EXPLORING UNIVERSAL SINGING SPEECH LANGUAGE IDENTIFICATION USING SELF-SUPERVISED LEARNING BASED FRONT-END FEATURES

Xingming Wang[1], *Hao Wu*[2], *Chen Ding*[2], *Chuanzeng Huang*[2], *Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, China
{xingming.wang,ming.li369}@dukekunshan.edu.cn

ABSTRACT

Despite the great performance of language identification (LID), there is a lack of large-scale singing LID databases to support the research of singing language identification (SLID). This paper proposed a over 3200 hours dataset used for singing language identification, called Slingua. As the baseline, we explore two self-supervised learning (SSL) models, WavLM and Wav2vec2, as the feature extractors for both SLID and universal singing speech language identification (ULID), compared with the traditional handcraft feature. Moreover, by training with speech language corpus, we compare the performance difference of the universal singing speech language identification. The final results show that the SSL-based features exhibit more robust generalization, especially for low-resource and open-set scenarios. The database can be downloaded following this repository: *https://github.com/Doctor-Do/Slingua*.

Index Terms— Singing Language Identification (SLID), Universal singing speech Language Identification (ULID), Music Database

1. INTRODUCTION

Knowing the singing language information of a song is beneficial for tasks such as lyrics transcription and music information retrieval. Assuming that the song's metadata, such as lyrics and song title, is available, it may be easily extracted using a text-based classifier. Unfortunately, the song's metadata is generally unavailable in many applications. Thus we need to determine the language information of the songs based on the audio signal.

There have been a few works on singing language identification (SLID) in the past few years. Renault et al. [1] use a phonotactic approach for SLID based on the DALI dataset [2] and achieved good performance. Choi et al. [3] achieve great performance on the Music4all [4] dataset using both the audio signal and the metadata of the song. However, none of the datasets mentioned before were initially designed for the SLID task and thus have some issues, such as uneven distribution of language tags and very limited data scale. In recent years, many corpora have been built in the form of Youtube crawls, e.g., Voxceleb [5], Jtubespeech [6]. Therefore, we propose the SLingua dataset based on Youtube, a dataset that focuses on the SLID task, covering 13 languages with a total of over 3200 hours of songs. This corpus is an aggregation of music playlists created by youtube users, all of which can be downloaded from youtube. In addition, for each audio, we also provide the corresponding results of voice activity detection (VAD). More specifically, this corpus is limited to non-commercial research only.

Recently, large scale self-supervised pre-trained models has been widely used for audio downstream tasks such as language identification (LID)[7], automatic speaker verification (ASV) [8], and emotion recognition [9], etc. Tjandra et al. [7] compared the performance of different transformer layer outputs using Wav2vec2 based model. The outstanding results demonstrate the suitability of the SSL model for LID. However, there is a lack of works on the task of SLID. Therefore, we compare the performance of two self-supervised learning based pre-trained models, Wav2vec2 based XLSR [10] and Hubert [11] based WavLM [12], with the traditional handcraft feature Fbank for SLID tasks as the benchmark of Slingua. Meanwhile, we compare the performance between traditional feature based systems and systems trained on SSL features with different training data scales. Moreover, we analyze the impact of different hidden layers of the self-supervised models regarding the final performance using integrated gradient attribution analysis [13]. To the best of our knowledge, this paper is the first large-scale open-source corpus and benchmark focusing on SLID in recent years.

Different from speech utterances in the LID task, polyphonic songs are characterized by significant overlapping between speech and background music, wide pitch variations, and longer vowel duration [14], making the SLID task more challenging. In addition, by introducing a speech language corpus Voxlingua107 [15], this paper also explores building a universal language recognition model for both speech and singing utterances.

The main contributions of this paper are summarized as follows.

- Releasing a large-scale corpus focusing on the SLID task.
- Comparing the performance of mainstream self-supervised pre-trained models for the SLID task as the benchmark on the Slingua dataset.
- Building a universal language identification system for both speech and singing input utterances.

