Interspeech 2024
1-5 September 2024, Kos, Greece

SA-WavLM: Speaker-Aware Self-Supervised Pre-training for Mixture Speech

Jingru Lin[1], *Meng Ge*[1,*], *Junyi Ao*[3], *Liqun Deng*[2], *Haizhou Li*[1,3]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Huawei Noah's Ark Lab [3]School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
jingrulin@u.nus.edu, {gemeng, haizhou.li}@nus.edu.sg

Abstract

It was shown that pre-trained models with self-supervised learning (SSL) techniques are effective in various downstream speech tasks. However, most such models are trained on single-speaker speech data, limiting their effectiveness in mixture speech. This motivates us to explore pre-training on mixture speech. This work presents SA-WavLM, a novel pre-trained model for mixture speech. Specifically, SA-WavLM follows an "extract-merge-predict" pipeline in which the representations of each speaker in the input mixture are first extracted individually and then merged before the final prediction. In this pipeline, SA-WavLM performs speaker-informed extractions with the consideration of the interactions between different speakers. Furthermore, a speaker shuffling strategy is proposed to enhance the robustness towards the speaker absence. Experiments show that SA-WavLM either matches or improves upon the state-of-the-art pre-trained models.

Index Terms: self-supervised learning, extraction, separation, enhancement, speech recognition

1. Introduction

Self-supervised learning (SSL) based pre-training have greatly advanced over the past few years [1, 2, 3]. By leveraging on a large amount of unlabeled speech data, the pre-trained models can produce universal speech representations that benefit a wide range of downstream applications, such as speech recognition, speaker verification, acoustic word embeddings etc [4, 5, 6]. However, many existing pre-trained models, e.g., wav2vec 2.0 [7] and HuBERT [8], are designed for clean speech, limiting their effectiveness in handling the mixture speech.

Mixture speech, where multiple speakers can speak simultaneously, presents a notably challenging scenario. WavLM [9] emerged as an early attempt to address this limitation. Given simulated mixture speech, WavLM is expected to perform denoising and masked prediction of the clean speech. A constraint of mixing portion to be less than 50% is imposed during mixture simulation, and a primary speaker is defined based on a longer speech duration. During pre-training, WavLM focuses only on masked prediction of the primary speaker's speech while disregarding other interfering speakers. Similar approaches, such as those seen in [10, 11], define the primary speaker as the solo speaker in non-speech background noises. Consequently, the resulting representations contain only information about a single speaker. Although these strategies help pre-trained models generalize better to mixture speech, subsequent research has suggested that these pre-trained models are tailored for single-speaker speech and may be suboptimal for tasks involving mixture speech [12], where all speakers are equally important.

The capability to process mixture speech is crucial for the well-known cocktail party problem [13, 14]. It is therefore important to design pre-trained models to handle mixture speech. Recently, several studies have attempted to extend the applicability of pre-trained models to mixture speech. In [15], both single-speaker and two-speaker mixture speech are simulated as input during the pre-training, and the pre-trained model learns to predict the masked timestamps of all clean speeches given the input mixture speech. A similar training strategy is extended to mixture speech with even more speakers in [12], by using a permutation invariant training (PIT) loss [16]. Both studies devise a pre-training scheme to model the mixture representations given the mixture input. As they are concerned about all speakers in the mixture speech, the pre-trained model's ability to handle mixture speech is greatly enhanced.

While the above-mentioned methods are effective, this work explores another promising direction: leveraging additional speaker information. The use of speaker information has show efficacy in various tasks, including speaker-attributed automatic speech recognition (ASR) [17, 18] and speaker diarization (SD) tasks [19]. In [17], speech recognition and speaker identification are jointly performed for multi-speaker speech scenarios. In [19], the speech activities of all speakers in the input mixture speech are estimated given the speaker profile of each speaker. These methods demonstrate the effectiveness of integrating the speaker information into the model for addressing mixture speech challenges.

