INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving Speaker Verification with Self-Pretrained Transformer Models

Junyi Peng[1], *Oldřich Plchot*[1], *Themos Stafylakis*[2], *Ladislav Mošner*[1], *Lukáš Burget*[1], *Jan Černocký*[1]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]Omilia - Conversational Intelligence, Athens, Greece

Abstract

Recently, fine-tuning large pre-trained Transformer models using downstream datasets has received a rising interest. Despite their success, it is still challenging to disentangle the benefits of large-scale datasets and Transformer structures from the limitations of the pre-training. In this paper, we introduce a hierarchical training approach, named self-pretraining, in which Transformer models are pretrained and finetuned on the same dataset. Three pre-trained models including HuBERT, Conformer and WavLM are evaluated on four different speaker verification datasets with varying sizes. Our experiments show that these self-pretrained models achieve competitive performance on downstream speaker verification tasks with only one-third of the data compared to Librispeech pretraining, such as VoxCeleb1 and CNCeleb1. Furthermore, when pre-training only on the VoxCeleb2-dev, the Conformer model outperforms the one pre-trained on 94k hours of data using the same fine-tuning settings.

Index Terms: speaker verification, pre-trained speech transformer model, pre-training,

1. Introduction

In recent years, deep neural networks based speaker recognition systems, such as x-vector [1], ECAPA-TDNN [2], and ResNet [3], have achieved state-of-the-art performance on various speaker verification tasks. These models are typically trained in a supervised manner from scratch on a large-scale dataset like VoxCeleb [4, 5] and CNCeleb [6], which contain millions of speech segments from thousands of speakers. However, due to their heavy parameterization, with tens of millions of parameters, optimizing these models can be challenging in data-restricted scenarios such as a completely new channel, language, or both in a low-resource domain.

A promising solution is to leverage general-purpose pre-trained Transformer models such as Wav2Vec [7], HuBERT [8], and WavLM [9], which are pre-trained on the large unlabelled test using self-supervised objective, like masked frame prediction, to learn general acoustic representations. To adapt these pre-trained models to specific downstream tasks, a common approach is to perform fine-tuning of the entire pre-trained model with a task-oriented back-end using labeled downstream

The work was supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Czech Ministry of Education, Youth and Sports project no. LTAIN19087 "Multi-linguality in speech technologies", and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports through the e-INFRA CZ (ID:90140).

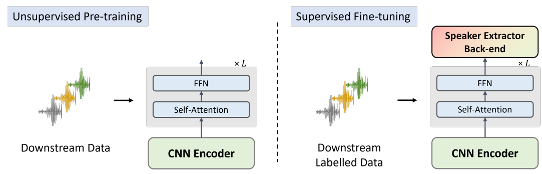

Figure 1: *In contrast to generalist pretraining, where a pre-trained model is first trained on a large-scale corpus (e.g., Librispeech) before fine-tuning on downstream task, self-pretraining utilizes the model pretrained on the fine-tuning dataset itself.*

datasets. For example, in [10], an average pooling and a fully-connected layer were cascaded to the top of the Wav2vec 2.0 model to extract speaker and language embedding. A stronger performance was achieved on the speaker verification task using back-end, i.e. ECAPA-TDNN, in [11], which fed the frame-by-frame input to the back-end as a weighted combination of the outputs of the individual layers of a pre-trained Transformer model. To shorten the training time, a more lightweight back-end was developed in [12], which consists of an attention layer and a linear layer to extract speaker representations. These simple approaches have led to impressive results in the fields of language recognition [13] and speaker verification tasks. Despite their success, it remains challenging to disentangle the advantages offered by large-scale datasets and Transformer structures from the limitations of the pre-training paradigm [14].

