

TS-SUPERB: A Target Speech Processing Benchmark for Speech Self-Supervised Learning Models

Junyi Peng^{*,1}, Takanori Ashihara^{*,2}, Marc Delcroix², Tsubasa Ochiai², Oldrich Plchot¹, Shoko Araki², Jan Černocký¹

¹Brno University of Technology, Czechia, ²NTT Corporation, Japan

Abstract—Self-supervised learning (SSL) models have significantly advanced speech processing tasks, and several benchmarks have been proposed to validate their effectiveness. However, previous benchmarks have primarily focused on single-speaker scenarios, with less exploration of target-speaker tasks in noisy, multi-talker conditions—a more challenging yet practical case. In this paper, we introduce the Target-Speaker Speech Processing Universal Performance Benchmark (TS-SUPERB), which includes four widely recognized target-speaker processing tasks that require identifying the target speaker and extracting information from the speech mixture. In our benchmark, the speaker embedding extracted from enrollment speech is used as a clue to condition downstream models. The benchmark result reveals the importance of evaluating SSL models in target speaker scenarios, demonstrating that performance cannot be easily inferred from related single-speaker tasks. Moreover, by using a unified SSL-based target speech encoder, consisting of a speaker encoder and an extractor module, we also investigate joint optimization across TS tasks to leverage mutual information and demonstrate its effectiveness.¹

Index Terms—Self-supervised learning, target-speaker speech processing, speech recognition, speech enhancement, voice activity detection

I. INTRODUCTION

Many speech processing tasks, including automatic speech recognition (ASR), speaker verification (SV), and speech enhancement (SE), have been significantly advanced by self-supervised learning (SSL) [1]–[4]. This paradigm enables downstream tasks to achieve remarkable performance by exploiting general-purpose features from large SSL models trained on large-scale unlabeled datasets, even with lightweight task-oriented decoders trained on limited amounts of labeled data [5]–[9].

To quantitatively evaluate the capabilities of SSL models across various speech tasks, benchmarks such as the Speech processing Universal PERformance Benchmark (SUPERB) and its variants have been proposed [10]–[15]. SUPERB offers a comprehensive comparison of SSL models across a wide range of downstream speech tasks, such as ASR and SV. Its downstream tasks consist mostly of single-speaker speech processing tasks.

However, in daily conversational situations, the speech of interest is often interfered with by other speakers (and background noise). To address the complex yet practical conditions, several target-speaker (TS) speech processing tasks, including target speech extraction (TSE), personalized SE (PSE), target-speaker ASR (TS-ASR), and personal voice activity detection (PVAD) [16]–[19], have been developed. For instance, TSE, which aims to extract the TS’s speech signal from a mixture, can be used to develop recording devices that focus on the desired speaker for, e.g., hearing aids/hearables or

teleconferencing systems applications [20]. TS-ASR, which aims to transcribe only the desired user’s speech, can be used to develop personalized ASR systems, potentially for use in smart speakers or smart watches [21].

Standard TS systems exploit TS’s clues, such as speaker embeddings derived from a pre-recorded enrollment utterance, to extract information about the speaker from a multi-talker mixture. Overall, TS tasks are more challenging than conventional speech processing tasks because they address two problems simultaneously: identifying the TS and extracting information about the speaker from the speech mixture. We, therefore, characterize the TS tasks, such as TS-ASR, as having a “dual objective” in contrast to conventional tasks, such as ASR, that address only one problem, i.e., have a “single objective,” because they deal with single-speaker and thus do not require identifying the TS.

Recently, several studies have explored combining TS systems with SSL models [22]–[25]. Despite the growing interest in TS tasks with SSL models, existing benchmarks have not fully assessed and compared their capabilities in addressing TS tasks. To advance research in these areas and promote the development of more versatile SSL models, the present study proposes Target Speaker SUPERB (TS-SUPERB). TS-SUPERB includes four widely accepted downstream tasks—TS-ASR, PSE, TSE, and PVAD—with a focus on extracting content (ASR) and fine or coarse acoustic characteristics (SE and VAD, respectively).