Inspired by the ideas discussed above, we propose SA-WavLM, a novel speaker-aware self-supervised pre-trained model designed to better handle the mixture speech. In contrast to WavLM [9], which focuses solely on a primary speaker that might neglect others, our SA-WavLM addresses all speakers within the input mixture speech. To achieve this, SA-WavLM adopts an "extract-merge-predict" pre-training pipeline. Beginning with the "extract" phrase, SA-WavLM obtains individual representations of each speaker in the mixture speech, conditioned on the corresponding speaker information. The speaker's information is provided by speaker embeddings, which are injected into SA-WavLM by conditional layer normalization. Next, in the "merge" phrase, SA-WavLM combines the individual representations and fosters interactions between different speakers via a Speaker Merge Block. Finally, SA-WavLM will "predict" pseudo labels for each speaker. Furthermore, we introduce a speaker shuffling strategy to ensure SA-WavLM's

This work is supported by 1) Huawei Noah's Ark Lab; 2) Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002); 3) Shenzhen Science and Technology Program ZDSYS20230626091302006; 4) National Research Foundation Singapore under its AI Singapore Programme grant number AISG2-TC-2022-004. * Corresponding author.

10.21437/Interspeech.2024-787

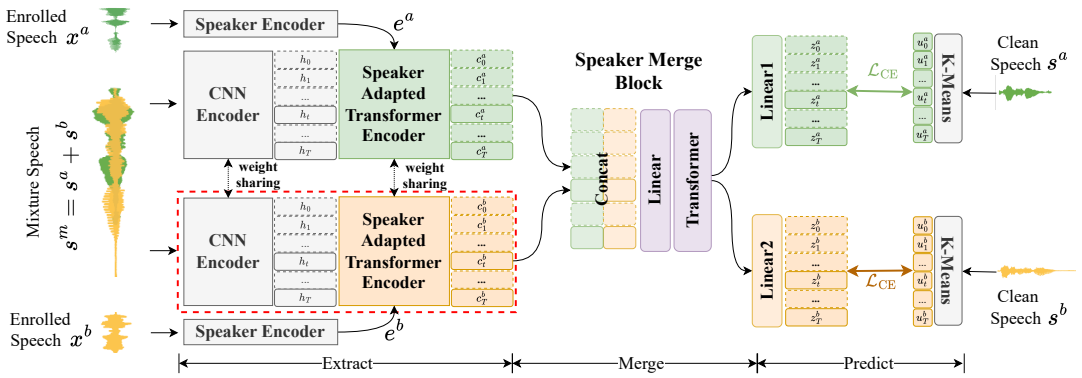

Figure 1: *The overview of SA-WavLM architecture and the "extract-merge-predict" pipeline. Given the input mixture speech, the proposed model first extracts the individual representations for each speaker in the input. Subsequently, the Speaker Merge Block merges the individual representations and models their interactions before making the final prediction. In downstream tasks, only the "extract" stage is used, which is the part within the red dotted box.*

invariance to speaker order and absence. This pre-training pipeline eliminates the need for constraints on simulated mixture data, as seen in WavLM, and empowers SA-WavLM with enhanced interference removal ability, even in scenarios where the target speaker is non-primary in the mixture.

2. WavLM

Our work is most related to WavLM [9], a self-supervised speech pre-trained model that combines denoising and masked speech prediction during pre-training. WavLM consists of a convolutional neural network (CNN) $\mathcal{F}(\cdot)$ and a Transformer encoder $\mathcal{G}(\cdot)$. Given a mixture speech s^m simulated from the clean speeches s^a of speaker a and s^b of speaker b, WavLM first extracts the frame-level speech representations $H^m = [h_1^m, h_2^m, ..., h_T^m]$ through $\mathcal{F}(\cdot)$. These representations are the passed through a partial masking process $\mathcal{M}(\cdot)$ and fed into $\mathcal{G}(\cdot)$. Let us assume speaker a is the primary speaker whose speech s^a is longer than s^b. WavLM performs denoising for speaker a to get the contextual representations $C^a = [c_1^a, c_2^a, ..., c_T^a]$ that corresponds to speech s^a. The whole pipeline can be formalized as

$$C^a = [c_1^a, c_2^a, ..., c_T^a] = \mathcal{G}\left[\mathcal{M}(\mathcal{F}(s^m))\right] \tag{1}$$