Corresponding Author: Ming Li. This research is funded in part by the National Natural Science Foundation of China (62171207) , Science and Technology Program of Guangzhou City (202007030011), Science and Technology Program of Suzhou City(SYC2022051) and Bytedance. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

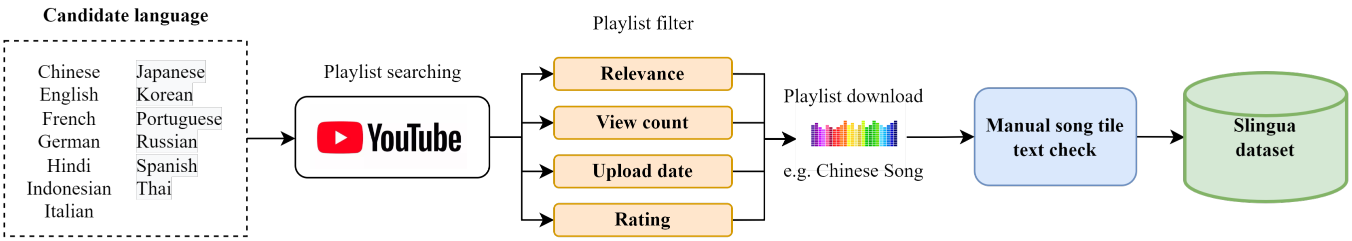

Fig. 1. *The data collection pipeline of the Slingua dataset.*

2. THE SLINGUA DATASET

2.1. Dataset Description

Due to the absence of a large-scale audio corpus for SLID, we collected the **Slingua**[1] dataset based on Youtube audios, which is designed for the SLID task. Since the *VideoIDs* are available, much other valuable information, such as singer information or channel information, can be found on the website. The Slingua dataset provides both the Youtube *VideoID* and the language label for each video. The whole audio dataset can be automatically downloaded and divided using given scripts.

Table 1. *The data distribution of the Slingua training set.*

Language	Num.	Hours	Language	Num.	Hours
Chinese	6659	617	Japanese	3927	305
English	2258	140	Korean	4377	322
French	3314	285	Portuguese	2068	163
German	2692	175	Russian	1796	156
Hindi	2301	187	Spanish	2403	222
Indonesian	3789	284	Thai	1806	162
Italian	2098	191	**Total**	**39488**	**3209**

2.2. Dataset Collection

Fig. 1 summarizes the collection process for the entire Slingua dataset. The construction detail adopt the following steps:

Step.1 **Candidate singing language list.** Taking into account the number and distribution of users, we made the Slingua dataset include a total of 13 languages. The corresponding languages are listed in Table 1.

Step.2 **Audio searching and downloading.** In brief, we first determine the keywords. Then we search playlists on Youtube by given keywords. Finally, all the downloadable audios in the selected playlists form the Slingua dataset. For example, for the target language French, we firstly set the keyword as $french$. By searching $french\ songs$ on Youtube, the top 20 playlists are filtered according to four Youtube official sort methods: relevance, time, number of views, and rating. Eventually, there will be 50-80 playlists corresponding to each language. All audios in playlists are downloaded and resampled to 16000 Hz using *yt-dlp*[2].

Step.3 **Manual detection of text.** After downloading, we deduplicate and manually retrieve the playlists according to the playlists text and video titles. We first tried to use the fastText [16] tool to identify text language. However, considering that many playlists' text and video titles are composed of English, while the actual audio content contains the target language, we made a rough manual correction based on the text. Therefore some mislabeled audios was removed. Note that we have only checked the text, so we cannot guarantee that the labels of audios are 100% accurate. Eventually, over 3200 hours of singing data for 13 languages are collected, making up the Slingua dataset.

2.3. Dataset Post-processing

After data collection, we make all audios go through an internal VAD model to remove the non-vocal part of the audios. The corresponding VAD result for each audio can be found in our repository. One hundred singing clips per language were sliced into 60-second segments and set as the evaluation set. Part of the evaluation set (about 20%) has been manually labeled by listening to the original audios. The remaining singing clips make up the Slingua training set, as shown in Table 1.