In our benchmark, the downstream models are kept simple, with a structure similar to the related tasks in SUPERB, and consist of two modules: an SSL-based *target speech encoder* and a *task-oriented decoder* as shown in Fig. 1. The target speech encoder includes an SSL-based extractor that obtains the TS features, informed by speaker representations derived from enrollment speech via an SSL-based speaker encoder. Since the downstream tasks in TS-SUPERB are related, we can use a similar configuration for the downstream model across all TS tasks, except for the prediction head (or decoder module). The decoder module is task-specific, aligning with the model architecture of related single-speaker tasks in SUPERB.

TS-SUPERB extends the evaluation framework beyond the single-objective SUPERB benchmark by emphasizing various aspects critical for overlapped multi-talker speech processing. The key contributions of the proposed TS-SUPERB are:

Comprehensive Evaluation: The proposed TS-SUPERB provides a standardized framework for evaluating SSL-based TS processing models across four tasks, highlighting SSL’s potential in extracting the TS information from mixtures. We will release our benchmark with an evaluation code to encourage further research in SSL-based TS speech processing.

Comparative Analysis: We benchmark seven leading speech SSL models with the proposed TS-SUPERB. Additionally, we analyze the relationship between tasks by computing the layer-wise weight distribution and further explore correlations between TS tasks and other related tasks (e.g., ASR, SV, and speech separation (Sep)).

The work was supported by European Defense Fund’s project No. 101168083 “ARCHER”, and Czech Ministry of Interior project No. VJ01010108 “ROZKAZ”, Ministry of Education, Youth and Sports of the Czech Republic (MoE) through the OP JAK project “Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications” (ID:CZ.02.01.01/00/23_020/0008518). Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports through the e-INFRA CZ (ID:90254).

¹The code is available at https://github.com/BUTSpeechFIT/TS_SUPERB
* Equal contributors.

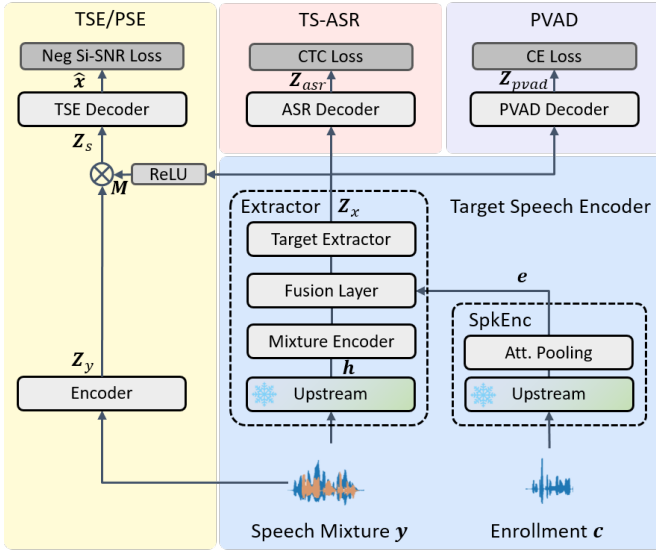


Fig. 1. Architecture of proposed TS-SUPERB system. All downstream models use the same architecture for the target speaker encoder, followed by the TSE, PSE, TS-ASR, or PVAD decoder, depending on the task.

Shared architecture: The target speech encoder, which processes both the input mixture and enrollment speech, shares the same architecture across tasks. By sharing the parameters of the target speech encoder, we can explore jointly training it using multi-task learning across different TS downstream tasks. We demonstrate experimentally that such joint-training can improve performance.