The challenges can be summarized as three-fold considering the mask prediction objective is the optimization target for Transformer models. (1) To what extent large-scale datasets are crucial for performance improvements? Particularly, in cases where sufficient downstream data is available, is there still a need to employ pre-trained models? For instance, the HuBERT base is pretrained on 960-hour Librispeech, while the fine-tuning dataset, VoxCeleb2, comprises over 2,000 hours of speech. (2) How does domain shift affect pre-trained models? Currently, most pre-trained models are trained on English read-aloud data, raising questions about their effectiveness when fine-tuned in non-English language and under in-the-wild conditions. For example, a pre-trained WavLM Base model may not perform well when fine-tuned on a Chinese speaker recognition corpus [6, 15]. The phonetic and linguistic properties of different languages can significantly vary, and models may not capture these nuances effectively. (3) What is the impact of the Transformer structure on downstream speaker verification tasks? The widely-used HuBERT structure is inherited from BERT [16], which is originally designed for natural language processing tasks. It still remains unclear whether this structure is optimal for downstream speaker verification tasks. In addition, recent studies have shown performance improvement in

 10.21437/Interspeech.2023-453

speech recognition using alternative architectures such as Conformer [17] and WavLM [9].

To tackle these challenges, in this paper, we explore a hierarchical training paradigm, named *self-pretraining*, which aims to reduce the domain shift between pre-training and fine-tuning stages and improve the discriminability of extracted speaker representations. Specifically, as shown in Fig 1, we construct the training process based on the following steps: firstly, the Transformer model is directly trained from scratch on the downstream datasets with the HuBERT-style mask prediction loss following the self-supervised manner. Then, a learnable task-oriented back-end is introduced to cascade the pre-trained Transformer model to predict the speaker identity based on a given utterance while the parameters of the pre-trained Transformer model are frozen. Finally, the entire Transformer model is jointly fine-tuned with the back-end to further boost the performance. Overall, our contributions are the following:

- We demonstrate the effectiveness of self-pretraining across four different pre-training/fine-tuning datasets, with two cascaded back-end structures (ECAPA-TDNN and Multi-head Factorized Attentive Pooling (MHFA) [12]) for SV tasks.
- A detailed analysis of cross-language performance of models pretrained on one language and then fine-tuned on another language for a SV task.
- We conduct a comprehensive study on the SV performance of pre-trained models with different architectures including HuBERT, Conformer and WavLM (Transformer with gated position encoding and data augmentation).
- Extensive experiments on both VoxCeleb and CNCeleb corpus show that our methods obtain a significant improvement over Librispeech pre-trained systems.

2. Self-pretraining for Speaker Verification

In this section, we formalize each of the self-pretraining components in detail, as illustrated in Fig 1. Our self-pretraining approach explores the potential of pre-trained models by utilizing the same dataset for both pre-training and fine-tuning while minimizing domain-shift between two datasets. Furthermore, the model is expected to learn in-domain representations that effectively capture the phonetic and acoustic characteristics of speech, as well as contextual information that enables it to predict masked frames. These learned representations can potentially be beneficial in the field of speaker verification, where the model is supposed to accurately distinguish between subtle speech signal variations to identify the speaker. Although similar observations have been reported in computer vision [14] and natural language processing [18], it is essential to validate these findings in the field of speaker verification, where the use of pretrained models has attracted increasing attention.

2.1. Unsupervised Pre-training

In this study, we utilize the learned representations from three different architectures, including HuBERT, Conformer, and WavLM, respectively, as input features to extract speaker embeddings. These models take raw waveforms as input, and the backbone consists mainly of a convolutional neural network (CNN) encoder and a Transformer encoder [19].

During pre-training, the model consumes masked frame-level features to predict a predetermined discrete target. This masked speech prediction objective is applied only to the masked frame-level features, with the model expected to correctly infer the targets of masked frames through the remaining unmasked ones. Hence, the model is enforced to learn acoustic and phonetic information over continuous input speech.

We follow the iterative re-clustering and re-training approach described in [8]. In the initial iteration, targets are assigned by clustering the MFCC features of the training data. In subsequent iterations, a new set of training targets is generated by clustering the latent representations produced by the first iteration trained model.

2.2. Supervised Fine-tuning

Transformer models trained on thousands of hours of speech data have been shown to effectively represent the structure of speech and generalize well to various downstream tasks [20, 21]. However, recent studies have suggested that the advantage of fine-tuning pre-trained models over deep speaker extractors trained from scratch is insignificant or even non-existent, based on analysis of the last layer's representations [10, 22, 23]. A potential explanation is that the speech prediction objective of the pre-training stage encourages the model to discover and internally represent various acoustic units that naturally correspond to context-dependent phones in the last layers [24]. As a result, the top layers, which are closer to the objective of the pre-training stage, are typically most beneficial for automatic speech recognition. In contrast, speaker verification tasks mainly rely on low- and mid-level features that carry most of the information about speaker identity. Thus, attempting to obtain the speaker representations from the last Transformer layer's output may be sub-optimal for speaker verification.