3. BENCHMARK SETTING

3.1. Front-end Feature extractor

For SSL models, we utilize and compare two start-of-the-art architectures, Wav2vec-based models and WavLM-based models. Specially, we have used XLS-R model[22] and WavLM-large model[12] as feature extractors. Both models consist of a CNN-based feature encoder and a transformer based context encoder, using raw waveform as input. More detailed information about these two SSL models can be found in Table 2. Both models are trained on cross-lingua corpus and have a similar

[1] https://github.com/Doctor-Do/Slingua

[2] https://github.com/yt-dlp/yt-dlp

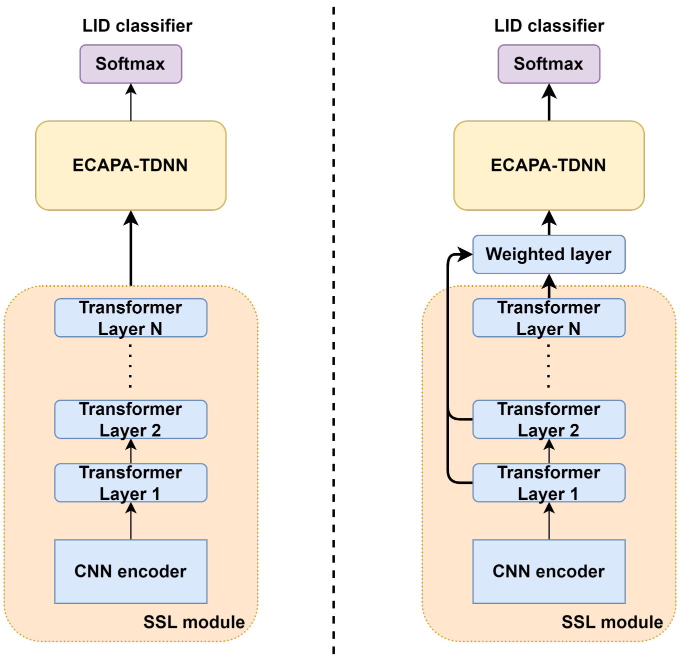

Fig. 2. *The illustration of the SSL-based language identification system. The left denotes using the last transformer output only. The right denotes the weighted feature using all hidden layer outputs.*

Table 2. Details of the adopted self-supervised models.

SSL Model	Training data	Parameters	Output dim.
`W2V-XLSR`	LibriSpeech [17], CommonVoice [18], BABEL	317 M	1024
`WavLM large`	Libri-Light [19], GigaSpeech [20], Voxpopuli [21]	315 M	1024

number of parameters. Thus, we consider them as comparable front-end feature extractors. We compare the feature using the last transformer layer's output only and the weighted feature using all hidden layer outputs for each model. As seen from Fig. 2, the left denotes using the last transformer layer's output only, while the right denotes using the weighted feature.

3.2. Downstream architecture

ECAPA-TDNN[23] has recently achieved great success in speaker verification by introducing the channel attention mechanism. The squeeze-excitation (SE) [24] module is also used in ECAPA-TDNN. The backbone feature extractor is followed by an attentive statistic pooling (ASP) layer[25] in order to extract utterance-level representation. The pooling layer is followed by a linear layer and softmax as a classifier for language classification.

4. EXPERIMENTS SETUP

4.1. Data Usage

4.1.1. Singing dataset

The collected Slingua dataset mentioned in section 2 was used for training and evaluation in our experiments. More details about the distribution of the Slingua training set can be found in Table 1. We also use another internal proprietary evaluation set from Bytedance, called **Saro** in this paper. The Saro evaluation set contains over four thousand labeled songs and can be defined as an out-of-domain (OOD) test set. The Saro dataset contains only seven languages: English, Spanish, Hindi, Korean, Japanese, Indonesian and Portuguese. All these seven languages are included in the Slingua dataset.

4.1.2. Speech dataset

We use voxlingua107 [15], a large corpus used for spoken language recognition, as an auxiliary dataset for universal singing speech language identification. The official voxlingua107 development set is used for evaluation. For utterances in the Voxlingua107, we use the utterances only if those language tag is included in the 13 languages of the Slingua dataset. Therefore, the training and evaluation sets are both subsets of the official Voxlingua107 dataset.

