2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-7451-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSPW62465.2024.10627502

PROBING SELF-SUPERVISED LEARNING MODELS WITH TARGET SPEECH EXTRACTION

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

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

ABSTRACT

Large-scale pre-trained self-supervised learning (SSL) models have shown remarkable advancements in speech-related tasks. However, the utilization of these models in complex multi-talker scenarios, such as extracting a target speaker in a mixture, is yet to be fully evaluated. In this paper, we introduce target speech extraction (TSE) as a novel downstream task to evaluate the feature extraction capabilities of pre-trained SSL models. TSE uniquely requires both speaker identification and speech separation, distinguishing it from other tasks in the Speech processing Universal PERformance Benchmark (SUPERB) evaluation. Specifically, we propose a TSE downstream model composed of two lightweight task-oriented modules based on the same frozen SSL model. One module functions as a speaker encoder to obtain target speaker information from an enrollment speech, while the other estimates the target speaker's mask to extract its speech from the mixture. Experimental results on the Libri2mix datasets reveal the relevance of the TSE downstream task to probe SSL models, as its performance cannot be simply deduced from other related tasks such as speaker verification and separation.

Index Terms— Target speech extraction, self-supervised learning, SUPERB

1. INTRODUCTION

Transformer models, empowered by self-supervised learning (SSL) [1, 2, 3, 4], have recently marked significant achievements in the field of speech processing, including automatic speech recognition (ASR) [5], speaker verification (SV) [6, 7, 8], and speech enhancement (SE) [9, 10]. The robustness and generalization abilities of these models are attributed to their capacity to extract general-purpose features through the SSL paradigm on large-scale datasets [11].

To quantitatively evaluate the SSL models for various speech tasks, benchmarks such as the Speech processing Universal PERformance Benchmark (SUPERB) and its multilingual variant have been proposed [12, 13, 14]. In these benchmarks, SSL models are evaluated on several downstream tasks using lightweight task-specific models that rely on input features derived from the layer-wise outputs of the frozen pre-trained SSL models. SUPERB covers diverse downstream tasks, including ASR, SV, SE, speech separation, etc.

Recently, the problem of Target Speech Extraction (TSE), defined as the process of isolating the speech signal of a target speaker from a multi-talker mixture using auxiliary cues [15], has attracted significant interest [16, 17]. This task not only requires speech separation but also the precise identification of the target speaker. Such a dual requirement makes TSE a valuable candidate for evaluating the capabilities of SSL models in extracting fine-grained (acoustic) features and understanding speaker-specific context. However, SUPERB does not include a TSE downstream task.

In this paper, we introduce a novel TSE downstream task, following SUPERB principles. Specifically, we build an SSL-based *extractor* model that processes a speech mixture to estimate the target speech. This process is conditioned on a target speaker embedding obtained by a *speaker encoder* using the enrollment speech of the target speaker. Both the extractor and speaker encoder are derived from the same pre-trained SSL model. With this new downstream task, we aim to evaluate pre-trained SSL models from a new perspective and answer the following research question: *Is TSE performance governed by the performance of SSL models on the related SV and separation tasks?*

There are a few works that use pre-trained SSL models for TSE [18, 19]. In [18], a pre-trained SSL model was explored to extract target speaker embeddings from enrollment speech for TSE, but not for the extraction module. It yielded a marginal improvement over FBANK features. In [19], an SSL model was employed for encoding both mixtures and speaker enrollment. While it briefly introduced a SUPERB-style downstream model, it predominantly focused on the integration of SSL representations into existing TSE systems (i.e., TD-SpeakerBeam [20]). Compared to that work, this paper makes the following key contributions:

- **SUPERB-TSE System:** We introduce a novel TSE task for evaluating pre-trained SSL models following principles from SUPERB. With this new task, we investigate various implementation choices for the downstream model, highlighting SSL's potential in extracting the target speaker's speech from mixtures.
- **Comparative Analysis:** We benchmark nine well-known large-scale SSL speech models and three Whisper models [21] with our proposed TSE downstream task.
- **Performance Correlation:** Our observations reveal that the performance of TSE tasks cannot be simply inferred from the performance on the isolated SV and Separation downstream tasks, suggesting a more intricate relationship between these tasks.
- **Comparison with TD-SpeakerBeam:** We compare the performance of the SUPERB-TSE model with a strong TSE system (TD-SpeakerBeam), in terms of training time and performance. It shows that while the SSL-based system enables fast training, there is significant room for further improvement.

