INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Acoustic Word Embeddings for Untranscribed Target Languages with Continued Pretraining and Learned Pooling

Ramon Sanabria, Ondrej Klejch, Hao Tang, Sharon Goldwater

The University of Edinburgh
r.sanabria@ed.ac.uk

Abstract

Acoustic word embeddings are typically created by training a pooling function using pairs of word-like units. For unsupervised systems, these are mined using k-nearest neighbor (KNN) search, which is slow. Recently, mean-pooled representations from a pre-trained self-supervised English model were suggested as a promising alternative, but their performance on target languages was not fully competitive. Here, we explore improvements to both approaches: we use continued pre-training to adapt the self-supervised model to the target language, and we use a multilingual phone recognizer (MPR) to mine phone n-gram pairs for training the pooling function. Evaluating on four languages, we show that both methods outperform a recent approach on word discrimination. Moreover, the MPR method is orders of magnitude faster than KNN, and is highly data efficient. We also show a small improvement from performing learned pooling on top of the continued pre-trained representations.

Index Terms: acoustic word embeddings, semi-supervised learning, continued pre-training, low-resource languages, unwritten languages

1. Introduction

Acoustic Word Embeddings (AWE) are vector representations of variable length speech segments (*i.e.*, words) [1, 2]. Ideally, AWEs abstract away from non-linguistic information such as speaker gender and voice quality, so that instances of the same word cluster together in the embedding space. AWEs can be applied to a wide variety of search-intensive tasks such as query-by-example [3] or semantic speech retrieval [4], as well as in unsupervised word segmentation and clustering systems [5]—one step toward creating speech technology for unwritten languages. Since our eventual goal targets this task, this paper focuses on models that do not rely on transcribed speech in the target language—the unsupervised setting—and we also assume limited *unlabeled* target language data (up to 50 hours).

Most previous work on constructing unsupervised AWEs has approached the problem using *learned pooling*, where positive training pairs of similar speech segments (assumed to be the same word or n-gram) are used to learn a pooling function, based on a reconstruction [6, 7, 8] or contrastive [9, 10] objective. Despite good AWE quality, these methods rely on identifying positive training pairs from a corpus using k-nearest-neighbors methods [10, 11]. Even with approximate search, such methods are computationally and memory intensive.

Recently, an alternative method for constructing AWEs was proposed [12], which does not rely on positive samples nor complex pooling mechanisms, but simply mean-pooling the frame-level representations from a pre-trained self-supervised model. The authors show that HuBERT [13] representations (pre-trained on English) work well for English AWEs, but less well on other languages, presumably because the model is not adapted to those target languages.

In this work, we propose solutions to the aforementioned limitations and perform experiments to directly compare the two methods individually and in combination. Specifically, for learned pooling, we show that high-quality positive pairs can be found efficiently by transcribing the target language data using a multilingual phone recognizer (MPR) trained on high-resource languages, then selecting matching phone n-grams. For average pooling of a pre-trained model, we show how to adapt English HuBERT to the target language using continued pretraining [14, 15, 16]. This entails an extra step of reconstructing HuBERT's k-means clusters, not needed for continued pretraining of wav2vec 2.0 [17], but is worth doing since HuBERT has been shown to be more effective at word discrimination [12].

We evaluate our AWE representations using the same-different word discrimination task (*same-diff*) [18] on four languages: French (which we use as a development language to select hyper-parameters and run analyses), Mandarin, German, and Xitsonga (with the latter three acting as unseen test languages from a variety of language families). We experiment with continued pretraining, learned pooling using a contrastive objective, and their combination. Our experiments show that:

- Using continued pretraining with 50 hours of target language data improves the performance of average-pooled HuBERT representations considerably, and most of the benefit is achieved with only 20 hours of data;
- For the contrastive-learning model, using MPR to identify positive pairs yields a large number of high-quality pairs, resulting in better word discrimination scores than a previous approach [10] while being orders of magnitude faster;
- With 50h of data, continued pretraining and contrastive learning have similar performance, but contrastive learning is more data-efficient, and achieves nearly the same results with only one hour of target language data;
- Combining both methods yields a small further improvement.

2. Task Overview

In this section, we formally define our task and experimental framework. Suppose we have two utterances x^1 and x^2, both being sequences of frames. The task (*same-diff*) is to tell whether two word segments $x^1_{s:t}$ and $x^2_{s':t'}$ belong to the same word type or not.