Based on the aforementioned, to fully utilize the hierarchical representations, we follow the method introduced in [25] and we assign a set of learnable weights $\mathbf{w} = \{w_l\}_{l=0}^{L}$ to the layer-wise outputs as:

$$\mathbf{y} = f_b\left(\sum_{l=0}^{L} w_l \mathbf{H}_l; \mathbf{\Theta}_b\right), \tag{1}$$

where l-th layer's outputs are $\mathbf{H}_l \in \mathbb{R}^{T\times D}$ ($\mathbf{H}_0$ denotes the CNN outputs), L denotes the total number of Transformer layers, T is the length of frames, D is the feature dimension, $\mathbf{\Theta}_b$ are the parameters of the backend, and $\mathbf{y}$ is the extracted speaker embedding. Regarding the back-end, in this study, we utilize two different structures: ECAPA-TDNN and MHFA. In detail, for ECAPA-TDNN we replace the original FBank features inputs with hierarchical representations obtained from the pre-trained Transformer model. For MHFA, we modify eq. (1) so that *two sets of layer-wise weights* ($\mathbf{w}^K$ and $\mathbf{w}^V$) are employed to generate attention maps and compressed features, respectively [12]. Then, the speaker embedding $\mathbf{y}$ is formed by aggregating over frames and projecting the vector to a lower-dimensional space using a linear layer.

3. Experiments

3.1. Setup

Data-sets: The SV performance is evaluated on the VoxCeleb [4, 5] and CNCeleb [15, 6] corpora, both are widely used text-independent speaker verification datasets. For VoxCeleb, the training set is VoxCeleb1, and the development set of VoxCeleb2, respectively. The performance is evaluated on *VoxCeleb1-O*, *VoxCeleb1-E*, and *VoxCeleb1-H* trials. For CNCeleb, the model is trained on different training datasets, namely, *CNCeleb1* and *CNCeleb1+2*, containing 800

Table 1: *Performance comparison of HuBERT Base models pretrained on various datasets, including Librispeech, VoxCeleb and CNCeleb.* ***RandomInit*** *indicates both Transformer and back-end are jointly trained from scratch on a fine-tuning dataset.* ***Librispeech*** *denotes the models are pre-trained on the Librispeech dataset from Huggingface.* ***Self-Pretrain*** *means model pre-training and finetuning on the same dataset.* ***Frozen*** *suggests the pretrained models are kept frozen during the fine-tuning, while* ***Learnable*** *denotes the pre-trained model jointly optimized with the back-end. It is noted that here we use* ***ECAPA-TDNN*** *as the back-end for all systems.*

Fine-Tuning Dataset	Hours(hr)	Test Dataset	RandomInit	Librispeech (960 hr)		Self-pretraining	
				Frozen	Learnable	Frozen	Learnable
VoxCeleb1	338	VoxCeleb1-O	10.14	2.43	2.22	2.14	1.91
VoxCeleb2-dev	2500	VoxCeleb1-O	1.88	1.36	1.18	1.13	1.04
CNCeleb1	270	CNCeleb-E	21.08	14.64	11.99	11.90	10.86
CNCeleb1+2	1200	CNCeleb-E	15.33	10.65	9.70	8.80	8.89

