INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Label Aware Speech Representation Learning For Language Identification

Shikhar Vashishth, Shikhar Bharadwaj, Sriram Ganapathy, Ankur Bapna,
Min Ma, Wei Han, Vera Axelrod, Partha Talukdar

Google Research

{shikharv,shikharop,srigana,ankurbpn,minm,weihan,vaxelrod,partha}@google.com

Abstract

Speech representation learning approaches for non-semantic tasks such as language recognition have either explored supervised embedding extraction methods using a classifier model or self-supervised representation learning approaches using raw data. In this paper, we propose a novel framework of combining self-supervised representation learning with the language label information for the pre-training task. This framework, termed as **L**abel **A**ware **S**peech **R**epresentation (LASR) learning, uses a triplet based objective function to incorporate language labels along with the self-supervised loss function. The speech representations are further fine-tuned for the downstream task. The language recognition experiments are performed on two public datasets – FLEURS and Dhwani. In these experiments, we illustrate that the proposed LASR framework improves over the state-of-the-art systems on language identification. We also report an analysis of the robustness of LASR approach to noisy/missing labels as well as its application to multi-lingual speech recognition tasks.

Index Terms: speech representation learning, supervision and self-supervision, language identification.

1. Introduction

The conventional approach for deriving speech representations for non-semantic speech tasks, such as speaker and language recognition, involved the use of training deep neural models with a statistics pooling layer. Some of the popular methods in this direction include d-vectors [1] and x-vectors [2], where a deep neural model is trained to classify the speaker/language labels on a large corpus of supervised data. However, recent trends in speech processing has seen a paradigm shift towards self-supervision based representation learning, mirroring the efforts in computer vision [3] and natural language processing [4]. Some popular examples of such approaches include contrastive predictive coding (CPC) [5], wav2vec family of models [6, 7], and hidden unit BERT (HuBERT) [8]. These methods primarily rely on learning speech representations at the frame-level with its impact reported on semantic tasks such as low-resource speech recognition [8, 9] or zero resource spoken language modeling [10]. These representations have also been investigated for speaker and language recognition tasks [11] through various benchmarks such as SUPERB [12] and NOSS [13].

In many learning paradigms, it is plausible to have portions of pre-training data along with the corresponding meta-data. In the broad spectrum of representation learning, where supervised and self-supervised frameworks constitute the two-ends of the spectrum, we hypothesize that a combination of supervision and self-supervision based methods may be more optimal than either of the two frameworks in isolation, for scenarios where parts of the pre-training have additional meta-data in the form of labels. In this paper, we propose a framework for **L**abel **A**ware **S**peech **R**epresentation learning (LASR) for such scenarios. To the best of our knowledge, this is the first attempt to combine label information with a self-supervision loss for non-semantic speech tasks. The contributions from this work are as follows.

1. We propose LASR, a framework for incorporating label information in self-supervised speech representation learning.
2. We demonstrate the effectiveness of LASR for language identification task and establish its efficacy even with missing and noisy labels.
3. Our findings demonstrate that inclusion of language information in the pre-training phase results in state-of-art-results on the FLEURS dataset [14].

2. Related Work

Supervised Learning: Deep learning methods for non-semantic speech tasks initially explored speech recognition models in the unsupervised i-vector framework [15]. Further, the embeddings derived from a classifier model, trained on large amounts of supervised pre-training data, showed promising results for speaker [1] and language recognition [16]. The initial architecture based on time-delay neural network (TDNN) [2] has since been improved with factorization [17], residual networks [18] and more recently with channel attention based TDNN [19]. Most of these approaches use a pooling layer to convert frame level representations to an utterance level embedding followed by a cross-entropy based classification objective. However, our work investigates the combination of self-supervision objectives along with the supervised labels.

Speech Self-Supervised Learning: Prior research in the field of speech self-supervised learning can largely be classified into two major categories: contrastive and predictive. The contrastive approaches learn by maximizing the similarity of an anchor with the positive samples, while simultaneously minimizing its similarity with the negative samples. The class of wav2vec models [6, 20] fall in this category. On the other hand, predictive methods are based on masked language modeling (MLM) objective [4]. The examples include Discrete-BERT [21], w2v-BERT [7], HuBERT [8], and BEST-RQ [22]. Our proposed framework enables integration of label information in both categories of methods.