In this paper, we focus on an approach that uses self-supervised speech representations. We assume we have a self-supervised model f to compute representations of both utter-

10.21437/Interspeech.2023-268

ances, i.e., $z^1 = f(x^1)$ and $z^2 = f(x^2)$. We then use a pooling function g to compute a fixed-dimensional vector, also known as an AWE, to represent a word segment. Specifically, we compute $g(z^1_{s:t})$ to represent $x^1_{s:t}$ and $g(z^2_{s':t'})$ for $x^2_{s':t'}$. The question of whether two word segments $x^1_{s:t}$ and $x^2_{s':t'}$ are of the same word type or not becomes measuring the similarity between $g(z^1_{s:t})$ and $g(z^2_{s':t'})$ compared to other AWEs. The cosine similarity is typically used, and the mean average precision (MAP), i.e., the area under the ROC curve when we sweep a threshold, is used as the evaluation metric.

Prior work [12] has shown strong results when f is an off-the-shelf HuBERT model trained on English and g is a simple averaging. In this work, we explore the setting where we have untranscribed speech of a target language to continue pretraining the self-supervised model f. In addition, we also explore pooling functions with trainable parameters, such as in [6, 7, 10]. We follow [9, 10, 19] and train the pooling function g with a contrastive loss. Specifically, we use NTXent [20] which is defined as

$$\ell_{\text{NTXent}}(c, c^+, N) = -\log \frac{\exp(\cos(c, c^+)/\tau)}{\sum_{c^- \in N} \exp(\cos(c, c^-)/\tau)}, \quad (1)$$

where c and c^+ are both AWEs, c^+ is a positive example for c, N is a set of negative examples for c, and τ is a temperature hyperparameter. This loss requires mining positive and negative pairs. For example, in [6, 7, 10, 11], nearest neighbors are considered as positive examples, leaving others as negative examples. Nearest neighbors are known to be slow to compute when the number of examples becomes large. As we will see in later sections, we will explore a different approach to mining positive and negative pairs.

3. Continued Pre-training

In previous work, Sanabria et al. [12] showed that for constructing AWEs on different languages, HuBERT representations (which are trained only on English) perform better than both English-trained wav2vec 2.0 [17] and multilingually-trained XLS-R [21]. We extend their work to the setting where some untranscribed target language data is available. When there is a mismatch between training and test conditions, a common approach is to continue pretraining self-supervised models on the test condition [14, 15, 16], which motivates us to continue pretraining HuBERT on the untranscribed target language.

The task of HuBERT pretraining is masked prediction, where parts of the input are masked, and the goal is to predict the quantized speech frames of the masked parts. To perform continued pretraining on a different language, it is not immediately obvious what training targets to use. Similar to the original HuBERT, we run k-means on the hidden vectors from one of the HuBERT layers, and use the cluster IDs of hidden vectors as targets (described further in Section 3.1). We then continue to pretrain HuBERT on the target language, and finally (following [12]), mean-pool the hidden vectors to create the AWE.

3.1. Experimental Setting

We evaluate our approach with the *same-diff* word discrimination task [18] on French, German, Mandarin, and Xistonga. The French, German, and Mandarin sets are from Task 2 of Zerospeech 2017 [22], and Xitsonga is from NCHLT [23]. Following [24], we only use words that are at least 5 characters (or 2 characters, for Mandarin) and 0.5 seconds long, and we report mean average precision (MAP) scores.[1] Table 1 summarizes the statistics of each test set.[2]

Table 1: *Numbers of lexical entries (word types) and word occurrences (instances) in the test set for each language.*

Language	# word types	# word instances
French	15354	39934
German	20286	45839
Xitsonga	1795	6384
Mandarin	3565	4132

In contrast to [10], we avoid pretraining on the set we evaluate on, and sample 50 hours of untranscribed data from multilingual LibriSpeech [25] for French and German, AIShell [26] for Mandarin, and a separate set in NCHLT [23] for Xitsonga. We use randomly sampled utterances to increase speaker diversity, which has been shown to be important for pre-trained models [27]. We use HuBERT BASE,[3] a 12-layer Transformer, implementated in Fairseq. We feed the untranscribed data from the target language to HuBERT, and run k-means with 500 clusters on the hidden vectors from the 10th layer of HuBERT. We use the cluster IDs of each frame as targets for continued pretraining.[4] We observe that after epoch 15, the performance stabilizes, so we do early stopping at epoch 15 for all languages. After continued pretraining, we average the hidden vectors from the 9th layer of HuBERT to construct the AWEs. The hyperparameters are tuned on the French dataset, and we will evaluate how well they generalize to other languages.

