoder layers, thus failing to take advantage of the hierarchical representations acquired by models pre-trained on large-scale datasets.

To tackle the aforementioned challenges, this paper presents a systematic approach that leverages multi-scale representations from SSL models for TSE tasks. We construct two modules based on a frozen pre-trained Transformer model named the *speaker encoder* (SpkEnc) and the *Adaptive Input Enhancer* (AIE). The SpkEnc module computes target speaker embeddings from the enrollment by following prior works on using SSL models for SV [17]. The AIE module is designed to extract features from the mixture. In particular, it adjusts the time resolutions across intermediate layers of the CNN Encoders and the transformer of the SSL model, thereby allowing the exploitation of multi-scale feature representations. Finally, the entire SSL model is jointly fine-tuned with the TSE system to enhance performance further. Overall, our contributions are as follows:

- We propose a new TSE downstream task, developed in line with SUPERB principles, which goes beyond single-objective downstream tasks by emphasizing the multi-faceted capabilities of pre-trained models, such as SE feature extraction and speaker encoding.
- We introduce two sub-modules, i.e., SpkEnc and AIE, designed to function as plug-in units, enabling flexible integra-

The work was partly supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports through the e-INFRA CZ (ID:90254).

tion of the pre-trained SSL model into a powerful TSE architecture.

- We conduct a comprehensive study on the Libri2mix dataset and demonstrate that exploiting pre-trained SSL representation can boost the performance of a powerful TSE system, outperforming prior systems such as SpEx+ [18] and TD-SpeakerBeam [19].

2. PRIOR WORKS

Pre-trained SSL models have been used for SE tasks such as denoising and speech separation. In [14], SSL model representations are used to estimate the time-frequency mask for the Short-Time Fourier Transform (STFT) of the input signal, refined by BLSTM layers. SSL-based approach demonstrated superior performance compared to FBANK-based methods on LibriMix [20] and Voicebank-DEMAND [21]. To further improve SE performance, various strategies have been explored, such as the fusion of SSL and STFT features [7], and the introduction of a regression-based training objective [22]. However, unlike our proposal, they do not exploit the features from the CNN layers, although they are probably the most relevant for SE tasks.

The only prior work using SSL for TSE is [9], where a pre-trained SSL model is only employed to derive the speaker embeddings from the enrollment, resulting in a limited improvement (0.3 dB) compared to using FBANK features.

SSL models have also been used for target speaker-ASR, which focuses on transcribing a specific speaker from a segmented utterance containing multi-talker speech using enrollment speech for that speaker. In [16], a speaker embedding is prepended to the input features of the SSL model's Transformer blocks. The entire model is then fine-tuned using a CTC loss. This approach differs from TSE as it outputs a character sequence instead of the target speech signal.

3. CONVENTIONAL NEURAL TSE

Let us first describe the overall architecture of a typical TSE system [10]. A neural TSE system consists of four main blocks: the encoder, decoder, SpkEnc, and extractor. The encoder transforms the input speech mixture $\mathbf{y}$ into higher-dimensional features $\mathbf{Z}_y$ as $\mathbf{Z}_y = \text{Encoder}(\mathbf{y})$, which could be either spectral features obtained via STFT or learned features derived from 1D convolutional layers operating on the raw waveform. SpkEnc is responsible for extracting speaker embedding $\mathbf{e}$ that captures the voice characteristics of the target speaker, usually derived from an enrollment speech $\mathbf{c}$, as $\mathbf{e} = \text{SpkEnc}(\mathbf{c})$. The extractor estimates the target speech from the mixture in the feature domain $\mathbf{Z}_y$, where $\mathbf{Z}_s = \text{Extractor}(\mathbf{Z}_y, \mathbf{e})$, given the target speaker embeddings $\mathbf{e}$. Finally, the decoder transforms the processed feature $\mathbf{Z}_s$ into the estimated target speech $\hat{\mathbf{x}}$, where $\hat{\mathbf{x}} = \text{Decoder}(\mathbf{Z}_s)$.