Non-Semantic Speech Representations: For tasks such as language identification, speaker diarization, and emotion detection, it is essential to also capture the non-semantic aspect of speech. TRILL [13] utilizes temporal proximity as

10.21437/Interspeech.2023-854

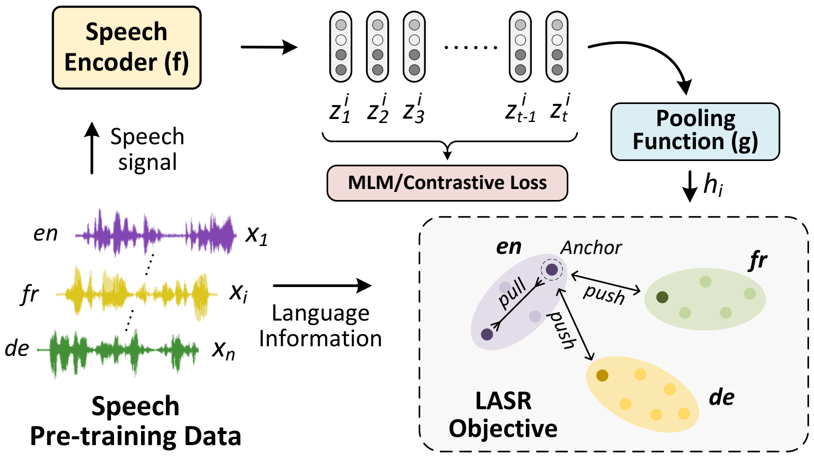

Figure 1: *Overview of LASR framework. Given a batch of multilingual speech samples, for each sample x_i, LASR utilizes a speech encoder (f) to obtain frame-level representations $z_1^i, z_2^i, ..., z_t^i$. These are used for computing self-supervised loss and are fed to a pooling function (g) to derive utterance-level embedding h_i. The language labels of samples and h_i's are used to compute LASR loss.*

supervision signal to learn non-semantic representation, with promising results on NOSS (non-semantic speech) benchmark. Further, methods such as FRILL [23] and TRILLsson [24] have enhanced the performance and efficiency of these models. Another approach named COLA [25] modifies the negative sampling scheme to learn more general purpose audio representation. All these works are specifically designed for contrastive techniques, whereas LASR can be integrated with any self-supervised speech representation learning method.

Joint learning: Talnikar et. al. [26] explored the combination of supervised (connectionist temporal cost (CTC)) and self-supervised (contrastive prediction loss (CPC)) losses for speech recognition. Similarly, UniSpeech [27] used CTC labeling and phonetically-aware contrastive learning in a multi-task learning framework. Bai et. al. [28] used the self-supervised MLM loss and the speech recognition loss for a multi-lingual speech recognition system. However, all these approaches learn frame-level representations for a semantic task. In our work, the LASR framework combines utterance-level label supervision with frame-level self-supervision.

3. LASR Framework

A comprehensive illustration of the LASR framework is depicted in Figure 1. A self-supervised speech encoding model $f : \boldsymbol{x} \rightarrow \boldsymbol{\mathcal{Z}}$, such as wave2vec-2.0 [6] or w2v-BERT [7], transforms a raw audio waveform $\boldsymbol{\mathcal{X}}$ into the frame-level speech representations $\boldsymbol{\mathcal{Z}} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_T]$. In our proposed LASR framework, the pre-training dataset is denoted as $\mathcal{D} = \{(\boldsymbol{\mathcal{X}}_1, l_1), (\boldsymbol{\mathcal{X}}_2, l_2), ..., (\boldsymbol{\mathcal{X}}_n, l_n)\}$, where each speech utterance $\boldsymbol{\mathcal{X}}_i$ is accompanied by its corresponding language label l_i. The remaining unlabeled samples will solely be utilized for optimizing the self-supervised objective.

Subsequently, we employ an aggregation function $g : \boldsymbol{\mathcal{Z}} \rightarrow \boldsymbol{h}$ to obtain an utterance level embedding $\boldsymbol{h} = g(\boldsymbol{\mathcal{Z}})$. Here, g can, in general, take the form of a neural network such as LSTM or an attention model [29]. In our case, g is an average pooling, i.e., $\boldsymbol{h} = g(\boldsymbol{\mathcal{Z}}) = \frac{1}{T}\sum_t \boldsymbol{z}_t$.