3.2. Results

Figure 1 shows the results for continued pretraining (HuBERT CP) on French, German, Mandarin, and Xitsonga. We compare to the original HuBERT (HuBERT EN) and to pretraining from scratch on a particular target language (HuBERT LANG). We observe that continued pretraining substantially outperforms others. In addition, in three of the four cases, pretraining from scratch on the target languages underperforms the HuBERT model pretrained on English.

These results further support the claim in [12] that by pretraining on 960 hours of English data, HuBERT `BASE` is able to learn a considerable amount of language-independent information that improves AWEs beyond what is learned from a smaller amount of target language data alone. However, it also indicates that a relatively small amount of target language data can successfully adapt the pre-trained English model and improve this type of mean-pooled AWEs.

[1] Note that Xitsonga and Mandarin are tonal languages and we do not explicitly model tones or consider them during evaluation (although HuBERT may implicitly capture some tonal information). In practice, there are very few pairs of words in our data that only differ in tones, so they should not have much effect in the evaluation.

[2] The list of test words, along with models and other materials used in our experiments, are available at `https://github.com/ramonsanabria/awe_ssl`.

[3] `https://github.com/facebookresearch/fairseq/tree/main/examples/hubert`

[4] We also explored using k-means centroids on MFCCs or hidden vectors of HuBERT on English, but using the k-means centroids on target languages consistently outperformed other units.

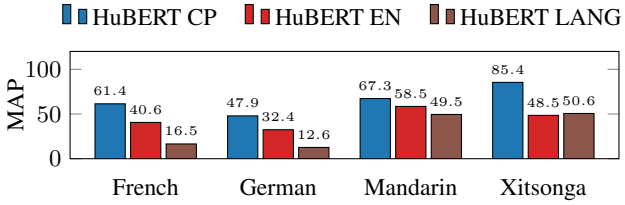

Figure 1: *Word discrimination performance on the Zerospeech 2017 Task 2 sets for French, German, Mandarin, and Xistonga, using mean-pooled HuBERT representations from the the HuBERT BASE model. The baseline model is pre-trained on English (HuBERT EN). Our method uses continued pretraining on 50 hours of speech from the target language (HuBERT CP), and we also compare to a model trained from scratch in each language (HuBERT LANG).*

4. Learned Pooling

As we have seen, AWEs can perform well by using simple pooling functions, such as mean-pooling. To improve the system further, we study the options of learning a pooling function for constructing AWEs. As we have detailed in Section 2, the goal is to learn a pooling function g, such that $g(z_{s:t})$ represents the segment between time s and t given the frame-level representation z of an utterance. Adhering to [9, 11, 19, 28], we focus on training the pooling function g with a contrastive loss — specifically, NTXent in (1). This loss function requires positive and negative examples that either need to be obtained from labelled data or to be mined with unsupervised approaches. Prior unsupervised work uses nearest neighbor search for mining contrastive examples, where positive examples are taken from the near neighbors and negative examples are taken from the complement. This approach can achieve a strong MAP, but nearest neighbor search is slow to compute and does not scale well when the data set becomes large.

We propose to use a multilingual phone recognizer (MPR) to label the untranscribed data with timings for each phone segment. Two speech segments are considered as positive pairs if they have the same phone sequence. Though the MPR system requires additional compute and data to train, we argue that the requirement is not as stringent as it seems, especially as pretrained models on high-resource languages are becoming more widely available; the use of HuBERT is one example, and the use of an external phone recognizer for unsupervised ASR in [29] is another.

4.1. Experimental Setting

We compare two approaches to mining contrastive examples for training the pooling function: the k-nearest neighbor (KNN) approach used in previous work, and the proposed MPR approach. For the nearest neighbor search, we first collect a set of random speech segments, ranging from 80 to 310 ms and being at least 80 ms apart. We represent each speech segment with mean-pooled HuBERT representations, and build an approximate nearest neighbor graph [11, 30] using dot-product as distance metric with FAISS [31].