In the masked prediction process, the generated C^a should be able to predict the pseudo labels $U^a = [u_1^a, u_2^a, ..., u_T^a]$ for the masked frames in s^a. The masked speech prediction loss during the pre-training is the cross-entropy (CE)

$$\mathcal{L}_m = \mathcal{L}_{\text{CE}}(C^a, U^a) = \sum\nolimits_{t \in O} \log\, p_t(\, u_t^a \mid c_t^a) \tag{2}$$

where O denotes the set of masked indices in H^m. An offline clustering model (i.e. k-means) is used to get the pseudo labels.

In Eq. (1), s^a is assumed to have a longer speech duration in s^m. In fact, during the mixture simulation for pre-training WavLM, a constraint is applied such that the mixing portion must be less than 50%. This constraint ensures that the speech from the primary speaker is always longer than the other speech, thereby informing the model about the primary speaker (to retain) and the interfering speaker (to remove). In practice, alternative primary speaker definitions, such as the speaker with higher energy or the solo speaker in noisy speech (clean speech with background non-speech noises), are viable too.

3. Proposed SA-WavLM

In WavLM, only the primary speaker is concerned. However, the definition of a primary speaker may not align with multi-speaker speech tasks, including speech diarization, speech separation, and multi-speaker speech recognition, where all speakers in the mixture speech are equally important.

Building upon WavLM, we propose SA-WavLM to address this limitation. Figure 1 illustrates the overview of SA-WavLM architecture and the "extract-merge-predict" pipeline. Given the mixture speech $s^m = s^a + s^b$, where s^a and s^b are two clean speeches from speakers a and b, we first "extract" representations of speakers a and b individually by the Speaker Adapated Transformer Encoder (SATE) denoted as $\mathcal{G}_{\text{SATE}}(\cdot)$, replacing the $\mathcal{G}(\cdot)$ in WavLM. The individual representations are then "merged" using a Speaker Merge Block (SMB), which facilitates their interactions. Finally, the training objective of SA-WavLM is to "predict" the pseudo labels of speeches s^a and s^b. In addition, to ensure the model's invariance towards speaker order and absence, we introduce a speaker shuffling strategy.

3.1. Speaker Adapted Transformer Encoder (extract)

Speaker Adapted Transformer Encoder (SATE) is designed to extract the contextual representation for the target speaker. Given speaker embedding e^k of the target speaker $k \in \{a, b\}$ and masked frame-level representation $H^m = [h_1^m, h_2^m, ..., h_T^m]$ of s^m extracted by $\mathcal{F}(\cdot)$, SATE, denoted by $\mathcal{G}_{\text{SATE}}$, is trained to extract the contextual representation $C^k = [c_1^k, c_2^k, ..., c_T^k]$ that corresponds to s^k. This is formulated as

$$C^k = \left[c_1^k, c_2^k, ..., c_T^k\right] = \mathcal{G}_{\text{SATE}}\left[\mathcal{M}(\mathcal{F}(s^m)), e^k\right], k \in \{a, b\} \tag{3}$$

SATE consists of a Speaker Adapated Transformer Layer (SATL) and multiple Vanilla Transformer Layers (VTLs). Here the VTL in SATE has the same structure as the Transformer layer in the Transformer encoder $\mathcal{G}(\cdot)$ of WavLM. The main extraction process of SATE is performed through SATL, which replaces the traditional layer normalization in VTL with conditional layer normalization (CLN). In VTL, the layer normalization operation is given by

$$H_{\text{VTL}} = \frac{H^m - E\left[H^m\right]}{\sqrt{Var\left[H^m\right] + \epsilon}} \otimes \gamma + \beta, \tag{4}$$

where $E\left[H^m\right]$ and $Var\left[H^m\right]$ are the mean and variance of the input representation H^m, and γ and β are learnable weight and bias for applying the element-wise affine transformation.