II. RELATED WORK

Benchmarking SSL models is essential to drive progress in the field. Current speech SSL benchmarks cover a wide variety of downstream tasks [10]–[15], [26]–[31], but none include TS speech processing tasks. It is thus difficult to compare the effectiveness of recently proposed SSL schemes for TS tasks. For example, there has been an increasing focus on using SSL models for TS speech processing [25], [32]. In [25], the enrollment speech is incorporated into the pre-training stage as an auxiliary input, achieving superior TS-ASR performance. In [32], the authors proposed appending speaker embeddings to the Transformer encoder’s input within an SSL model and then fine-tuning the entire model for TS-ASR. However, these studies focused on a single TS task. It is thus unclear if the learned representation by these models can also be effective for other TS tasks. In this paper, we propose a framework to tackle this issue. Our proposed TS-SUPERB assembled datasets to allow fair comparison between SSL models. We also provide an experimental framework, allowing us to easily benchmark the performance of SSL models for various downstream TS tasks. While the TS tasks are relatively complex compared to single-speaker tasks, our benchmark uses a simple architecture for the downstream model, consisting of a few bidirectional long short-term memory (BLSTM) layers, similar to those used in conventional speech SSL benchmarks, such as SUPERB [10] or SUPERB-SG [11].

III. TS-SUPERB

In this section, we provide details on the datasets for the multi-talker scenario and the downstream models/tasks within the TS-SUPERB framework, including TSE, PSE, PVAD, and TS-ASR.

TABLE I
STATISTICS OF SPEECH DATASETS USED FOR THE DIFFERENT
DOWNSTREAM TASKS IN TS-SUPERB.

TASK	Datasets	Train/Val/Test (hrs)
TSE	Libri2Mix-min clean [33]	43.2 / 4.5 / 4.1
PSE	Noisy SparseLibri2Mix [33]	70.1 / 6.1 / 5.8
PVAD	Noisy SparseLibri2Mix [33]	70.1 / 6.1 / 5.8
TS-ASR	Libri2Mix-max clean [33]	56.3 / 7.6 / 7.0
TSE+TS-ASR	Libri2Mix-max clean [33]	56.3 / 7.6 / 7.0
PSE+PVAD	Noisy SparseLibri2Mix [33]	70.1 / 6.1 / 5.8

A. Datasets

The training/testing datasets for each task, along with their statistical information, are listed in Table I. For the TSE task, we use the Libri2Mix-min clean dataset, which consists of two-speaker mixtures truncated to the duration of the shortest utterance to produce 100% overlap ratio. For PSE and PVAD, we use the Noisy SparseLibri2Mix dataset [33]. The dataset simulates overlapped speech in conversational audio signals, by mixing speech from LibriSpeech [34] and noise from the WHAM! dataset [35]. Similar to LibriCSS [36], the overlap ratio in the training set ranges from 0 to 40%. The training set consists of 24 000 mixtures, totaling 70 hours. The test dataset contains four overlap conditions (i.e. 0%, 20%, 40%, and 60%). For TS-ASR, we use the Libri2Mix-max clean dataset, which is similar to the Libri2Mix-min except that mixture corresponds to the length of the longest utterance.

B. Target Speech Encoder

Figure 1 presents the unified architecture, emphasizing the integration of pre-trained SSL models. This system aims to isolate the target-speaker information from a mixture y , modeled as $y = x + i$, where i represents the interference from other speakers and background noise and x denotes the clean speech. The process utilizes an enrollment utterance c , to identify the TS. The system is structured around two main components: a unified *target speech encoder* that we describe below and a *task-oriented decoder*, which we detail in the next subsections.

The target speech encoder comprises an extractor and a speaker encoder (SpkEnc), as shown in Fig. 1. SpkEnc employs an attentive pooling (i.e., multi-head factorized attention pooling (MHFA) [7]) to compute speaker vector e by processing SSL features based on enrollment speech, c . Subsequently, the extractor computes the target speech feature, Z_x , from the upstream model’s representations, h , and the TS embedding, e , as $Z_x = \text{Extractor}(h, e)$. Note that, as in SUPERB, h consists of the weighted sum of the output of the Transformer layers of the SSL model. Different from other SUPERB tasks, here SSL models are used twice in the downstream model to process the enrollment c and speech mixture y . Different from using a pre-trained speaker embedding extractor for the SpkEnc, our architectural choice for the SpkEnc allows for future comparisons with recent SSL schemes that handle mixture speech and enrollment speech/embeddings together [32].