The MPR system is a hybrid model based on time-delayed neural network [32] trained with lattice-free maximum mutual information [33]. The hybrid model is trained on English, Spanish, Russian, Polish, Portuguese, Bulgarian, Czech, Hausa, Swedish, and Ukrainian from GlobalPhone [34]—so it has not seen any of the languages that we evaluate the AWEs on. We collect speech segments with 2 to 5 phones. We define positive examples as the segments that have the same phone sequences, and negative examples as the complement. We sample a maximum of 300 n-gram instances for every n-gram type due to hard-drive limitation.

As opposed to wav2vec2.0 used in [10], we use HuBERT throughout the experiments due to its superior performance [12]. We use the same network architecture as in [10] to implement the pooling function. The network consists of a LayerNorm [35], a 1D convolution, and a transformer layer with 4 attention heads (including position embeddings) with a learning rate of 0.0001. The model finally max pools the frame-level representations to create the AWE. We train the pooling function (while fixing HuBERT) with a batch size of 150 for 5 epochs with a maximum 1000 iterations for each epoch.

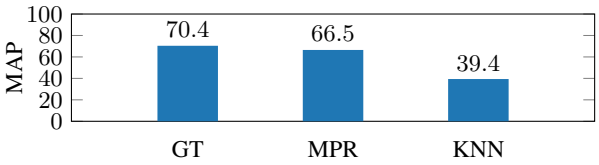

Figure 2: *Word discrimination performance on French, our development language, with a learned pooling operation using different methods for positive pair mining. All representations use HuBERT English frame-level representations from layer 7 as input. Ground Truth (GT) and Multilingual Phone Recognizer (MPR) use 2 to 5 gram pairs.*

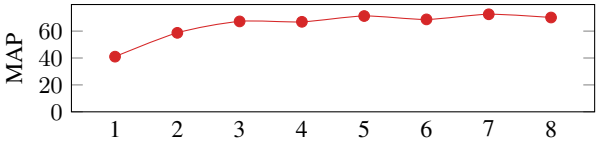

Figure 3: *Word discrimination performance of contrastive-based pooling trained on ground-truth phone ngrams from different sizes. Results are reported in our development set – French split from Zerospeech 2018 Task 2 set. We use HuBERT English frame-level representations from layer 7.*

4.2. Results and analysis

Results for the MPR approach and the KNN baseline are shown in Figure 2, along with an oracle approach that uses ground truth n-grams from forced alignment (GT). We evaluate the approaches on 5 hours of untrascribed French, with one million positive pairs. We find that our MPR approach performs nearly as well as using using ground truth n-grams, despite no training on the target language, and it works considerably better than the nearest neighbors approach. Moreover, for this 5h dataset, the training pairs took only five minutes to extract using MPR, versus 12 hours for the KNN approach.

The size of the speech segments we use for mining the positive and negative examples can potentially have a significant impact on the results. We perform a controlled experiment on the same 5 hour set with speech segments consisting of various number of forced-aligned phones to study the performance of the learned pooling function. Results are shown in Figure 3. In general, larger segment sizes perform better. Speech segments with at least 5 phones perform similarly, and we will use this setting for the rest of the experiments.

Table 2: *Word discrimination results (MAP, %) for the development language (French) and the test languages (DE=German; TS=Xitsonga; ZH=Mandarin). We compare continued pretraining (CP); learned pooling using a contrastive objective on positive pairs from a multilingual phone recognizer (LP); the combination of both (CP + LP); and the method from [10] (IKNN). Training uses 50h of data unless otherwise specified.*

	FR	DE	ZH	TS
CP	61.4	47.9	67.3	85.4
LP	66.6	70.2	76.1	89.8
CP + LP	69.5	**74.5**	**80.6**	**92.9**
IKNN (20h)	41.1	40.0	52.9	57.8
CP + LP (20h)	**69.6**	72.1	79.3	87.8

5. Comparing and Combining Methods

In the previous sections, we investigate configurations and show that our techniques achieve good performance. Now, we ask *how do both methods compare to each other, and can they be effectively combined?* We also include results of the iterative nearest neighbor (IKNN) approach proposed in [10] as a baseline. We use their implementation[5] but reduce the batch size from 250 to 150 to fit on a 12 GB GeForce GTX 1080 Ti. Instead of training with the test set as in [10], we use a separate training set to be more comparable to our own approach. Because IKNN requires large amounts of memory, we were not able to run it on the full 50h training set, so we report results for 20h of training data for IKNN and also for our own best system. We train our contrastive pooling model on 5-gram MPR positive pairs due to the results observed in Figure 3.