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).

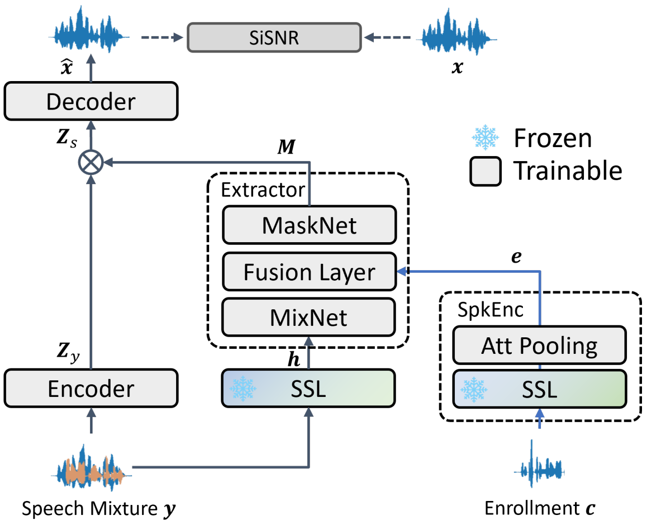


Fig. 1: Architecture of proposed SSL-based TSE system.

2. SUPERB-STYLE TARGET SPEECH EXTRACTION

In this section, we introduce the downstream TSE model used to probe SSL models. Figure 1 shows the architecture of the TSE model, emphasizing the pre-trained SSL models. The TSE problem consists of extracting the speech of a target speaker, $\mathbf{x}$, in a mixture, $\mathbf{y} = \mathbf{x} + \mathbf{i}$, where $\mathbf{i}$ consists of interference speakers and noise. We rely on an enrollment utterance, $\mathbf{c}$, to identify the target speaker.

The proposed SSL-based TSE system consists of four main blocks: the encoder, decoder, speaker encoder (SpkEnc), and extractor, as a typical TSE system [15]. The encoder transforms the input speech mixture $\mathbf{y}$ into a sequence of features $\mathbf{Z}_y$ by $\mathbf{Z}_y = \text{Encoder}(\mathbf{y})$. SpkEnc is responsible for computing a speaker embedding vector, $\mathbf{e}$, from the enrollment speech, $\mathbf{c}$, as $\mathbf{e} = \text{SpkEnc}(\mathbf{c})$. Subsequently, the extractor computes the target speech mask $\mathbf{M}$ within the feature domain $\mathbf{Z}_y$ from SSL features $\mathbf{h}$ and the target speaker embedding $\mathbf{e}$, as $\mathbf{M} = \text{Extractor}(\mathbf{h}, \mathbf{e})$. Here, following the standard SUPERB approach, the SSL features consist of a weighted sum of the outputs of the Transformer blocks. We then obtain a feature representation of the target speech, $\mathbf{Z}_s$, by applying the mask on the features of the mixture, as $\mathbf{Z}_s = \mathbf{M} \odot \mathbf{Z}_y$, where $\odot$ denotes Hadamard multiplication operation. Finally, the decoder converts $\mathbf{Z}_s$ back to the time domain, to obtain the target speech signal as, $\hat{\mathbf{x}} = \text{Decoder}(\mathbf{Z}_s)$.

2.1. Encoder/Decoder

The Encoder module operates on the raw waveform by utilizing a set of time-domain finite impulse response filters. We explore two options for the encoder and decoder.

First, we can use Short Time Fourier Transform (STFT) and inverse STFT (iSTFT) for the encoder and decoder, respectively. Alternatively, the filter bank can also be randomly initialized and then jointly optimized with the entire TSE system, allowing the filters to focus on task-related frequency bands [22]. Accordingly, the Decoder is implemented by using a deconvolution layer that up-samples the target speaker features $\mathbf{Z}_s$ back into the time-domain waveform.