To extract the representations of speech s^k given e^k, the CLN in SATL replaces γ with speaker-specific scaling, i.e.

$$H_{\text{SATL}} = \frac{H^m - E\left[H^m\right]]}{\sqrt{Var[H^m] + \epsilon}} \otimes \left[w(e^k) \cdot \gamma + \theta(e^k)\right] + \beta, \quad (5)$$

where $w(\cdot)$ and $\theta(\cdot)$ are linear projections to project the speaker embedding e^k to the feature dimension of H^m to perform the element-wise multiplication. In fact, CLN modulates the normalization output via a specific speaker's latent representation.

3.2. Speaker Merge Block (merge)

Motivated by TS-VAD [19], it is believed that interaction between different speakers is important for better extraction and separation abilities. Therefore, in SA-WavLM, we design a Speaker Merge Block (SMB) to allow interaction between different speakers. The structure of SMB is also shown in Figure 1.

SMB first concatenates C^a and C^b of speaker a and b, extracted respectively by SATE as shown in Eq. (3), along the feature dimension. Later, the concatenated representations are projected down by a linear layer. A single VTL, which has the same structure as the VTL in SATE, is deployed to model the down-projected representations and their interactions, generating the mixture contextual representations C^m

$$C^m = \text{VTL}(\text{Linear}(\text{Concat}(C^a, C^b))) \quad (6)$$

Note that the SMB is used during pre-training only to facilitate the interactions. In downstream applications, SMB is removed and only CNN and SATE are used.

3.3. Prediction and training objective (predict)

Unlike Cocktail HuBERT [12] where PIT [16] is needed to find the optimal alignment between the predictions and pseudo labels, SA-WavLM determines the order of predictions based on the order of speaker embeddings injected. That is, if the concatenation order in Eq. (6) is a followed by b, Linear1 predicts the pseudo labels of clean speech s^a and Linear2 predicts that of clean speech s^b and vice versa. Here, the former order, i.e. a followed by b, is used for illustration

$$Z^a = \text{Linear1}(C^m), \quad Z^b = \text{Linear2}(C^m) \quad (7)$$

The final masked speech prediction loss is the summation of Cross Entropy (CE) losses for the two speakers

$$\mathcal{L}_m = \sum_{k\in\{a,b\}} \mathcal{L}_{\text{CE}}(Z^k, U^k) = \sum_{k\in\{a,b\}} \sum_{t\in O^k} \log\, p_t(\, u_t^k \mid z_t^k) \quad (8)$$

3.4. Speaker shuffling strategy

To keep the model invariant to speaker order, we deploy a speaker shuffling strategy to shuffle the order of speaker embeddings and the corresponding pseudo labels. In our mixture simulation, we have two-speaker scenarios and one-speaker scenarios. Two-speaker scenarios include two-speaker overlapped speech $s^m = s^a + s^b$ and noisy two-speaker overlapped speech $s^m = s^a + s^b + n$, where n is the background noise. One-speaker scenarios include clean speech s^a and noisy single-speaker speech $s^n = s^a + n$. For two-speaker scenarios, the speaker embeddings e^a and e^b are used directly. While for one-speaker scenario, we will use speaker embeddings e^a and, either 1) e^c from a random speaker that is different from speaker a, with a probability of α, or 2) e^s, a learnable vector indicating non-speaker existence, with a probability of $(1-\alpha)$. Subsequently, we shuffle the order of all the speaker embeddings injected into SATE, which determines the order of the Concat operation in Equation 6. This strategy aims to enhance the model's insensitivity towards the speaker order and robustness towards the speaker absence, thereby improving the accuracy of extraction. The pseudo labels corresponding to e^c or e^s will be S, which is a vector of silence tokens.