4. EXPLOITING PRE-TRAINED SSL MODELS FOR TSE

Many SSL models including WavLM [15], Hubert [1], and wav2vec2.0 [2] consist of CNN and Transformer blocks producing the intermediate outputs denoted $\mathbf{H}_j^{\text{cnn}}$ and $\mathbf{H}_i^{\text{trf}}$, with $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, N\}$ indicating the index of CNN and Transformer blocks, respectively, where J and N denote the total number of CNN and Transformer blocks in the models. We can exploit the feature representation obtained by these models as input features for the extractor and for the SpkEnc.

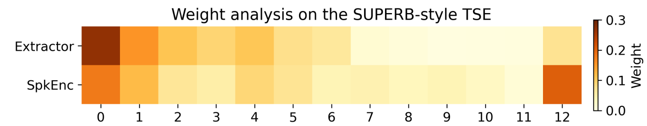

Fig. 1. Layer-wise weights of speaker encoder (SpkEnc) and extractor, using the BLSTM-based TSE downstream model, and WavLM Base Plus pretrained SSL model. Note that 0-th Transformer layer denotes the output of the CNN encoder, which is also the input of the 1st Transformer layer.

First, we propose a simple TSE downstream model following the style of the SUPERB evaluation to carry out preliminary experiments. We then discuss how to exploit SSL features within a more powerful TSE architecture.

4.1. SUPERB-style downstream TSE model

In line with SUPERB's methodology, we construct two lightweight modules that utilize two different set of weights to perform a weighted sum of SSL features as $\sum w_i^\nu \mathbf{H}_i^{\text{trf}}$, where $\sum w_i^\nu = 1$, where w_i^ν are the weights for the extractor or SpkEnc (i.e., $\nu = \{\text{SpkEnc}, \text{extractor}\}$) for layer i of the Transformer. The SpkEnc computes the target speaker embedding by averaging these weighted features over frames followed by a linear layer. The extractor comprises a three-layer BLSTM model that predicts a time-frequency mask. It accepts the SSL features as input. The processing is conditioned on the target speaker by multiplying the speaker embedding with the output of the first BLSTM [23]. We then multiply the STFT coefficients of the speech mixture with the estimated mask and apply inverse STFT to produce the waveform of the extracted speech. During the training, the SSL model is kept frozen while the sub-modules, including the weights w_i^ν are learnable. We provide experiments with such a model in Section 5.2.

Despite its simplicity, it succeeds in performing TSE, but the performance remains far behind current TSE systems trained from scratch (e.g., TD-SpeakerBeam [19]). Figure 1 plots the weights w_i^ν obtained with such a system. This reveals that lower layers of the Transformer blocks, particularly layer 0, corresponding to the output of the CNN encoder, are more critical for the extractor.

4.2. TD-SpeakerBeam extension with pre-trained SSL models

We then explore effectively exploiting SSL representation within a powerful TSE system, such as TD-SpeakerBeam [19]. TD-SpeakerBeam uses 1-D convolutions for the encoder and decoder and temporal convolutional network (TCN) blocks for the extractor and SpkEnc. Note that the time resolution of TD-SpeakerBeam is typically much higher than most layers of an SSL model.

Building upon the insights from the SSL weight visualization of the SUPERB-style TSE model in Fig. 1, we investigate the use of the CNN encoder module positioned before the Transformer blocks of the SSL model. We introduce an AIE module that integrates the output of the different layers of the SSL model. In particular, AIE performs progressive upsampling to adjust the time resolutions of the different layers of the CNN encoder and the Transformer, capturing multi-scale and fine-grained information from the SSL module. The output of the AIE module, $\mathbf{h}$, has the same time resolution as the output features of the encoder, $\mathbf{Z}_y$. Consequently, these two streams can be concatenated along the feature dimension and fed into the Extractor model as $\mathbf{Z}_s = \text{Extractor}(\text{concat}(\mathbf{Z}_y, \mathbf{h}))$, as shown in Fig. 2-(a)