For an anchor speech utterance $\boldsymbol{\mathcal{X}}_i$ with aggregate representation $\boldsymbol{h}_i$ and language label l_i, we select a positive and negative sample: $\boldsymbol{\mathcal{X}}_i^+$ and $\boldsymbol{\mathcal{X}}_i^-$ such that $l_i^+ = l_i$ and $l_i^- \neq l_i$. We use the triplet-loss objective, as proposed in [30], i.e.,

$$\mathcal{L}_{\texttt{tri}} = \sum_i \max\left[0, \gamma + d(\boldsymbol{h}_i, \boldsymbol{h}_i^+) - d(\boldsymbol{h}_i, \boldsymbol{h}_i^-)\right], \quad (1)$$

where γ is the margin and $d(\cdot, \cdot)$ is the distance metric employed. In this work, we use angular distance as the distance metric. We also explore the hard triplet mining strategy [31, 32], where the most distant positive and closest negative sample within the mini-batch are selected to form the triplet.

$$\mathcal{L}_{\texttt{hard}} = \sum_i \max[0, \gamma + \max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j) - \min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)] \quad (2)$$

Here, $j \in i^+$ denotes the set of utterances in the mini-batch that have the same label $l_j = l_i$ and $j \in i^-$ denotes the set of utterances with a different label, i.e., $l_j \neq l_i$. The total loss function used in the proposed LASR approach is given by,

$$\mathcal{L}_{\texttt{LASR}} = \mathcal{L}_{\texttt{SSL}} + \lambda \cdot \mathcal{L}_{\texttt{hard}}. \quad (3)$$

Here, $\mathcal{L}_{\texttt{SSL}}$ is the loss corresponding to the self-supervised speech encoding method f and λ decides the trade-off between the SSL objective and hard-triplet objective. In our experiments, we find that having the SSL objective is crucial for achieving the best language recognition performance. In Section 6, we also assess the significance of altering the parameter λ. In addition to the triplet loss (Eq. 3), we also examine generalized end-to-end loss [33].

$$\mathcal{L}_{\texttt{ge2e}} = \sum_i 1 - \sigma(\max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j)) + \sigma(\min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)). \quad (4)$$

4. Experimental Setup

4.1. Dataset

Pre-training Data: In our experiments, we employ a large set of open source speech data for pre-training, totaling about 429k audio hours. This consists of 372k hours of speech data across 23 languages from VoxPopuli dataset [34], 50k hours of speech from 25 languages in Common Voice dataset [35], 50k hours of read speech in 8 European languages from Multilingual LibriSpeech (MLS) corpus [36], and 1000 hours of telephone conversation data across 17 African and Asian languages from BABEL dataset [37]. Overall, this combined dataset has speech utterances from 75 languages.

Method	FLEURS					Dhwani				
	O (48)	NO (54)	Overall			O (5)	NO (17)	Overall		
	Acc	Acc	Acc	F1	EER	Acc	Acc	Acc	F1	EER
wav2vec 2.0 [6]	84.8	72.2	78.2	76.3	1.1	77.6	47.6	56.0	41.1	15.9
w2v-BERT [7]	87.7	69.6	78.0	77.7	0.5	78.8	49.9	58.0	42.6	15.4
BEST-RQ [22]	86.8	65.6	75.8	73.3	1.2	76.2	46.4	54.7	39.8	16.9
LASR + wav2vec 2.0	89.6	74.8	**81.9**	79.9	0.7	78.5	50.1	58.1	42.5	**15.2**
LASR + w2v-BERT	88.9	**74.3**	81.3	**80.4**	**0.5**	**78.9**	50.6	**58.6**	42.8	15.9
LASR + BEST-RQ	**90.6**	73.4	81.6	79.7	**0.5**	77.0	**50.8**	58.2	**43.2**	16.1

Table 1: *Language identification accuracy (%), macro-F1 and equal error rate (EER) for various approaches. O stands for languages that overlap with the pre-training data and NO are the non-overlapping languages. In parenthesis, we report the number of classes in each category. We find that methods trained with the LASR objective achieve better performance. Refer to Section 5 for details.*

Evaluation Data: In our experiments, we employ FLEURS [14] and Dhwani [38] datasets for spoken language identification. Additionally, we utilize Multilingual Librispeech dataset [36] for Automatic Speech Recognition (ASR). The FLEURS dataset consists of speech data for 102 languages, with approximately 12 hours of speech per language, derived from translated versions of 2009 English Wikipedia sentences. All the translations are human generated with training, development and test containing 1500, 150 and 359 sentences respectively. Each sentence was spoken by at least 3 native speakers of the language.