4. Experimental setups

4.1. Training datasets

Following [15], we pre-train our model on the data simulated from the 960-hour LibrSpeech [20] and the DNS Challenge's noise datasets [21]. Due to limited computational resources, only one- and two-speaker scenarios are simulated. This includes clean speech, noisy single-speaker speech, two-speaker overlapped speech, and noisy two-speaker overlapped speech, which are generated through a dynamic mixing strategy [22]. The enrolled speech of the target speaker is randomly selected from the LibriSpeech corpus, and it should be different from the one used to simulate the mixture speech. The speaker embedding is generated from the selected enrolled speech using the pre-trained CAM++ [23]. All speeches are sampled at 16kHz.

4.2. Models and training details

All pre-training experiments are conducted using Fairseq toolkit [24]. Our model is based on WavLM Base [9], while replacing a Vanilla Transformer Layer (VTL) with a Speaker Adapted Transformer Layer (SATL) and adding a Speaker Merge Block (SMB). SATL can replace any VTL in the Transformer encoder, but here we only replace the first layer. The parameters of CNN encoder and SATE are initialized from the pre-trained WavLM Base, except for w and θ of SATL. For the SMB, the weights are randomly initialized. Training takes 400k steps with a learning rate of 7e-5, using the pseudo labels generated from the 9th transformer layer of the HuBERT Base model [12]. The probability α used in the speaker shuffling strategy is set to 0.5. Other hyperparameters are consistent with those of the WavLM Base. Note that although SA-WavLM has the largest model size as shown in Table 1, the additional SMB is removed for downstream tasks. This leaves our model size with 94.97M parameters, which is comparable to other pre-trained models (e.g. WavLM Base, HuBERT Base etc).

4.3. Downstream evaluations

To evaluate the effectiveness of our models on the cocktail party problem, we evaluated our models on different mixture speech tasks including speech enhancement (SE), separation (SS), diarization (SD), extraction (Ext) and multi-speaker speech recognition (ASR). For SE, the dataset used is Voicebank-DEMAND [25]. For the remaining, the different subsets of Libri2Mix are used. In particular, SD uses the "mix-both max mode", ASR uses the "mix-clean max mode" and SS and Ext use the "mix-clean min mode". All the datasets in downstream evaluations have 16kHz sample rates.

For the baselines, we select the state-of-the-art pre-trained models including wav2vec 2.0 [7], HuBERT [8], WavLM [9], Wang et.al [15] and Cocktail HuBERT [12]. For all the baselines, Base sizes, which are comparable to SA-WavLM, are used. Among them, wav2vec 2.0 and HuBERT are trained on clean speech only, while the rest are trained on mixture speech.

5. Experiments

5.1. Results on SUPERB benchmark

We conduct evaluations on SUPERB [26], the well-known benchmark that unifies evaluations on various speech processing tasks. It provides lightweight downstream networks that make it easy to compare and draw insights across different pre-trained models. Assessments are made for the provided mixture speech tasks: SE, SS and SD. We report Perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) for SE, scale-invariant signal-to-distortion ratio

Available at https://github.com/JorisCos/LibriMix

Table 1: *Universal representation evaluation on SUPERB.*

Methods	#Param	SE		SS	SD
		PESQ↑	STOI↑	SI-SDRi↑	DER↓
FBANK	–	2.55	93.60	9.23	10.05
Wav2vec2.0 Base [8]	95.04M	2.55	93.9	9.77	6.08
HuBERT Base [8]	94.68M	2.58	93.90	9.36	5.88
Wang et.al [15]	95.02M	–	–	10.59	–
C-HuBERT Base [12]	96.00M	**2.63**	94.00	11.08	2.77
WavLM Base [9]	94.70M	2.58	94.00	10.37	4.55
SA-WavLM (Ours)	103.7M	2.62	**94.18**	**11.13**	**1.88**

Table 2: *Results on multi-speaker speech recognition.*

Pre-trained Model	Spk. Embedding	WER (%)	
		w/o LM	w/ LM
HuBERT Base [8]	✗	22.70	15.60
	✓	17.42	14.88
WavLM Base [9]	✗	15.97	10.38
	✓	13.30	11.36
SA-WavLM (Ours)	✓	**8.39**	**6.49**

improvement (SI-SDRi) for SS, and the diarization error rate (DER) for SD. In all the experiments, the pre-trained models are frozen and only the downstream networks are fine-tuned.