Table 1: Different fusion methods in TSE. FiLM: Feature-wise linear modulation, which uses two vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ obtained by projecting the embedding vector, $\mathbf{e}$, with two learnable linear layers.

Fusion Method	Implementation	SI-SDRi↑
Addition	$\mathbf{Z}_f = \mathbf{Z}_{mix} + \mathbf{e}$	9.13
Multiplication	$\mathbf{Z}_f = \mathbf{Z}_{mix} \odot \mathbf{e}$	**9.96**
Concatenation	$\mathbf{Z}_f = concat(\mathbf{Z}_{mix}, \mathbf{e})$	8.88
FiLM	$\mathbf{Z}_f = \mathbf{Z}_{mix} \odot \mathbf{e}_1 + \mathbf{e}_2$	8.16

2.2. SSL-based Speaker Encoder

In SUPERB-style SV downstream tasks, speaker representation is obtained using a time-delay neural network (TDNN)-based speaker extractor (i.e. X-vector [23]) from the weighted sum of layer-wise SSL features. To construct a more lightweight module, we proposed an attentive pooling called multi-head factorized attentive pooling (MHFA) to extract speaker information [8], as shown in Figure 1. The SSL-MHFA utilizes two sets of normalized layer-wise weights to generate attention maps and compressed features, respectively. These are designed to encode speaker-discriminative information and phonetic information, respectively. We obtain the speaker embedding vector by aggregating the compressed features over frames and then project the resulting vector to a lower-dimensional space using a linear layer. This method allows each attention head to focus on specific phonetic units, resulting in a speaker embedding that is robust to phonetic variability.

In the proposed downstream TSE model, we use the SSL-MHFA module to compute the target speaker embedding vector, $\mathbf{e}$. Note that this module is trained jointly with the other components of the TSE model without any speaker identification loss.

2.3. SSL-based Extractor

The extractor is composed of three sub-modules: a mixture encoder (MixNet), a fusion layer, and a target mask generator (MaskNet), as illustrated in Figure 1. It accepts the SSL features, $\mathbf{h}$, as input. The processing is conditioned on the target speaker through a fusion layer that combines the speaker embedding $\mathbf{e}$ and the output of the MixNet, $\mathbf{Z}_{mix}$, as $\mathbf{Z}_f = \text{Fusion}(\mathbf{Z}_{mix}, \mathbf{e})$. Finally, MaskNet computes the mask, $\mathbf{M}$ from $\mathbf{Z}_f$.

In this study, MixNet is implemented by a single BLSTM layer, while Masknet employs two BLSTM layers. We explore various options for the fusion layer [24], summarized in Table 1.

3. EXPERIMENTS

We perform three sets of experiments. First, we investigate the configuration for the TSE downstream model. We then probe various pre-trained SSL models on the proposed TSE downstream task. Finally, we compare the performance with a powerful TSE system.

3.1. Experiment Setup

Datasets: In this work, we conduct comparative experiments across three downstream tasks: TSE, SV, and Separation. For TSE, we use Libri2Mix [25], consisting of simulated mixtures of two speakers. Following the enrollment speech preparation in TD-SpeakerBeam[1] with 16kHz sampling rate, the dataset is partitioned into three subsets: train-100, valid, and test. Our proposed TSE downstream task relies on Libri2mix with train-100 for faster experimental turnover.

[1] https://github.com/BUTSpeechFIT/speakerbeam

Table 2: Evaluating different TSE model configurations in Libri2mix (16kHz-min). For a fair comparison, we use the layer-wise outputs of the pre-trained WavLM Base Plus model as SSL features.