For the mixture encoder within the target speech encoder, as shown in Fig. 1, a single BLSTM layer is used, while the target extractor utilizes two BLSTM layers. The dimension of the BLSTM layers is set to 512, except for the PVAD task, where it is 32. This adjustment is based on preliminary PVAD experiments, which showed no significant performance differences when using models with larger dimensions. The fusion layer uses broadcast multiplication to combine two features [16]. The MHFA is configured with four heads and a compression layer with a dimension of 128. To align

TABLE II

COMPARISON OF DIFFERENT SPEECH SELF-SUPERVISED MODELS FOR TS-SUPERB INCLUDING TSE, PSE, PVAD, AND TS-ASR DOWNSTREAM TASKS. ADDITIONALLY, THE PERFORMANCE ON THE CORRESPONDING SS, ASR AND SV TASKS IS INCLUDED FOR COMPREHENSIVE EVALUATION.

Upstream	TSE	PSE		PVAD		TS-ASR (WER)		Sep [37]	ASR [10]	SV [37]
	SI-SDRi↑	SI-SDRi↑	FR↓	mAP↑	m. tss↑	w/ LM↓	w/o LM↓	SI-SDRi↑	WER↓	EER↓
data2vec Base [2]	9.43	10.36	4.22	0.945	0.967	33.83	40.51	9.95	4.94	3.51
HuBERT Base [3]	9.62	10.36	4.98	0.928	0.953	34.57	41.75	10.01	6.42	3.06
WavLM Base [4]	10.03	11.01	3.58	0.951	0.971	23.72	29.13	10.80	6.21	2.71
WavLM Base+ [4]	11.04	10.96	3.58	0.942	0.961	23.45	29.09	11.41	5.59	2.03
data2vec Large [2]	9.55	10.37	4.80	0.954	0.970	28.93	35.84	10.81	3.36	2.59
HuBERT Large [3]	9.03	9.76	4.25	0.954	0.973	25.10	31.88	10.95	3.62	2.94
WavLM Large [4]	10.47	11.34	3.67	0.966	0.980	17.97	22.62	11.87	3.44	2.30

with SUPERB, we keep SSL models frozen by default throughout the training of all tasks (except when specifically mentioned).

C. Target Speech Extraction (TSE)

Task: TSE aims to estimate the speech of a TS from a multi-talker mixture [16]. It can be used to evaluate both the generative and speaker discrimination capabilities of SSL models.

Architecture: In addition to the aforementioned target speech encoder, TSE has a separate encoder that transfers the mixture \mathbf{y} into a sequence of features \mathbf{Z}_y . Subsequently, \mathbf{Z}_x are processed through a non-linear activation function, such as ReLU, to compute the target speech mask \mathbf{M} , in the feature domain of \mathbf{Z}_y . Finally, the decoder reconstructs the masked features $\mathbf{Z}_s = \mathbf{M} \odot \mathbf{Z}_y$ back into the time domain, resulting in the target speech signal, $\hat{\mathbf{x}} = \text{TSEdecoder}(\mathbf{Z}_s)$. The negative scale-invariant signal-to-noise ratio (SI-SNR) is used as the training objective. Regarding the implementation, we employ Conv1D and DeConv1D for the encoder and decoder, respectively, setting the kernel size to 1024, stride to 320, and the number of filters to 512.

Evaluation: We evaluate TSE performance in terms of scale-invariant signal-to-distortion ratio improvement (SI-SDRi) measured on the Libri2Mix-min, as detailed in Section III-A.