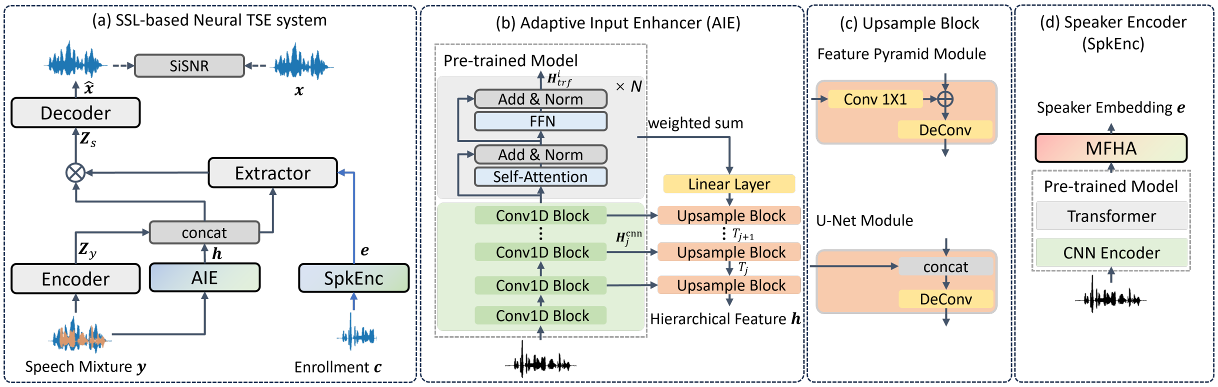

Fig. 2. (a) Overview of the proposed SSL model-based TSE system, and the details of (b) AIE module, (c) upsample blocks, and (d) SSL-based SpkEnc.

4.2.1. *Adaptive input enhancer*

The proposed AIE leverages a series of one-dimensional deconvolutional blocks formulated as $\mathbf{T}_j = \text{Upsample}(\mathbf{T}_{j+1}, \mathbf{H}_j^{\text{cnn}})$, where $\mathbf{T}_j$ is the output of j-th upsample block and has the same time resolution as $\mathbf{H}_j^{\text{cnn}}$, as shown in Fig. 2-(b). The output of AIE is the hierarchical feature obtained as $\mathbf{h} = \mathbf{T}_2$, since the second CNN layer has the same time resolution as $\mathbf{Z}_y$ for the typical setting of TD-SpeakerBeam. Note that when considering features from both the Transformer and the CNN, the topmost layer of the AIE is initialized as $\mathbf{T}_{top} = \text{Linear}(\sum_{i=1}^{N} w_i^{\text{AIE}} \mathbf{H}_i^{\text{trf}})$, where w_i^{AIE} are learnable scalar weights summing to one.

We experiment with two variants to implement the upsampling operation as shown in Fig. 2-(c).

Feature Pyramid Module: The upsampling operation is borrowed from the FPN architecture [24]. It is implemented as follows: $\mathbf{T}_j = \text{DeConv}(\text{Conv}(\mathbf{H}_j^{\text{cnn}}) + \mathbf{T}_{j+1})$, where $\text{DeConv}(\cdot)$ is a deconvolution operation that perform upsampling and $\text{Conv}(\cdot)$ is a convolution operation used to transform the output of the j-th CNN block.

Unet-Style Module: This upsampling alternative follows the U-Net architecture [25] and is implemented as: $\mathbf{T}_j = \text{DeConv}(\text{Concat}(\mathbf{H}_j^{\text{cnn}}, \mathbf{T}_{j+1}))$, where the concatenation is performed along the channel dimension.

The implementation of the upsampling slightly differs between these two configurations, but both approaches perform top-down processing to extract multi-scale representations from the CNN Encoder layers. We expect that such a hierarchical upsampling process can capture rich information from the pre-trained SSL model, yielding hierarchical features, $\mathbf{h}$, that can complement the features obtained from the TSE encoder, $\mathbf{Z}_y$.