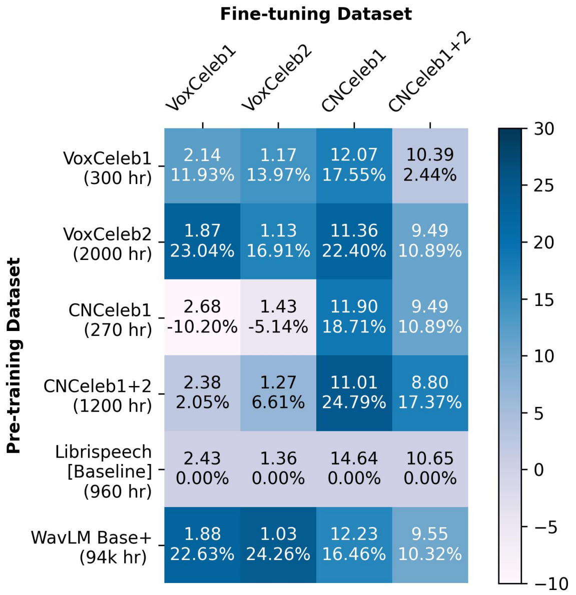

Figure 2: *Equal Error Rates (EERs) and improvements of models pretrained on various datasets, fine-tuned on downstream speaker verification tasks with different datasets. Each value represents the EER and relative improvement gains compared to the modeal readily available on HuggingFace pretrained on Librispeech. For systems fine-tuned on Voxceleb1 and Voxceleb2, the results are given on VoxCeleb1-O, while for systems fine-tuned on CNCeleb we report results on CNCeleb-E. During fine-tuning the pre-trained models are kept frozen.*

and 2800 speakers, respectively. *CNCeleb1+2* is a combination of CNCeleb1-dev and CNCeleb2. The evaluation part *CNCeleb-E* contains 18,849 utterances from 200 speakers.

Implementation details: In this work, we utilize three types of pre-trained BASE models: 1) The *HuBERT* models, consisting of a CNN encoder and 12 layers of Transformer with 94M parameters. The dimension of the Transformer output D_{hidden} is 768. 2) The *Conformer* models, which include 12 layers combining both convolutional neural networks and transformers, result in 170M parameters. 3) The *WavLM* models, a variant of the Transformer model that redesigns the position embedding and utilizes simulated noisy/overlapped speech to improve model's robustness. We then pre-train these models for 800k steps on VoxCeleb2-dev and CNCeleb1+2, and for 400k steps on VoxCeleb1 and CNCeleb1 using the labels generated by clustering the 6-th transformer layers outputs of 1st-iteration (250K steps) corresponding model following *fairseq* implementation. The pre-training stage takes 3-4 days on 8 A100 GPUs.

Table 2: *The performance of various Base models with different back-ends on VoxCeleb1-O. All models are self-pretrained on VoxCeleb2-dev.* ***WavLM*** *refers to a Transformer with gated relative position bias and simulated noisy/overlapped speech.*

Backbone/Param	Back-end	Finetune	EER (%)
HuBERT (94.6 M)	ECAPA-TDNN	✗	1.13
	ECAPA-TDNN	✓	1.04
	MHFA	✗	1.38
	MHFA	✓	0.88
Conformer (172.2 M)	ECAPA-TDNN	✗	1.17
	ECAPA-TDNN	✓	1.19
	MHFA	✗	1.21
	MHFA	✓	0.65
WavLM (94.7 M)	ECAPA-TDNN	✗	1.12
	ECAPA-TDNN	✓	1.09
	MHFA	✗	1.40
	MHFA	✓	0.87

For fine-tuning configurations, the total number of heads in MHFA is set to 64, the channel of ECAPA-TDNN is 1024, and the extracted speaker embedding dimension is 256 for both back-ends. We use AAM-softmax [26] in a fine-tuning stage with a margin of 0.2 and scaling of 30 for 10 epochs. To further boost performance, we adopt large margin tuning [27]. we input longer (5 seconds) waveforms and set the margin to 0.5 for additional 3 training epochs. The initial learning rate is $5e^{-4}$ and decreases by 10% per epoch for the Adam optimizer. All fine-tuning datasets are augmented by adding noise (MUSAN) and reverberation (RIR). Due to the GPU memory constraints, the mini-batch size of 100 is chosen for Base model training.

Performance Metrics: Both equal error rate (EER) and minimum detection cost function (minDCF) are employed to measure the performances of speaker verification systems. The prior target probability P_{tar} is set to 0.01 or 0.05, for DCF1 and DCF5, respectively. C_{fa} and C_{miss} are set to 1.0.

3.2. Comparison of Generalist and Self Pretraining

To investigate the performance of self-pretraining and other pretraining techniques. For each dataset, we first pretrain a HuBERT Base Transformer model from scratch and then fine-tune it on the same training dataset with speaker labels in Table 1.