Table 2 presents the results on all test languages. We observe that learned pooling with pre-trained (English) HuBERT features outperforms CP the four languages, and that a small further improvement is obtained by combining the two approaches. Results are almost as good with only 20h of data, and considerably outperform the comparison approach.

Since we are interested in a low-resource setting, we finally explore the data efficiency of each method, by reducing the amount of training data used. We create subsets of different sizes by randomly sampling utterances from the 50-hour dataset. We use the same data to train all components (CP, K-means, and the learned pooling). Since longer n-gram pairs may be limited for some of the smaller data sizes, we use all 2-5 grams in all settings.

Figure 4 (top) presents the results. We observe that while all models reach a similar performance when trained on 50 hours of data, learned pooling achieves nearly all of the gain with only about three hours of data, indicating far greater data efficiency. This result accords with previous work: e.g., [7] showed that for three learned pooling methods, performance had begun to level off by 50k training pairs (the maximum they tested). It is worth noting that until [10], it was assumed that systems should be trained on ground truth words or word-like discovered terms, leading to far fewer pairs than the n-grams used by [10] and this paper—thus, many systems were trained on only 5k-20k pairs [7, 8, 6]. With our approach, one hour of data yields about 1M pairs, and 10 minutes yields 24k pairs.

Inspired by the results above, we further explore the data

[5] https://gitlab.cognitive-ml.fr/ralgayres/ssemodel

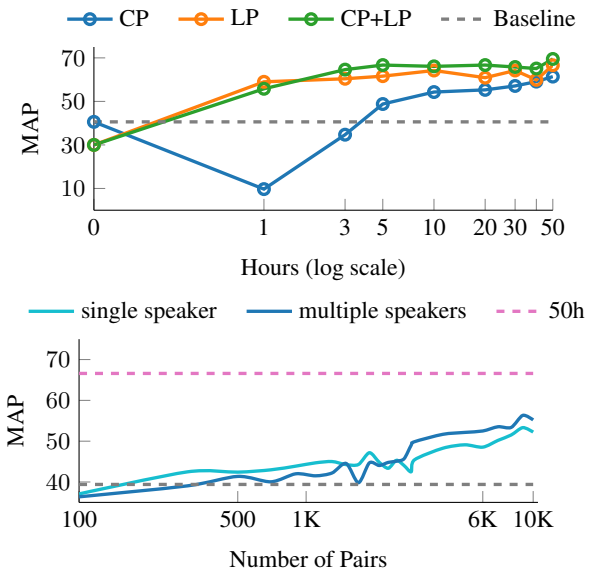

Figure 4: *Effects of training data size. (Top) Training on 1-50h of data, using mean-pooled HuBERT with continued pretraining (CP), non-CP HuBERT with learned pooling (LP), and CP using learned pooling (CP+LP); the baseline is mean-pooled (non-CP) HuBERT. All results are on the French dataset. (Bottom) Training the LP model on 100-10k pairs (from $<$ 10m of data), from either a single speaker or multiple speakers; the topline is LP trained on 50h.*

requirements of the learned pooling technique, by looking at very low data regimes and the effects of speaker diversity. We sample from 100 to 10k positive training pairs, which are either randomly sampled from the 50h multi-speaker dataset, or from a single speaker. Figure 4 (bottom) shows the results, indicating improvements over the baseline with just a few hundred training pairs, and little difference between single-speaker and multi-speaker training. These results suggest that the pre-trained HuBERT features are already doing a good job of speaker normalization, and only a relatively simple learned pooling function is needed to improve over mean-pooling. This contrasts with earlier work that learned AWEs using MFCC input features, where training pairs from multiple speakers were needed to help overcome speaker differences [6].

6. Conclusions

We propose two techniques to adapt English self-supervised acoustic word embedding representations to a target language with up to 50 hours of unlabeled data. We first propose to adapt English frame-level representations to a target language by continued pretraining (CP). Our results using mean-pooling show show that CP is highly effective and can outperform the original model with only 10 hours of data. Next, we show that one can achieve similar performance by training a pooling mechanism on top of the self-supervied representations, using contrastive-learning with positive phone n-gram pairs obtained by a multilingual phone recognizer. The MPR method is fast and returns a large number of high-quality pairs, leading to better word discrimination than a previous approach. It is also extremely data-efficient, requiring only a few hours of target language data to reach its best results, and outperforming the previous approach with less than 1h of data. Finally, we show that the two methods can be combined, leading to the best overall results.