4.2.2. *Speaker encoder based on pre-trained model*

Compared to SUPERB architecture, which simply adopts the average operation, we employ an advanced attentive pooling, named multi-head factorized attentive pooling (MHFA) [17], to enhance the quality of learned speaker representation, as shown in Fig. 2-(d). This SSL-MHFA employs two sets of normalized layer-wise weights to generate attention maps and compressed features, which are expected to encode speaker-discriminative information and phonetic information respectively. Then, the speaker embedding is formed by aggregating over frames and projecting the vector to a lower-dimensional space using a linear layer. In this way, each attention head is expected to aggregates information from a specific set of phonetic units, which leads to a robust speaker embedding.

5. EXPERIMENTS

5.1. Experiment Setup

Data-sets: We conduct experiments using the Libri2Mix dataset, consisting of simulated mixtures of two speakers[20]. Following the data preparation in TD-SpeakerBeam[1] with 16kHz sampling rate, dataset is partitioned into three subsets: train-100, dev, and test. We choose the model with the best SI-SNRi performance on the dev.

Implementation details: For *SUPERB-style setup*, we use a BLSTM-based TSE. The window size and the number of FFT points are set to 1024, the dimension of BLSTM is 512, and so is the speaker embedding. As a baseline, we use the same architecture with STFT or FBANK features.

For the *TD-SpeakerBeam setup*, we use an SSL-based speaker encoder using MHFA with a total of 8 heads as in [17]. The speaker embedding is of dimension 256. For TD-SpeakerBeam baseline, the configuration follows the specifications in [19]. We use Adam as the optimizer with an initial learning rate set to 10^{-3}. When the pre-trained model is unfrozen for joint optimization with the TSE system, the learning rate is set to 2×10^{-5}. When using TD-SpeakerBeam as downstream model, we use the WavLM Base Plus[2].

Performance Metrics: We measure performance in terms of source-to-distortion Ratio (SDR), scale-invariant SDR improvement (SI-SDRi), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and Failure rate (FR) [26]. FR measures the proportion of test samples with an SI-SDRi below 1 dB. Failures typically occur when the TSE system extracts the incorrect speaker or outputs the mixture.

5.2. Evaluation results following SUPERB's setup

Table 1 shows the performance of different SSL models using the SUPERB style downstream model describes in Section 4.1. We observe that SSL models significantly outperform the acoustic feature (STFT and FBANK) models across all three metrics: SI-SDRi, STOI, and PESQ in Table 1. This suggests the potential of utilizing SSL models for TSE tasks. WavLM models outperform Hubert and wav2vec versions, probably because the training style, including noise and interference speakers, provides more robust speech representations. Although WavLM Large achieves the best performance, we chose the more compact WavLM Base Plus in the remaining experiments. Note that the best model achieves an SI-SDRi of 10.3 dB, which is significantly lower than TSE models trained from scratch, such as SpecEx+ [18] or TD-SpeakerBeam [19], which attain an SI-SDRi of more than 13 dB.

[1] https://github.com/BUTSpeechFIT/speakerbeam
[2] https://huggingface.co/microsoft/wavlm-base-plus

Table 1. Performance comparison of various SSL models for target speech extraction tasks, following the SUPERB Challenge settings.

Upstream	Target Speech Extraction		
	SI-SDRi↑	STOI↑	PESQ↑
STFT	5.96	0.79	1.48
FBANK	5.18	0.78	1.41
HuBERT Base	9.18	0.86	1.77
wav2vec 2.0 Base	9.15	0.86	1.76
WavLM Base	9.69	0.87	1.95
WavLM Base Plus	9.96	0.88	1.97
WavLM Large	10.30	0.88	2.01

Table 2. Performance comparison for different approaches to exploit SSL models for TD-SpeakerBeam setup.