Our experimental results demonstrate that the self-pretrained models consistently outperform the fine-tuned Librispeech pretrained models obtained from Huggingface, even when the amount of training data is significantly lower than that of Librispeech. Since the self-pretraining models are fine-tuned on the same dataset used for pretraining, it is unlikely that the observed performance gains are due to transfer learning from pre-training on a different dataset. Instead, the improvements may be attributed to the hierarchical training approach, which enables the models to learn in-domain representations and cap-

Table 3: *Results for speaker verification on the Voxceleb1-O data-set and extended VoxCeleb1-E and VoxCeleb-H test sets. All models are fine-tuned on VoxCeleb2-dev.* ***Pre-train*** *denotes the pre-training dataset.*

Model	Pre-train	VoxCeleb1-O			VoxCeleb1-E			VoxCeleb1-H		
		EER(%)	DCF1	DCF5	EER(%)	DCF1	DCF5	EER(%)	DCF1	DCF5
ECAPA-TDNN [28]	-	0.90	-	0.081	1.11	-	0.077	2.32	-	0.155
WavLM Base Plus- ECAPA-TDNN [9]	Mix 94k hr	0.98	-	-	1.06	-	-	2.21	-	-
WavLM Base Plus- MHFA [12]	Mix 94k hr	0.66	0.074	0.045	0.89	0.097	0.056	1.90	0.190	0.119
HuBERT Base - MHFA [12]	LS 960 hr	0.92	0.091	0.059	1.19	0.136	0.078	2.52	0.252	0.159
HuBERT Base - MHFA	Vox2-dev 2k hr	0.88	0.097	0.056	1.06	0.118	0.068	2.11	0.223	0.136
WavLM Base - MHFA	Vox2-dev 2k hr	0.87	0.091	0.053	1.04	0.114	0.072	2.23	0.227	0.133
Conformer - MHFA	Vox2-dev 2k hr	0.65	0.063	0.045	0.93	0.100	0.058	1.86	0.193	0.117

Table 4: *Comparison of different models on CNCeleb-E test set. All models are trained on the CNCeleb1+2 dataset.*

Model	EER(%)	DCF1
ResNet34-DTCF [29]	14.84	0.596
MBFA-MW [30]	9.48	0.456
ECAPA-TDNN [31]	8.93	0.504
Conformer - MHFA	7.73	0.406

ture relevant acoustic and phonetic features of the downstream dataset during pretraining. By fine-tuning on the same dataset, the cascaded back-end is able to learn more speaker-related information based on these robust features, resulting in better generalization and performance improvement. Furthermore, when we jointly train the Transformer model and the followed back-end from scratch on speaker recognition tasks, referred to as RandomInit, the worst performance is attained. It is finally observed that when the training dataset is insufficient, such as in the case of VoxCeleb1 and CNCeleb1, the performance degrades dramatically.

3.3. Analysis of Cross-Language Fine-tuning

We investigate the generalizability of pre-trained models across a range of SV tasks with different scales and languages, as illustrated in Fig 2. In detail, we explore whether models pre-trained on a particular dataset are only useful for that specific dataset or whether they can be further applied to a broader range of conditions, such as cross-language scenarios. We first take four Transformer models pre-trained on VoxCeleb1, VoxCeleb2-dev, CNCeleb1 and CNCeleb1+2, respectively. Subsequently, we fine-tune the back-end of each pre-trained Transformer on all other datasets. For the approaches fine-tuned on VoxCeleb1 and VoxCeleb2-dev, the performance is evaluated on VoxCeleb1-O; while for CNCeleb1 and CNCeleb1+2, we evaluate their performance on CNCeleb-E. The relative performance improvements and corresponding EERs are shown as a heatmap in Fig 2.