System	Encoder/Decoder	Extractor	SpkEnc	Mask Type	Objective	SI-SDRi↑	STOI(%)↑	PESQ↑
1	STFT/iSTFT	STFT	STFT	Magnitude Mask	MSE	5.96	79.55	1.42
2	STFT/iSTFT	STFT	SSL	Magnitude Mask	MSE	7.42	81.75	1.51
3	STFT/iSTFT	SSL	STFT	Magnitude Mask	MSE	8.70	85.03	1.83
4	STFT/iSTFT	SSL	SSL	Magnitude Mask	MSE	9.96	87.79	**1.97**
5	STFT/iSTFT	SSL	SSL	Magnitude Mask	SI-SDR	10.66	88.74	1.91
6	STFT/iSTFT	SSL	SSL	Complex Mask	SI-SDR	10.61	88.84	1.91
7	Conv1D/DeConv1D	SSL	SSL	Encoder-domain Mask	SI-SDR	**11.04**	**89.47**	1.93

Table 3: Comparison of different general-purpose speech models for TSE, SV, and separation downstream tasks. For Whisper models, we only use the audio encoder. In the fine-tuning case, initializing from a converged model with a frozen SSL (e.g. WavLM Base Plus), we unfreeze the SSL and further train the entire system for 20 epochs.

Upstream	#Params	TSE				SV (MHFA)	SV (Xvector)	Separation
		SI-SDRi↑	STOI (%)↑	PESQ↑	FR(%)↓	EER(%) ↓	EER(%) ↓	SI-SDRi↑
Whisper-Base	20.59M	9.25	86.13	1.71	6.11	3.39	9.55	9.76
Whisper-Small	88.15M	10.28	88.79	1.82	4.45	2.55	9.19	11.06
Whisper-Medium	307.22M	9.94	87.26	1.85	6.50	4.22	8.66	10.94
Data2vec Base	93.84M	9.43	86.21	1.72	5.45	3.51	6.79	9.95
Data2vec Large	314.30M	9.55	86.11	1.77	7.46	2.59	7.61	10.81
wav2vec 2.0 Base	95.00M	9.52	86.34	1.72	5.23	3.53	6.10	10.01
wav2vec 2.0 Large	317.38M	8.40	84.24	1.74	9.31	3.04	6.38	10.31
Hubert Base	94.68M	9.62	86.69	1.74	4.56	3.06	5.30	10.01
Hubert Large	316.61M	9.03	85.41	1.88	8.73	2.94	5.82	10.95
WavLM Base	94.70M	10.03	87.99	1.84	3.71	2.71	5.36	10.80
WavLM Base Plus	94.70M	**11.04**	**89.47**	1.93	**3.45**	**2.03**	**4.39**	11.41
WavLM Large	316.62M	9.73	86.53	**2.04**	7.96	2.30	4.87	**11.87**
WavLM Base Plus [Fine-tuning]		11.51	90.08	2.01	3.23	-	-	-
TD-SpeakerBeam		13.03	90.63	2.21	4.85	-	-	-

Consequently, all experiments in the paper rely on this configuration, except when specifically specified in the exploration experiments in Section 3.4. Regarding SV, all models are trained using the VoxCeleb1 dataset and evaluated on VoxCeleb1-O [26]. The speakers in the training and evaluation sets are different. For Separation, we evaluate the performance of SSL models on Libri2mix.

Implementation details: In the system using STFT/iSTFT for Encoder/Decoder, the window size and the number of FFT points are set to 1024 with a stride of 320. This stride aligns with the downsampling rate in the SSL model. Additionally, the dimension of BLSTM is 512. It is noted that this configuration is coherent with the SUPERB's Separation downstream task hyper-parameters [13]. For the systems with learnable kernels, using Conv1D and DeConv1D, the kernel size is set to 1024 with a stride of 320, and the number of filters is 512.

For the speaker encoder, when the input consists of STFT, we process the magnitude of the STFT coefficients with a three-layer BLSTM followed by average pooling to derive the speaker embedding vectors. Conversely, for SSL features, we employ MHFA for speaker embedding extraction. The MHFA is configured with four heads and a compression layer with a dimension of 128.

We trained all modules of the TSE model jointly using the Mean Squared Error (MSE) or scale-invariant signal-to-distortion ratio (SI-SDR) loss between the reference and estimated target speech [27]. Note that the MSE is computed in the spectral domain, while SI-SDR is computed in the time domain. We used the Adam optimizer and trained the model for 200 epochs.