SpkEnc	AIE		SDR↑	SI-SDRi↑
	SSL Feature	Fusion Method		
TCN	-	-	13.69	13.03
SSL-MHFA	-	-	12.91	12.13
	Transformer	Weighted Sum	13.84	13.18
	Single- CNN	-	13.12	12.40
	Multi- CNN	Unet	14.03	13.67
		FPM	14.14	13.49
	Multi- CNN + Transformer	Unet	14.38	13.80
		FPM	14.65	14.01

5.3. Evaluation Results on TD-SpeakerBeam setup

Table 2 shows the extraction performance of different variants of TD-SpeakerBeam using or not using SSL features. We compare different versions of the SpkEnc, SSL features (from Transformer, CNN layers, or both), and fusion methods of the AIE module. The first row corresponds to the baseline TD-SpeakerBeam, which uses a TCN as SpkEnc. The second row replaces the SpkEnc with the SSL-MHFA introduced in Section 4.2.2. We observe that using SSL for SpkEnc degrades performance. One possible reason could be the difference in model architectures. The lightweight attentive pooling might be insufficient to effectively deal with the complex feature distributions captured by SSL-based models when paired with a TCN-based extractor that is randomly initialized.

Next, we investigate augmenting the input of the extractor with different SSL features. Using a weighted sum of the transformer layers (followed by a deconvolution layer for upsampling) or the single CNN layer that matches the time resolution of the extractor fails to improve the results significantly. To address this, we introduce the AIE module. Notably, adopting upsampling modules to obtain hierarchical features, including U-Net and FPM, outperforms the baseline. Finally, merging the feature representations from both the Transformer and CNN Encoder layers offers additional performance improvements. These results demonstrate that we can improve extraction performance with a pre-trained SSL model if we use a strategic integration method like our proposed AIE.

We also investigate the contributions of various components in our proposed architecture. Employing a UNet as an AIE with a traditional TCN-based SpkEnc yields an SI-SDRi of 12.63 dB, which is slightly lower than the baseline. This result demonstrates that while direct component replacement may not yield immediate benefits, strategic integration like AIE and SSL-MHFA provides significant performance gains. We also confirmed that a model with the same architecture as the AIE but randomly initialized performs worse than when using the pretrained SSL model (13.26 dB v.s. 13.67 dB). This highlights the importance of the SSL pre-training in capturing robust features beneficial for TSE tasks.

Table 3. SI-SDRi performance under different learning conditions. FR denotes Failure rate.

Fine-tuning	SDR↑	SI-SDRi↑	STOI↑	PESQ↑	FR(%)↓
✗	14.65	14.01	0.91	2.38	3.9
✓	**15.26**	**14.65**	**0.93**	2.45	3.0
TD-SpeakerBeam [19]	13.69	13.03	0.90	2.12	4.8
SpEx+ [18]	-	13.41	-	**2.93**	-
sDPCCN [27]	-	11.61	-	-	-

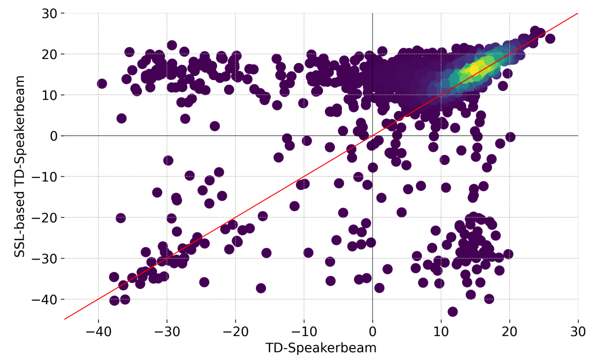

Fig. 3. Comparison of SI-SDRi scores of test set samples using TD-SpeakerBeam (X-axis) against the best SSL-based model (Y-axis).