In most cases, the self-pretrained models perform better than the Librispeech pretrained models. Moreover, we observe that the performance improvement is more significant when there is a larger amount of data in a similar scenario during fine-tuning. For example, pre-training on VoxCeleb2-dev and fine-tuning on VoxCeleb1 leading to a significant performance boost. However, it is noted that the Transformer model pre-trained on CNCeleb1 performs worse with negative benefits. This might be because the limited amount of training data (200 hours) may not be sufficient to train a Transformer model with 96 million parameters effectively. Moreover, compared to VoxCeleb1 with a similar size, CNCeleb1 is collected with a variety of challenging acoustic and phonetic conditions, such as regional accents, background noise, and music interference, which can make the pre-training task more difficult.

3.4. Ablation Studies on Architectures

To demonstrate the effectiveness of self-pretraining, in Table 2, we present an analysis of the impact of various backbone architectures, including the original Transformer model, Conformer model, and WavLM model, and two back-end models (i.e. ECAPA-TDNN and MHFA).

By replacing the Transformer model with the Conformer model, a substantial improvement is observed, especially when using the MHFA back-end to joint fine-tuning with the entire pre-trained model. This suggests that the depthwise separable convolution module in the Conformer model is more effective in capturing local acoustic features compared to the original Transformer model.

Additionally, the system using MHFA is able to extract more discriminative speaker representations, yielding better performance on the VoxCeleb1-O compared to the system using ECAPA-TDNN. This may be due to the fact that the pre-trained model has already learned rich and robust acoustic representations, and the lightweight back-end with fewer parameters gives the pre-trained model more flexibility, allowing it to better adapt to the speaker verification tasks. In contrast, a heavy, multi-layer back-end, designed to be a standalone network, cannot make full use of the backbone network and propagate gradients to finetune it. As Table 2 shows, the gains from unfreezing the backbone network are marginal when ECAPA-TDNN is used.

3.5. Comparison with State-of-the-art SV systems

We compare the proposed method with other state-of-the-art SV systems in Tables 3 and 4. All pre-trained model-based approaches achieve remarkable performance compared to conventional SV systems trained from scratch, indicating that pre-trained models can effectively capture relevant acoustic and phonetic information for downstream tasks. Among the pre-trained models, our self-pretraining approach consistently outperforms the generalist one. These results show the effectiveness of self-pretraining in capturing speaker-specific information, resulting in more discriminative embeddings.

4. Conclusion

In this paper, we propose a hierarchical training approach, named self-pretraining, for pre-training Transformer models on downstream speaker verification tasks. We have shown that this approach can capture in-domain and speaker-specific information, resulting in more discriminative speaker embeddings. Our experiments show that self-pretraining outperforms generalist pretraining and achieves competitive performance with 94k hours of pretraining on downstream speaker verification tasks with solely 2k hours of downstream data. Finally, we showed that the recently introduced MHFA is superior to ECAPA-TDNN as a backend for Transformer models.

5. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech2020*, 2020, pp. 1–5.

[3] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.

[4] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[5] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[6] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[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*, 2022.

[10] 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.

[11] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[12] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. 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.

[13] Y. Lee, C. Greenberg, E. Godard, A. A. Butt, E. Singer, T. Nguyen, L. Mason, and D. Reynolds, "The 2022 NIST Language Recognition Evaluation," *arXiv preprint arXiv:2302.14624*, 2023.

[14] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?" *arXiv preprint arXiv:2112.10740*, 2021.

[15] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[16] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

[18] K. Krishna, S. Garg, J. P. Bigham, and Z. C. Lipton, "Downstream datasets make surprisingly good pretraining corpora," *arXiv preprint arXiv:2209.14389*, 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.

[21] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[22] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[23] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[24] G. Laperrière, V. Pelloin, M. Rouvier, T. Stafylakis, and Y. Estève, "On the use of semantically-aligned speech representations for spoken language understanding," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 361–368.

[25] 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.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[27] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814–5818.

[28] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from VoxSRC 2020," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5809–5813.

[29] L. Zhang, Q. Wang, and L. Xie, "Duality temporal-channel-frequency attention enhanced speaker representation learning," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 206–213.

[30] Y. Qin, Q. Ren, Q. Mao, and J. Chen, "Multi-branch feature aggregation based on multiple weighting for speaker verification," *Computer Speech & Language*, vol. 77, p. 101426, 2023.

[31] C. Zeng, X. Wang, E. Cooper, X. Miao, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *ICASSP 2022*. IEEE, 2022, pp. 6717–6721.