Table 1 lists the evaluation results on the SUPERB benchmark. It can be observed that wav2vec 2.0 and HuBERT do not generalize well to the mixture speech tasks. With the denoising strategy, WavLM outperforms wav2vec 2.0 and HuBERT across all three tasks. Both Wang et. al and Cocktail HuBERT (referred to as C-HuBERT in Table 1) are designed for mixture speech, hence demonstrating superior performance on all three tasks. When more speakers are considered during pre-training, C-HuBERT shows further improvement on SS. Compared against all baselines, SA-WavLM either matches or outperforms across all the tasks. Notably, both Wang et.al and SA-WavLM are pre-trained with 2-speaker mixture speech only, yet SA-WavLM demonstrates better performance on SS, surpassing Wang et. al by 0.54dB SI-SDRi. This proves the effectiveness of the use of speaker cues and modeling of speaker interactions.

5.2. Multi-speaker speech recognition

Following the utterance group-based evaluation settings in [27], we evaluate SA-WavLM's performance on multi-speaker ASR. The character-level connectionist temporal classification (CTC) loss is used to fine-tune the pre-trained models except for the CNN encoder. We consider two settings: one without and one with speaker embeddings. In the former, permutation invariant training (PIT) [16] is used. In the latter, the same CLN operation as in SA-WavLM is applied to integrate speaker embeddings.

Table 2 reports the word error rate (WER) for all the pre-trained models, with and without the 4-gram language model (LM). For models trained using PIT, concatenated minimum-permutation word error rate (cpWER) [28] is reported. As shown in Table 2, WavLM, designed for mixture speech, obtains a much better performance compared to HuBERT which is designed for clean speech. When incorporating the speaker embeddings into the pre-trained models, both HuBERT and WavLM demonstrate reduced WER in settings without LM. Compared to WavLM, SA-WavLM achieves a more significant relative WER reduction of 37.4%. This can be attributed to SA-WavLM's consideration of speaker interactions, which improves its speaker discrimination ability.

5.3. Speech extraction and separation

While SUPERB offers a standard and comprehensive benchmark for the research community, many works have investigated to evaluate SSL models' capabilities in different ways, striving to push the performance boundaries further [5, 7, 8, 31, 32]. However, the ability of the pre-trained model on speech extraction and separation is relatively under-explored. In [33], discretization is first performed on the representations obtained from the frozen pre-trained model and the clean waveform is resynthesized from the discrete tokens. In our experiment, we use the representations from all layers of the frozen pre-trained models directly, as discretization could potentially result in the loss of information. The pre-trained representations of different layers are weighted-sum, upsampled and then concatenated with the supervised model's encoded features before feeding into separation/extraction modules in these models.

Table 3: *Results on speech extraction. The default stride for BSRNN is 8ms. We changed it to 5ms so that BSRNN's features can be aligned with the pre-trained features of 20ms stride.*

Extraction Model	Stride	Pre-trained Model	SDRi↑	SI-SDRi↑
BSRNN [29]	8ms	–	14.0	13.01
	5ms	–	12.39	10.06
		HuBERT Base [8]	14.83	14.11
		WavLM Base [9]	14.64	15.47
		SA-WavLM (Ours)	**17.74**	**17.3**

Table 4: *Results on speech separation, including 1%, 10%, and 100% of training data (13,900 utterances for 100%).*

Train Data	Pre-trained Model	Separation Model	SDRi↑	SI-SDRi↑
1%	–	ConvTasNet [30]	3.05	2.59
	HuBERT Base [8]		4.45	3.97
	WavLM Base [9]		7.56	7.09
	SA-WavLM (Ours)		**8.81**	**8.42**
10%	–	ConvTasNet [30]	9.73	9.16
	HuBERT Base [8]		11.08	10.67
	WavLM Base [9]		11.93	11.58
	SA-WavLM (Ours)		**13.93**	**13.62**
100%	–	ConvTasNet [30]	14.53	14.12
	HuBERT Base [8]		14.99	14.62
	WavLM Base [9]		15.95	15.6
	SA-WavLM (Ours)		**16.55**	**16.22**

Table 3 shows the results for speech extraction using BSRNN [29] that is integrated with different pre-trained models. The results indicate that BSRNN benefits from integrating with the pre-trained speech representations. SA-WavLM stands out among all the pre-trained models, achieving 3.74dB SDRi and 4.29dB SI-SDRi improvements from BSRNN (8ms).