For the SV task, we use AM-softmax loss with a scale of 30 and a margin of 0.4 as the objective function. We set the number of heads to 32 for MHFA, resulting in a model size of 2.23M parameters compared to the 5.71M parameters of Xvector.

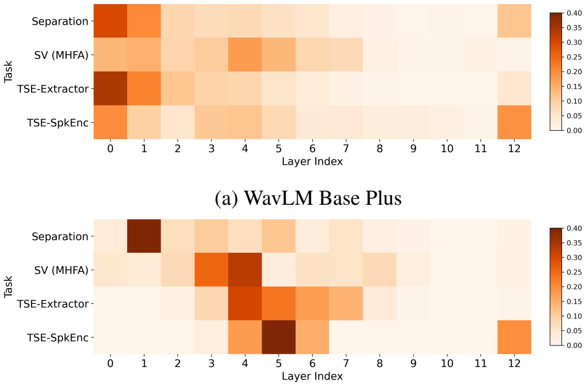
For the separation experiments, we simply employ the default configuration in SUPERB, which consists of a three-layer BLSTM.

Performance Metrics: For TSE, we measure performance in terms of scale-invariant signal-to-distortion ratio (SI-SDR) improvement (SI-SDRi), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and Failure rate (FR) [28]. 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. For SV, we calculate the Equal Error Rate (EER).

3.2. Analysis of downstream model configuration

We first perform several experiments to find an effective configuration for the TSE downstream model. Here, we use WavLM Base Plus as the upstream model and STFT/iSTFT as the encoder/decoder unless specified.

In Table 1, we evaluate the effectiveness of various fusion strategies employed in TSE. We observe that the multiplication strategy



(a) WavLM Base Plus

(b) Whisper Small

Fig. 2: The weight distribution of Transformer layers. Note that 0-th Transformer layer denotes the output of the CNN encoder, which is also the input of the 1-st Transformer layer.

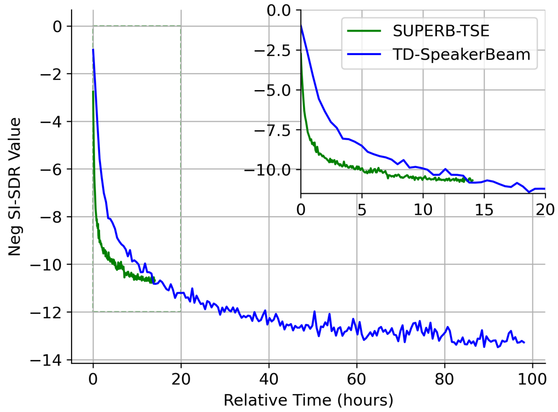


Fig. 3: Comparison of SUPERB-TSE and TD-SpeakerBeam training curves, i.e., the validation loss (negative SI-SDR) as a function of the training time (time × # GPUs).

outperforms other approaches. Consequently, we use it in all subsequent experiments.

In Table 2, we present an extensive evaluation of different TSE model configurations on the Libri2mix dataset. First, we compare using SSL features with STFT coefficients for the extractor and SpkEnc (systems 1 to 4). We observe that using SSL features for both the extractor and SpkEnc (system 4) results in significantly higher extraction performance. Next, we explore the use of the time domain SI-SDR loss and using magnitude or complex masks (systems 5 and 6). We found that using SI-SDR loss significantly improves SI-SDRi and STOI scores. However, using complex masks (system 6) had little impact on the results. Finally, we replaced the STFT/iSTFT encoder/decoder modules with learnable ones (system 7). This again improved SI-SDR and STOI scores. We use that configuration in subsequent experiments.

3.3. Performance Comparison of various SSL models

In Table 3, we evaluate the performance of various SSL models (and Whisper encoder) across three tasks: TSE, SV, and speech separation. The models are trained with LibriSpeech data and LibriLight [29]. Note that the difference between WavLM Base and Base Plus is the amount of training data, i.e., WavLM Base is trained only with LibriSpeech 960 hours of training data, while WavLM Base Plus is trained using the same data as WavLM Large, i.e., 94k hours.