D. Personalized Speech Enhancement (PSE)

Task: Different from TSE, which focuses on processing highly overlapped speech, PSE is more concerned with extracting the TS's voice in conversational scenarios with sparsely overlapped ratios and background noise. The PSE task probes the generative and speaker discrimination capabilities of SSL models with noisy multi-talker inputs.

Architecture: The downstream model architecture is the same as for TSE.

Evaluation: We use the SparseLibri2Mix dataset, described in Section III-A, to evaluate the PSE task. The final performance is measured in terms of SI-SDRi, perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and Failure rate (FR) [38], where FR is defined as the percentage of test samples with SDRi below 1 dB.

E. Personalized VAD (PVAD)

Task: PVAD is designed to detect the voice activity of a TS in a mixture at frame level [19]. Unlike VAD, which simply detects the presence of any speech or not, PVAD aims to detect the TS's voice in a multi-talker recording.

Architecture: The PVAD decoder employs a classification layer (i.e., a linear layer with softmax activation function) to predict frame-level labels with cross-entropy loss between actual labels. Each utterance is annotated with frame-level labels covering three categories, i.e., target speaker speech (tss), non-target speaker speech (ntss), and none-speech (ns), such as noise and silence.

Evaluation: The dataset used for PVAD aligns with that for PSE, i.e., noisy SparseLibri2Mix. We measure performance using mean average precision (mAP) and mAP of target speaker speech (m. tss) [19].

F. Target-speaker ASR (TS-ASR)

Task: TS-ASR transcribes target-speaker speech into words given the multi-talker mixture. This task serves to evaluate both the linguistic and speaker discrimination capabilities of SSL models.

Architecture: The decoder within TS-ASR employs a single BLSTM layer with 512 dimensions. It estimates the probability of characters using the connectionist temporal classification (CTC) loss. The trained model is decoded with the official LibriSpeech 4-gram LM, making use of KenLM [39] and the Flashlight toolkit [40].

Evaluation: The performance of TS-ASR is evaluated on the Libri2Mix-max dataset. The evaluation metric is the word error rate (WER).

G. Multi-task Learning

All TS-SUPERB tasks share the same architecture for the target speech encoder. To leverage the potential mutual information between different tasks and further enhance the learned features, we explore a multi-task learning framework for joint optimization across tasks. Here we investigate multi-task learning between two tasks (TSE+TS-ASR and PSE+PVAD). The optimization objective can be formulated as $\mathcal{L} = \alpha\mathcal{L}_i + (1 - \alpha)\mathcal{L}_j$, where $0 \leq \alpha \leq 1$ denotes the task weight. \mathcal{L}_i and \mathcal{L}_j are the loss functions of the two tasks, i.e., $(i, j) \in \{(\text{TSE}, \text{TS-ASR}), (\text{PSE}, \text{PVAD})\}$.

IV. RESULTS AND ANALYSIS

A. Performance comparison of various SSL models

The results of evaluating upstream models on TS-SUPERB are presented in Table II. Additionally, to explore the correlations between tasks, we also include results from related ASR, SV, and Sep tasks following implementation in S3PRL, within the same table. Specifically, for PSE, we report the averaged SI-SDRi and FR metrics across four different overlap ratio conditions. Overall, no single SSL model consistently leads across all tasks. Meanwhile, the WavLM Base+ and Large models exhibit competitive performance in TS tasks, likely due to their use of a training data augmentation strategy, which leverages mixed utterances from multiple speakers, making them well-suited for the multi-speaker and noisy conditions of TS-SUPERB.

In detail, WavLM Base shows the best performance in the PVAD task, followed by data2vec Base and WavLM Base+ among all BASE models, though the lead is marginal. For TS-ASR, the Large model significantly outperforms the Base model, attributed to its stronger modeling of phoneme-related hidden units during pre-training. In denoising tasks (PSE and TSE), WavLM models obviously outperform

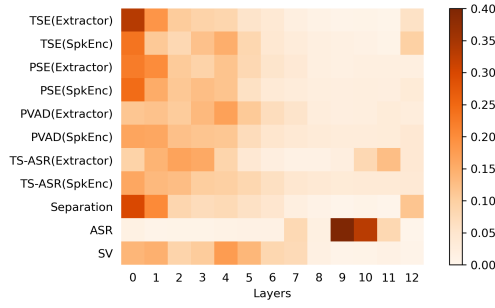


Fig. 2. The weight distribution of SSL layers. Note that the 0th layer denotes the input of the 1st Transformer encoder layer.