The Dhwani dataset encompasses multilingual speech data from 40 Indian languages, downloaded from YouTube and the news platform `newsonair`. For our experiments, we use only the publicly accessible YouTube split, which consists of 12.6k hours of speech in 22 Indian languages. Unlike the FLEURS dataset, the Dhwani dataset is highly noisy and also contains substantial amounts of code-mixing, which challenges the label information based learning in the proposed LASR method.

4.2. Baseline systems

We compare LASR framework with several other established benchmarks namely, (i) **wav2vec-2.0 (w2v)** model [6] pre-trained with SSL contrastive loss, (ii) **w2v-BERT** [7] model, trained using the SSL MLM loss, and (iii) **BEST-RQ** [22] model, which uses a random quantizer with the MLM loss. Since the LASR approach is agnostic to the choice of the SSL objective function, we explore the combination of wav2vec-2.0, w2v-BERT and BEST-RQ model with the hard-triplet based LASR objective. All models are fine-tuned on the respective training split of the downstream task before evaluation.

Implementation details: Most of the hyper-parameters are directly adopted from prior works [7, 22]. All the SSL baseline systems are pre-trained for 1.5M epochs. For LASR training, the pre-trained SSL model at 1M epochs is used as initialization, followed by 0.5M steps of training with the LASR objective. All the models are fine-tuned on the supervised training data for an additional 50k epochs, with a batch size of 64. We choose λ from $\{4, 8, 16\}$. The Adam optimizer [39] is used in conjunction with a Transformer learning rate scheduler [4] that has 40k warm-up steps. The learning rate is increased to $6e^{-4}$, followed by an inverse square root decay. We report mean of three runs for all the results.

Initialization	FLEURS			Dhwani		
	Acc	F1	EER	Acc	F1	EER
Random	52.8	46.5	2.1	55.3	25.1	14.9
BEST-RQ	81.4	73.4	0.9	61.9	30.0	17.9
+ LASR	83.8	73.4	0.9	62.2	34.3	18.0

Table 2: *Language recognition performance in supervised case. Fully supervised models have higher macro-F1 and EER.*

5. Results

The language recognition performance is measured using accuracy, equal error rate (EER) and macro-F1 score. These results are reported in Table 1. The languages in the test set (FLEURS/Dhwani) are split into two categories - a) the set of languages which overlap with the ones in the pre-training (denoted as O, 48 classes in the FLEURS dataset and 5 classes in the Dhwani dataset), and b) the set of languages which do not have any overlap with the set of languages in the pre-training data (denoted as NO, 54 classes in the FLEURS dataset and 17 classes in the Dhwani dataset). Further, the overall results are also reported. The following are the key takeaways from the results reported in Table 1.

- On the FLEURS dataset, the LASR approach improves the BEST-RQ model relatively by 7.7%, 8.7%, and 58.3% in terms of accuracy, F1 and EER metrics, respectively. Similarly, on Dhwani dataset, the relative improvements from LASR for BEST-RQ are 6.4%, 8.5%, and 5.0% on the above metrics. This trend is consistent across other pre-training methods as well. Thus, LASR framework improves over the baseline SSL results for both the datasets and for all the pre-training models (wav2vec-2.0, w2v-BERT and BEST-RQ).
- The improvements observed for the LASR approach are also consistent with the overlap and the non-overlap subsets of the test data, and on all the three metrics reported.
- For all the systems compared, the performance on the overlap set is consistently better than the non-overlap set. This indicates that, even when the pre-training objective did not explicitly use language labels (baseline SSL approaches), the language information is implicitly captured by the models.