4.2. Model configurations

4.2.1. Front-end Feature Extractor

For SSL models, we followed the official configuration. We compared both fixed front-end and fine-tuning front-end. Fine-tuning denotes training the SSL model with the downstream model. For Fbank, the logarithmical Mel-spectrogram is extracted by applying 80 Mel filters on the spectrogram computed over Hamming windows of 20ms shifted by 10ms.

4.2.2. Downstream Model

For ECAPA-TDNN, the number of feature channels was set as 1024 to scale up the network. The dimension of the bottleneck in the SE-Block is set to 256. The backbone feature extractor is followed by an ASP layer to extract utterance-level representation. A cross-entropy loss function is used for training the model given the one-hot language labels.

4.2.3. Training configurations

During training, all audios were split into 3-second chunks based on VAD results. The learning rate is set as 1e-4 during the training of the SSL-based model while 1e-3 during the training of the Fbank based model. All models are trained using the Adam optimizer.

5. RESULTS AND DISCUSSION

5.1. Singing language identification

For SLID, we compare different configurations of the SSL model and analyze weighted features using integrated gradient attribution for weighted features. In addition, we compare the performance of traditional feature based system with systems trained on SSL features with different training data scales.

Table 3. *Results of different front-end feature extractors for SLID, the downstream models are all ECAPA-TDNN.*

Front-end	Slingua eval		Saro	
	F1	ACC	F1	ACC
80d Mel(Baseline)	0.892	0.893	0.714	0.687
wavlm-last layer-fixed	0.922	0.912	**0.785**	**0.782**
wavlm-last layer-finetuned	0.921	0.914	0.737	0.73
w2v2-last layer-fixed	0.031	0.071	0.05	0.02
w2v2-last layer-finetuned	0.921	0.915	0.731	0.693
wavlm-weighted-fixed	0.913	0.908	0.716	0.709
wavlm-weighted-finetuned	**0.936**	**0.932**	0.732	0.725
w2v2-weighted-fixed	0.915	0.910	0.681	0.701
w2v2-weighted-finetuned	0.915	0.908	0.696	0.688

5.1.1. *SSL model selection*

As shown in Table 3, systems with different SSL models show significant differences on the performance of the SLID task. Even for the same model, adopting the weighted scheme or using the output of the last transformer layer as features also achieve different results. Overall, the SSL-based model outperforms the traditional handcraft feature based model except for the w2v2-last layer-fixed based model. The WavLM-last layer-fixed achieves outstanding performance, especially on the Saro evaluation set. This demonstrates the generalizability of the SSL front-end for OOD data compared with the traditional handcraft feature. The final results on the Saro evaluation set also show that compared to the Wav2vec2-based model, the WavLM-based model achieved better performance generally. this finding is similar to [12] that WavLM based pre-trained features are more robust in downstream audio classification tasks.

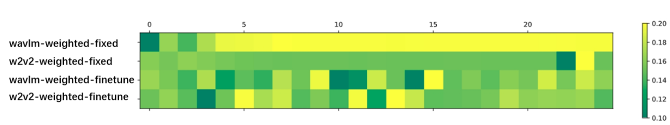

Fig. 3. *The integrated gradient attributions for four weighted models. The y-axis represents different models, while the x-axis indicates different transformer layers.*

Inspired by [26], we use the Integrated Gradient (IG) attribution analysis approach to study the weight values. The IG considers the gradient distribution, thus indicating a better model contribution. Fig. 3 presents the IG attributions for four weighted models. Before model fine-tuning, the weight of the last layer of w2v2-weighted-fixed occupies a smaller percentage, indicating why w2v2-last layer-fixed has poor performance in the SLID task. After fine-tuning, we can observe that the high weights are distributed across various layers on both models. This weights distribution presents that shallow and deep features contain language-related information, not just the last layer.

5.1.2. *Training data scaling*

Table 4 shows the performance of different training data scales. The results demonstrate the superiority of SSL front-end features on limited training data compared to traditional handcraft features. For 5 hours of training data, the WavLM-based feature improves performance by nearly 20% over the traditional Mel spectrum in the unseen Saro scenario. Although there is no significant improvement on the Slingua test set, there is still a remarkable discrepancy between the model trained with the full amount of data and those trained with 50 hours of data on the Saro evaluation set.