First, let us look at the SV and separation performance. For SV, we observe that MHFA outperforms Xvector with fewer parameters (Params: 2.31M v.s. 5.71M), suggesting its effectiveness. For SV with MHFA, Large SSL models tend to perform best, except for WavLM where WavLM Base Plus achieves significantly better performance. For separation, WavLM Large outperforms other models, and large SSL models constantly perform better than base ones.

Intuitively, we would expect that SSL models achieving high scores in terms of separation and extraction should be good for TSE. However, we observe that there is a more intricate relation, e.g., Base models usually perform better than Large models for TSE although they perform worse in terms of SV and separation. Compared to other SSL models, WavLM Base Plus achieves the best performance for SV and TSE tasks. This suggests the importance of data augmentation in the pre-training stage to capture robust speech and speaker representations.

These experiments demonstrate that the performance of TSE models is not directly correlated with their performance in Separation and SV tasks. Moreover, as illustrated in Figure 2, the distributions for Separation and Extractor, as well as SV and SpkEnc, are distinctly different. However, the two TSE sub-modules show a similar pattern, which might be a result of their joint optimization. This observation further emphasizes the uniqueness of TSE and the necessity for task-specific evaluation of SSL models rather than assuming a universally effective model across various speech tasks.

3.4. Comparison with powerful TSE system

Finally, we compare the WavLM-based TSE system (hereafter called SUPERB-TSE), with a strong TSE model, TD-SpeakerBeam [20]. First, Figure 3 compares the training speed of these two models. The SUPERB-TSE system demonstrates faster convergence, achieving a 10 dB SI-SDR within 2-3 hours of training time, in contrast to TD-SpeakerBeam, which requires over 10 hours to achieve similar performance. Moreover, TD-SpeakerBeam reaches full convergence in 100 hours, whereas SUPERB-TSE achieves this in only 14 hours. Moreover, as seen in Table 3, the failure rate is lower with the WavLM-based TSE system (FR: 3.45 vs 4.85 %)), which may indicate better speaker identification capabilities. However, despite these benefits, there remains a significant performance gap between SUPERB-TSE and TD-SpeakerBeam systems (SI-SDRi: 11.04 dB v.s. 13.03 dB as shown in Table 3). We attribute part of this gap to the very simple model architecture of the downstream model, as well as to the low time resolution of the SSL model, which may not be optimal for speech enhancement [19].

4. CONCLUSIONS

In this work, we introduce a novel SSL-based TSE aligned with SUPERB principles. Our comprehensive experiments on Libri2Mix datasets demonstrate that TSE performance cannot be directly inferred from SV and Separation tasks, justifying the importance of including the TSE downstream task when probing SSL models. We showed that with careful implementation choices, we can build a relatively strong TSE downstream model, which achieves fast convergence. However, such a simple downstream model still lags behind more powerful TSE systems trained from scratch, such as TD-SpeakerBeam. However, fine-tuning the SSL model for TSE as well as enhancing the temporal resolution of SSL models, may constitute promising research directions to boost TSE performance [19].

5. 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] 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.

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[5] Zhengyang Li, Thomas Graave, Jing Liu, Timo Lohrenz, Siegfried Kunzmann, and Tim Fingscheidt, "Parameter-efficient cross-language transfer learning for a language-modular audiovisual speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, et al., "Ml-superb: Multilingual speech universal performance benchmark," *arXiv preprint arXiv:2305.10615*, 2023.

[15] 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.

[16] 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.

[17] Naoyuki Kamo, Marc Delcroix, and Tomohiro Nakatani, "Target Speech Extraction with Conditional Diffusion Model," in *Proc. INTERSPEECH 2023*, 2023, pp. 176–180.

[18] 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.

[19] Junyi Peng, Marc Delcroix, Tsubasa Ochiai, Oldřich Plchot, Shoko Araki, and Jan Černocký, "Target speech extraction with pre-trained self-supervised learning models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[20] 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.

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[22] 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.

[23] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP 2018*. IEEE, 2018, pp. 5329–5333.

[24] 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.

[25] 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.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[27] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[28] 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.

[29] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.