Fully-Supervised Setting: In Table 2, we report the performance for the scenario where a supervised pre-training is performed using the combined data of all the languages (pre-training and training data) with the cross-entropy loss. The label set is the union of the languages in the pre-training data and the fine-tuning data. The supervised model architecture is

Method	Accuracy
MLM	75.8
Hard-Triplet	74.3
MLM + Triplet (Eq. 1)	75.1
MLM + GE2E (Contrastive) (Eq. 4)	76.2
MLM + Hard-Triplet (Eq. 2)	80.4

Table 3: *Language identification accuracy with various loss functions. Hard-triplet loss performs best among all.*

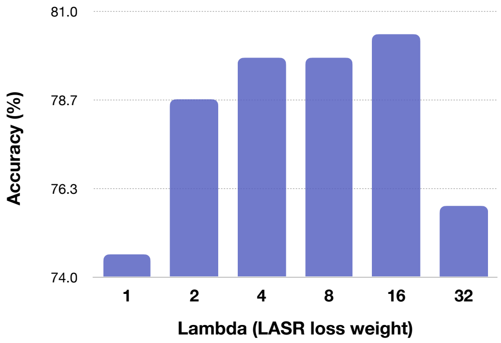

Figure 2: *Accuracy for different choice of λ in LASR objective (Eq. 3). Slightly higher value of $\lambda(=16)$ is beneficial.*

identical to the SSL and LASR models reported in Table 1. We also experiment with three different initialization choices for this supervised model - i) random initialization, ii) BEST-RQ model trained with SSL, and iii) BEST-RQ model trained with LASR objective. Our findings show that, on both the datasets, the fully-supervised setting does not achieve satisfactory results without weight initialization using a pre-trained model. While the accuracy of the supervised model improves over the SSL and LASR models in Table 1, EER and F1 scores are substantially worse for the supervised models. Nevertheless, the performance in this setting also improves with LASR initialization.

6. Discussion

Effect of different optimization objectives - Table 3 compares various supervised loss functions. These experiments used the BEST-RQ [22] model evaluated on the FLEURS dataset. The first two experiments of Table 3 use only the MLM loss (SSL loss) or only the supervised loss (Hard-triplet loss). The remaining experiments use the combined LASR loss (Eq. 3). As seen here, the hard-triplet loss improves over other choices of semi-hard triplet loss or GE2E loss.

Choice of supervised loss weight $\boldsymbol{\lambda}$ - For the hard-triplet loss in the LASR objective function (Eq. 3), we have experimented with different choices of λ. These results are reported in Fig. 2. The optimal choice of λ is found to be 16, which indicates that a higher weight for the supervised component is beneficial for the language recognition performance. However, a larger value, for example, $\lambda = 32$, degrades the performance.

Pre-training with missing/noisy labels - All experiments reported thus far used the language label information for the entire pre-training data. We experiment with the robustness of the LASR approach for cases where the label information is either missing or noisy. For these experiments reported in Fig. 3, we assume $p\%$ of the pre-training data to either have missing labels or have noisy labels (randomly corrupted to other language labels in pre-training set). As expected, the language recognition performance degrades as p increases. However, even when 75% of the pre-training data labels are missing, LASR is significantly better than the baseline approach. The experiments highlight that the LASR approach can also yield performance improvements on pre-training data with noisy/missing labels.

Method	Languages						Avg
	de	en	es	fr	it	nl	
w2v-BERT	4.0	6.2	4.0	4.7	8.9	10.6	7.2
+ LASR	4.0	6.2	4.8	4.8	8.9	10.0	7.2
BEST-RQ	3.9	6.2	3.8	4.8	8.8	9.3	7.0
+ LASR	4.1	6.2	4.3	4.8	9.0	9.6	7.1

Table 4: *WER (%) for ASR on Multilingual LibriSpeech. Adding the non-semantic LASR objective during pre-training does not degrade performance on semantic tasks such as ASR.*

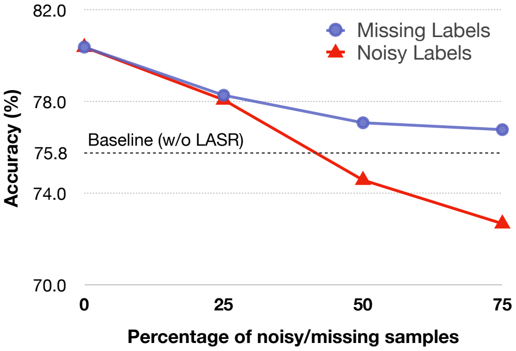

Figure 3: *Pre-training with missing or noisy labels. LASR is robust to missing and noisy language information in the data.*