Table 4. *Results of different training data scale for SLID, the downstream models are all ECAPA-TDNN.*

Data scale	Front-end	Slingua eval		Saro	
		F1	ACC	F1	ACC
5 hours	80d Mel	0.651	0.655	0.474	0.513
5 hours	wavlm-last layer-fixed	0.786	0.79	0.588	0.616
50 hours	80d Mel	0.884	0.883	0.669	0.676
50 hours	wavlm-last layer-fixed	0.916	0.913	0.734	0.730
3209 hours (ALL)	80d Mel	0.892	0.893	0.714	0.687
3209 hours (ALL)	wavlm-last layer-fixed	0.922	0.912	0.785	0.782

5.2. Universal singing speech language identification

Table 5. *Results of different front-end feature extractors for ULID, the Downstream models are all ECAPA-TDNN. The Voxlingua107 training and evaluation sets used here are both subsets as mentioned in 4.1.2*

Front-end	Training data	Slingua eval		Saro		Voxlingua107	
		F1	ACC	F1	ACC	F1	ACC
wavlm-last layer-fixed	Slingua	0.922	0.912	0.785	0.782	-	-
80d Mel	Slingua	0.892	0.893	0.714	0.687	-	-
wavlm-last layer-fixed	Slingua+Voxlingua107	0.908	0.894	**0.812**	**0.858**	0.982	0.972
80d Mel	Slingua+Voxlingua107	0.904	0.901	0.742	0.784	0.934	0.945

As can be seen in Table 5, it is possible to build a ULID system. By simply using speech and singing data for training, the Saro set's performance has dramatically improved relevant 10%. Both models perform well on the voxlingua107 development subset, probably due to the fact that the speech evaluation subset is relatively easy.

6. CONCLUSION

This paper proposes a large-scale corpus for singing language identification, which contains over 3200 hours of singing data, named Slingua. Moreover, we explore the performance of two different self-supervised learning based front-end feature extractors for SLID. The WavLM-large model performs best in our experiments. The results demonstrate that the SSL-based feature performs better on limited low-resource training data than the traditional handcraft feature. Moreover, for open-set scenarios, the SSL-based feature exhibits more robust generalization. In addition, we build an effective universal singing speech language identification system by combining singing and speech data during the training phase.

7. REFERENCES

[1] Lenny Renault, Andrea Vaglio, and Romain Hennequin, "Singing language identification using a deep phonotactic approach," in *Proc. ICASSP*. IEEE, 2021, pp. 271–275.

[2] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters, "Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm," in *Proc. ISMIR*, 2019.

[3] Keunwoo Choi and Yuxuan Wang, "Listen, read, and identify: multimodal singing language identification of music," *arXiv preprint arXiv:2103.01893*, 2021.

[4] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al., "Music4all: A new music database and its applications," in *Proc. IWSSIP*. IEEE, 2020, pp. 399–404.

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

[6] Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe, "Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[7] Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli, "Improved language identification through cross-lingual self-supervised learning," in *Proc. ICASSP*. IEEE, 2022, pp. 6877–6881.

[8] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.

[9] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *Proc. SLT 2021*. IEEE, pp. 373–380.

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

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

[12] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017, pp. 3319–3328.

[14] Chen Zhang, Jiaxing Yu, LuChin Chang, Xu Tan, Jiawei Chen, Tao Qin, and Kejun Zhang, "Pdaugment: Data augmentation by pitch and duration adjustments for automatic lyrics transcription," *arXiv preprint arXiv:2109.07940*, 2021.

[15] Jörgen Valk and Tanel Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *Proc. SLT 2021*. IEEE, pp. 652–658.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[18] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[19] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. ICASSP*. IEEE, 2020, pp. 7669–7673.

[20] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech 2021*, pp. 3670–3674.

[21] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 993–1003.

[22] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, pp. 2278–2282.

[23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, pp. 3830–3834.

[24] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[25] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[26] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Zhuo Chen, Peidong Wang, Gang Liu, Jinyu Li, Jian Wu, Xiangzhan Yu, and Furu Wei, "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?," in *Proc. Interspeech 2022*, pp. 3699–3703.