We further investigate the performance of our proposed model after fine-tuning all parameters, including the SSL model. The results are summarized in Table 3. The best performance is achieved when both AIE and SpkEnc are fine-tuned, reaching an SI-SDRi of 14.65. In side experiments, we confirmed that constraining the same SSL model for both AIE and SpkEnc results in about 0.1 dB SI-SDRi degradation but significantly reduces the number of model parameters. The proposed system significantly outperforms the baseline methods, including TD-SpeakerBeam, SpEx+, and sDPCCN, underscoring the advantages of incorporating SSL models into the TSE framework.

Finally, we analyze the performance improvement achieved by incorporating SSL into the TSE system. Figure 3 plots the SI-SDRi of TD-SpeakerBeam versus that of the proposed extension of TD-SpeakerBeam with a fine-tuned SSL model, where each dot represents the performance of one test sample. We observe that it not only improves the quality of the already well-extracted samples but also improves performance when TD-SpeakerBeam performs relatively poorly (SI-SDRi between -10 and 10dB). This translates into having a significantly lower FR value (3.0% v.s. 4.8%) as shown in Table 3, which indicates a more accurate identification of the target speaker.

6. CONCLUSIONS

In this work, we proposed using pre-trained SSL models for TSE. We introduced a new downstream task, following SUPERB, as a benchmark for evaluating the performance of TSE models. Besides, we explored using SSL models with more powerful TSE systems. Our extensive experiments on Libri2mix demonstrate the importance of exploiting both CNN and Transformer layers of the SSL model and properly upsampling the representation. After fine-tuning SSL-based components, we improved significantly over existing systems trained from scratch. Future work will include exploring similar AIE modules for other SE tasks and reducing the number of parameters of the models.

7. REFERENCES

[1] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[3] A Arunkumar and Srinivasan Umesh, "Joint Encoder-Decoder Self-Supervised Pre-training for ASR," in *Proc. Interspeech 2022*, 2022, pp. 3418–3422.

[4] Nik Vaessen and David A Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[5] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[6] Hyungchan Song, Sanyuan Chen, Zhuo Chen, Yu Wu, Takuya Yoshioka, Min Tang, Jong Won Shin, and Shujie Liu, "Exploring WavLM on Speech Enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 451–457.

[7] Kuo-Hsuan Hung, Szu wei Fu, Huan-Hsin Tseng, Hsin-Tien Chiang, Yu Tsao, and Chii-Wann Lin, "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 186–190.

[8] Junyi Peng, Oldřich Plchot, Themos Stafylakis, Ladislav Mosner, Lukáš Burget, and Jan "Honza" Černocký, "Improving Speaker Verification with Self-Pretrained Transformer Models," in *Proc. INTERSPEECH 2023*, 2023, pp. 5361–5365.

[9] Xiaoyu Liu, Xu Li, and Joan Serrà, "Quantitative evidence on overlooked aspects of enrollment speaker embeddings for target speaker separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[11] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[12] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, et al., "SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8479–8492.

[13] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[14] Zili Huang, Shinji Watanabe, Shu-wen Yang, Paola García, and Sanjeev Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6837–6841.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Zili Huang, Desh Raj, Paola García, and Sanjeev Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] Junyi Peng, Oldřich Plchot, Themos Stafylakis, Ladislav Mošner, Lukáš Burget, and Jan Černockỳ, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 555–562.

[18] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410.

[19] Marc Delcroix, Tsubasa Ochiai, Katerina Zmolikova, Keisuke Kinoshita, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki, "Improving Speaker Discrimination of Target Speech Extraction With Time-Domain Speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 691–695.

[20] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[21] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[22] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech.," in *SSW*, 2016, pp. 146–152.

[23] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan 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.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[26] Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Katerina Zmolikova, Hiroshi Sato, and Tomohiro Nakatani, "Listen only to me! How well can target speech extraction handle false alarms?," in *Proc. Interspeech 2022*, 2022, pp. 216–220.

[27] Jiangyu Han, Yanhua Long, Lukáš Burget, and Jan Černockỳ, "DPCCN: Densely-connected pyramid complex convolutional network for robust speech separation and extraction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7292–7296.