Fig. 3. The Spearman's rank correlation between tasks.

others, possibly because of denoising strategies incorporated during their pre-training stage as data augmentation.

Note that this paper aims to compare SSL models over various TS tasks and not to achieve top performance on these tasks. However, for reference, we compared the results of Table II with prior TSE and TS-ASR works. SSL-based models demonstrate comparable performance with state-of-the-art systems trained from scratch like TD-SpeakerBeam [41], which achieves SI-SDRi of 13.03 dB and 10.71 dB on the TSE and PSE tasks, respectively, suggesting the promising applications of SSL models in denoising. However, the performance of the TS-ASR systems is significantly behind that reported in [23], which employed a pre-trained speaker encoder, conditioned the pre-trained SSL model on the TS embeddings, and performed full fine-tuning of the SSL model, resulting in a WER of 12.32 % with a different configuration for enrollment speech. This indicates the potential for future improvement and investigation.

B. Layer-wise and task correlation analysis

The layer weights distributions of the WavLM Base+ model across various tasks are shown in Fig. 2. In TS tasks, the contributions of the bottom layers in both the extractor and the SpkEnc are more significant than for tasks such as SV, Sep, and ASR. SpkEnc's weight distribution resembles that in SV, indicating the bottom layer of the Base model contains rich speaker-related information. In addition, both the TS-ASR extractor and ASR show a peak at the 10th layer, suggesting that semantic information is encoded in the top layers. However, unlike ASR, TS-ASR also requires information from the lower layers to identify the TS in the mixture.

Furthermore, we investigate the correlations between tasks within TS-SUPERB and related tasks to verify if task relationships confirm existing hypotheses in speech processing. After converting all metrics to a higher-is-better format, we then calculate Spearman's correlation coefficients among all tasks [42]. As illustrated in Fig. 3, all TS-SUPERB tasks exhibit a strong correlation with Sep. One

TABLE III
INVESTIGATION OF JOINT-TRAINING OF TSE AND TS-ASR ON THE LIBRI2MIX-MAX CLEAN DATA.

Model	SI-SDRi \uparrow	STOI(%) \uparrow	PESQ \uparrow	WER \downarrow
TSE+TS-ASR	11.74	90.60	2.14	33.26
- Only TSE	11.16	89.40	1.97	-
- Only TS-ASR	-	-	-	29.09
TSE+TS-ASR [FT]	12.70	91.66	2.24	26.29

TABLE IV
PERFORMANCE ON NOISY SPARSELIBRI2MIX. ALL RESULTS ARE AVERAGES OF FOUR DIFFERENT OVERLAP RATIOS (I.E. 0%, 20%, 40%, 60%) TESTING DATA. #PARAMS DENOTES THE TRAINABLE PARAMETERS.

Model	#Params	SI-SDRi \uparrow	mAP \uparrow
PSE+PVAD	7.3 M	11.09	0.962
- Only PSE	7.3 M	10.96	-
- Only PVAD	0.7 M	-	0.942
PSE+PVAD [FT]	101M	11.49	0.961

possible reason is that these tasks share a similar goal: extracting specific speaker information from mixed speech signals. SV shows a high correlation with the TS tasks, except for PVAD, suggesting that PVAD does not require strict speaker representations. TSE, PSE, and TS-ASR show a high correlation, whereas PVAD has a weaker correlation with TSE and PSE but a stronger connection with TS-ASR and ASR. One possible explanation is that PVAD, TS-ASR, and ASR focus on frame-level feature processing, which does not require the fine-grained temporal resolution needed by TSE and PSE. It is observed that TS-ASR has a surprisingly weak correlation with ASR. This could be attributed to the additional requirement of TS-ASR to separate the mixture speech, which complicates the task using the simple ASR decoder module, making it challenging to simultaneously deal with both tasks effectively.