Impact on downstream ASR tasks - In this section, we fine-tune LASR models on a semantic task, namely ASR. In particular, we run experiments for multilingual ASR on the MLS dataset. We follow a similar setup for ASR fine-tuning as was done in [22]. To be more specific, we use the RNN-transducer model [40], where the decoder uses unidirectional LSTM. We do not employ shallow fusion with an external language model. The ASR WER (%) results are reported in Table 4. As seen in this table, the LASR based objective does not degrade the overall ASR performance even when the label information used in the LASR loss is an utterance-level non-semantic label. Thus, the representations learned using the LASR approach improve the language recognition tasks without any degradation on semantic tasks such as ASR.

7. Conclusion

In this paper, we introduce a method for enhancing self-supervised speech representation learning by incorporating non-semantic language label information. Our proposed approach, Label Aware Speech Representation (LASR) learning, utilizes a triplet-based objective in addition to the self-supervised loss function. The results from language recognition experiments demonstrate that the LASR approach provides substantial overall improvements, particularly on subsets of test data that do not overlap with pre-training languages. Additionally, experiments on the automatic speech recognition (ASR) task indicate that the LASR model produces speech representations that do not compromise performance for semantic tasks.

8. References

[1] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE ICASSP*. IEEE, 2016, pp. 5115–5119.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5329–5333.

[3] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of 2019 NAACL-HLT*, Jun. 2019.

[5] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th NeurIPS Conference*, 2020.

[7] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *2021 IEEE ASRU Workshop*, pp. 244–250, 2021.

[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," 2021.

[9] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[10] T. Maekaku, X. Chang, Y. Fujita, and S. Watanabe, "An exploration of hubert with large number of cluster units and model assessment using bayesian information criterion," in *ICASSP 2022-2022*. IEEE, 2022, pp. 7107–7111.

[11] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[12] 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 *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[13] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[14] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *2022 IEEE Spoken Language Technology Workshop*, pp. 798–805, 2022.

[15] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 ICASSP*. IEEE, 2014, pp. 1695–1699.

[16] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, vol. 2018, 2018, pp. 105–111.

[17] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*.

[18] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

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

[21] A. Baevski, M. Auli, and A. rahman Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *ArXiv*, vol. abs/1911.03912, 2019.

[22] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *Proceedings of the 39th ICML*, 2022, pp. 3915–3924.

[23] J. Peplinski, J. Shor, S. P. Joglekar, J. Garrison, and S. N. Patel, "Frill: A non-semantic speech embedding for mobile devices," in *Interspeech*, 2020.

[24] J. Shor and S. Venugopalan, "TRILLsson: Distilled Universal Paralinguistic Speech Representations," in *Proc. Interspeech 2022*, 2022, pp. 356–360.

[25] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021 - 2021 IEEE*, 2021, pp. 3875–3879.

[26] C. Talnikar, T. Likhomanenko, R. Collobert, and G. Synnaeve, "Joint masked cpc and ctc training for asr," in *ICASSP 2021-2021*. IEEE, 2021, pp. 3045–3049.

[27] 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 *ICML*. PMLR, 2021.

[28] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. C. Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual asr," in *ICASSP 2022-2022*. IEEE, 2022.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd ICLR 2015, Conference Track Proceedings*, 2015.

[30] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE CVPR 2015*, 2015, pp. 815–823.

[31] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In defense of the classification loss for person re-identification," in *Proceedings of the IEEE/CVF CVPR Workshops*, June 2019.

[32] V. Mingote, D. Castan, M. McLaren, M. K. Nandwana, A. Ortega, E. Lleida, and A. Miguel, "Language Recognition Using Triplet Neural Networks," in *Proc. Interspeech 2019*, 2019, pp. 4025–4029.

[33] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE ICASSP 2018*. IEEE, 2018, pp. 4879–4883.

[34] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. M. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.

[35] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2019.

[36] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. Interspeech 2020*, 2020, pp. 2757–2761.

[37] M. J. F. Gales, K. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Workshop on Spoken Language Technologies for Under-resourced Languages*, 2014.

[38] T. Javed, S. Doddapaneni, A. Raman, K. S. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Towards building asr systems for the next billion users," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022 (to appear).

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[40] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012.