Table 4 presents the results for speech separation using ConvTasNet [30]. In this experiment, we further investigate the effectiveness of the pre-trained models in low-resource scenarios (e.g. 1%). Consistent with the findings in Table 3, ConvTasNet benefits from the pre-trained representations across all settings (i.e., 1%, 10% and 100%). The benefits are particularly prominent in low-resource scenarios. This can be explained as pre-trained representations provide more noise-invariant contextual information. Again, among all the pre-trained models, SA-WavLM demonstrates the most promising performances.

6. Conclusion

This paper proposes SA-WavLM, a novel speaker-aware self-supervised pre-trained model for mixture speech. Pre-trained with the designed "extract-merge-predict" pipeline, SA-WavLM considers the presence of all speakers in the mixture speech and models the interactions between them. In addition, a speaker shuffling strategy is introduced to ensure the model's invariance towards speaker order and absence. Experimental results show that SA-WavLM has an enhanced ability to remove interference across various mixture speech tasks.

7. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[2] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.

[3] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 50 221–50 236.

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[5] 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 *Proc. ICASSP*. IEEE, 2022, pp. 6147–6151.

[6] J. Lin, X. Yue, J. Ao, and H. Li, "Self-Supervised Acoustic Word Embedding Learning via Correspondence Transformer Encoder," in *Proc. INTERSPEECH 2023*, 2023, pp. 2988–2992.

[7] 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.

[8] 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.

[9] 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.

[10] Q.-S. Zhu, L. Zhou, J. Zhang, S.-J. Liu, Y.-C. Hu, and L.-R. Dai, "Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[11] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *Proc. ICASSP*. IEEE, 2022, pp. 7097–7101.

[12] M. Fazel-Zarandi and W.-N. Hsu, "Cocktail hubert: Generalized self-supervised pre-training for mixture and single-source speech," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[13] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 40–63, 2018.

[14] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[15] T. Wang, X. Chen, Z. Chen, S. Yu, and W. Zhu, "An adapter based multi-label pre-training for speech separation and enhancement," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[16] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*. IEEE, 2017, pp. 241–245.

[17] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-End Speaker-Attributed ASR with Transformer," in *Proc. Interspeech 2021*, 2021, pp. 4413–4417.

[18] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *Proc. ICASSP*. IEEE, 2022, pp. 8082–8086.

[19] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. Interspeech 2020*, 2020, pp. 274–278.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[21] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[22] W. Zhang and Y. Qian, "Weakly-supervised speech pre-training: A case study on target speech recognition," in *Proc. Interspeech 2023*, 2023.

[23] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," in *INTERSPEECH*, 2023.

[24] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, W. Ammar, A. Louis, and N. Mostafazadeh, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[25] C. Veaux, J. Yamagishi, and S. 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.

[26] S. wen 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. tik 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 *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[27] Z. Huang, D. Raj, P. García, and S. Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[28] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[29] J. Yu, H. Chen, Y. Luo, R. Gu, and C. Weng, "High Fidelity Speech Enhancement with Band-split RNN," in *Proc. Interspeech 2023*, 2023, pp. 2483–2487.

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

[31] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.

[32] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. ICASSP*. IEEE, 2022, pp. 6922–6926.

[33] J. Shi, X. Chang, T. Hayashi, Y.-J. Lu, S. Watanabe, and B. Xu, "Discretization and re-synthesis: an alternative method to solve the cocktail party problem," *arXiv preprint arXiv:2112.09382*, 2021.