C. Analysis of multi-task learning and fine-tuning

To investigate how the target speech processing tasks complement each other, we evaluate the performance of joint-training TSE and TS-ASR in Table III and PSE with PVAD in Table IV, respectively. The first experiments are conducted using Libri2Mix-max, while the second utilizes noisy SparseLibri2Mix. The α is set to 0.5 for all experiments. WavLM Base+ was used for the upstream model.

When jointly training TSE and TS-ASR tasks (Table III), the system achieves improved TSE performance in terms of SI-SDRi, STOI, and PESQ, compared to TSE alone but worse WER than standalone TS-ASR. The fine-tuning of the whole system results in further improvement in all metrics.

In Table IV, the jointly optimized model improves in both PSE and PVAD performance. This indicates that speaker detection and denoising tasks can mutually benefit each other. Notably, training with an unfrozen SSL model leads to further improvements, especially for the TSE+TS-ASR task.

V. CONCLUSION

In this paper, we propose a new benchmark, named TS-SUPERB, for target-speaker speech processing, comprising four new tasks: TSE, PSE, TS-ASR, and PVAD. We further explore the multi-task learning within those tasks to utilize their mutual information. Through the comprehensive experiments with seven SSL models evaluated on Libri2Mix and noisy SparseLibri2Mix datasets, our results demonstrate the uniqueness of target speech processing tasks as their performances cannot be simply deduced from SV, ASR, and Sep tasks alone.

REFERENCES

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [2] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *ICML*, 2022, pp. 1298–1312.
- [3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.
- [4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. 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.
- [5] Z. Li, T. Graave, J. Liu, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, "Parameter-efficient cross-language transfer learning for a language-modular audiovisual speech recognition," in *ASRU*, 2023, pp. 1–8.
- [6] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP*, 2022, pp. 6147–6151.
- [7] J. Peng, O. Plhot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *SLT*, 2023, pp. 555–562.
- [8] H. Song, S. Chen, Z. Chen, Y. Wu, T. Yoshioka, M. Tang, J. W. Shin, and S. Liu, "Exploring WavLM on speech enhancement," in *SLT*, 2023, pp. 451–457.
- [9] K.-H. Hung, S. wei Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting self-supervised embeddings for speech enhancement," in *INTERSPEECH*, 2022, pp. 186–190.
- [10] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERFORMANCE Benchmark," in *INTERSPEECH*, 2021, pp. 1194–1198.
- [11] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERFORMANCE benchmark for semantic and generative capabilities," in *ACL*, 2022, pp. 8479–8492.
- [12] J. Shi, D. Berrebbi, W. Chen, E.-P. Hu, W.-P. Huang, H.-L. Chung, X. Chang, S.-W. Li, A. Mohamed, H. yi Lee, and S. Watanabe, "ML-SUPERB: Multilingual Speech Universal PERFORMANCE Benchmark," in *INTERSPEECH*, 2023, pp. 884–888.
- [13] Y.-H. Wang, H.-Y. Chen, K.-W. Chang, W. Hsu, and H.-Y. Lee, "Minisuperb: Lightweight benchmark for self-supervised speech models," in *ASRU*, 2023, pp. 1–8.
- [14] Y. Tseng, L. Berry, Y.-T. Chen, I.-H. Chiu, H.-H. Lin, M. Liu, P. Peng, Y.-J. Shih, H.-Y. Wang, H. Wu, P.-Y. Huang, C.-M. Lai, S.-W. Li, D. Harwath, Y. Tsao, A. Mohamed, C.-L. Feng, and H.-Y. Lee, "AV-SUPERB: A multi-task evaluation benchmark for audio-visual representation models," in *ICASSP*, 2024, pp. 6890–6894.
- [15] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, "EMO-SUPERB: An in-depth look at speech emotion recognition," *arXiv preprint arXiv:2402.13018*, 2024.
- [16] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.
- [17] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *ICASSP*, 2022, pp. 356–360.
- [18] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings," in *SLT*, 2021, pp. 809–816.
- [19] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Odyssey*, 2020, pp. 433–439.
- [20] B. Veluri, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Look once to hear: Target speech hearing with noisy examples," in *CHI*, 2024, pp. 1–16.
- [21] T. Moriya, H. Sato, T. Ochiai, M. Delcroix, and T. Shinozaki, "Streaming target-speaker ASR with neural transducer," *arXiv preprint arXiv:2209.04175*, 2022.
- [22] J. Peng, M. Delcroix, T. Ochiai, O. Plhot, S. Araki, and J. Černocký, "Target speech extraction with pre-trained self-supervised learning models," in *ICASSP*, 2024.
- [23] Z. Huang, D. Raj, P. García, and S. Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *ICASSP*, 2023, pp. 1–5.
- [24] J. Peng, O. Plhot, T. Stafylakis, L. Mosner, L. Burget, and J. H. Černocký, "Improving speaker verification with self-pretrained transformer models," in *INTERSPEECH*, 2023, pp. 5361–5365.
- [25] W. Zhang and Y. Qian, "Weakly-supervised speech pre-training: A case study on target speech recognition," in *INTERSPEECH*, 2023, pp. 3517–3521.
- [26] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quiry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *INTERSPEECH*, 2020, pp. 140–144.
- [27] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark: A reproducible framework for assessing self-supervised representation learning from speech," in *INTERSPEECH*, 2021, pp. 1439–1443.
- [28] T. Parcollet, H. Nguyen, S. Evain, M. Z. Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli *et al.*, "LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of French speech," *Computer Speech & Language*, vol. 86, p. 101622, 2024.
- [29] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "HAER: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, 2022, pp. 125–145.
- [30] A. Conneau, A. Bapna, Y. Zhang, M. Ma, P. von Platen, A. Lozhkov, C. Cherry, Y. Jia, C. Rivera, M. Kale, D. van Esch, V. Axelrod, S. Khanuja, J. Clark, O. Firat, M. Auli, S. Ruder, J. Riesa, and M. Johnson, "XTREME-S: Evaluating cross-lingual speech representations," in *INTERSPEECH*, 2022, pp. 3248–3252.
- [31] T. Ashihara, M. Delcroix, Y. Ijima, and M. Kashino, "Unveiling the linguistic capabilities of a self-supervised speech model through cross-lingual benchmark and layer-wise similarity analysis," *IEEE Access*, vol. 12, pp. 98 835–98 855, 2024.
- [32] Z. Huang, D. Raj, P. García, and S. Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *ICASSP*, 2023, pp. 1–5.
- [33] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.
- [34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.
- [35] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," in *INTERSPEECH*, 2019, pp. 1368–1372.
- [36] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP*, 2020, pp. 7284–7288.
- [37] J. Peng, M. Delcroix, T. Ochiai, O. Plhot, T. Ashihara, S. Araki, and J. Černocký, "Probing self-supervised learning models with target speech extraction," in *ICASSP SASB Workshops*, 2024.
- [38] M. Delcroix, K. Kinoshita, T. Ochiai, K. Zmolikova, H. Sato, and T. Nakatani, "Listen only to me! How well can target speech extraction handle false alarms?" in *INTERSPEECH*, 2022, pp. 216–220.
- [39] K. Heafield, "KenLM: Faster and smaller language model queries," in *WMT*, 2011, pp. 187–197.
- [40] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2Letter++: A fast open-source speech recognition system," in *ICASSP*, 2019, pp. 6460–6464.
- [41] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP*, 2020, pp. 691–695.
- [42] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.