7. References

[1] A. L. Maas, S. D. Miller, T. M. O'neil, A. Y. Ng, and P. Nguyen, "Word-level Acoustic Modeling with Convolutional Vector Regression," in *International Conference on Machine Learning (ICML) Workshop*, 2012.

[2] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional Acoustic Embeddings of Variable-length Segments in Low-resource Settings," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.

[3] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example Search with Discriminative Neural Acoustic Word Embeddings," in *Interspeech*, 2017.

[4] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic Speech Retrieval with a Visually Grounded Model of Untranscribed Speech," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2018.

[5] H. Kamper, A. Jansen, and S. Goldwater, "A Segmental Framework for Fully-unsupervised Large-vocabulary Speech Recognition," in *Computer Speech and Language*, 2017.

[6] H. Kamper, "Truly Unsupervised Acoustic Word Embeddings Using Weak Top-down Constraints in Encoder-decoder Models," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[7] P. Peng, H. Kamper, and K. Livescu, "A Correspondence Variational Autoencoder for unsupervised Acoustic Word Embeddings," in *Advances in Neural Information Processing Systems (NeurIPS) Wokshop*, 2020.

[8] L. Van Staden and H. Kamper, "A Comparison of Self-supervised Speech Representations as Input Features for Unsupervised Acoustic Word Embeddings," in *Spoken Language Technology Workshop (SLT)*, 2021.

[9] C. Jacobs and H. Kamper, "Multilingual Transfer of Acoustic Word Embeddings Improves when Training on Languages related to the Target Zero-resource Language," in *Interspeech*, 2021.

[10] R. Algayres, A. Nabli, B. Sagot, and E. Dupoux, "Speech Sequence Embeddings using Nearest Neighbors Contrastive Learning," in *Interspeech*, 2022.

[11] A. Jansen and B. Van Durme, "Efficient Spoken Term Discovery Using Randomized Algorithms," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[12] R. Sanabria, H. Tang, and S. Goldwater, "Analyzing Acoustic Word Embeddings from Pre-trained Self-supervised Speech Models," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[13] 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," in *Transactions on Audio Speech and Language Processing (TASLP)*, 2021.

[14] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *Association for Computational Linguistics (ACL)*, 2020.

[15] J.-H. Lee, C.-W. Lee, J.-S. Choi, J.-H. Chang, W. K. Seong, and J. Lee, "CTRL: Continual Representation Learning to Transfer Information of Pre-trained for Wav2vec 2.0," in *Interspeech*, 2022.

[16] K. Nowakowski, M. Ptaszynski, K. Murasaki, and J. Nieuważny, "Adapting Multilingual Speech Representation Model for a new, Underresourced Language Through Multilingual Fine-tuning and Continued Pretraining," in *Information Processing and Management*.

[17] 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 (NeurIPS)*, 2020.

[18] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid Evaluation of Speech Representations for Spoken Term Discovery," in *Interspeech*, 2011.

[19] C. Jacobs, Y. Matusevych, and H. Kamper, "Acoustic Word Embeddings for Zero-resource Languages using Self-supervised Contrastive Learning and Multilingual Adaptation," in *Spoken Language Technology Workshop (SLT)*, 2021.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International Conference on Machine Learning (ICML)*, 2020.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," in *Interspeech*, 2020.

[22] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017.

[23] E. Barnard, M. H. Davel, C. van Heerden, F. De Wet, and J. Badenhorst, "The NCHLT Speech Corpus of the South African Languages," in *Workshop Spoken Language Technologies for Under-resourced Languages (SLTU)*, 2014.

[24] H. Kamper, "Unsupervised Neural and Bayesian Models for Zero-resource Speech Processing," in *Thesis*, 2017.

[25] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIshell-1: An Open-source Mandarin Speech Corpus and a Speech Recognition Baseline," in *Oriental COCOSDA*.

[27] R. Sanabria, W.-N. Hsu, A. Baevski, and M. Auli, "Measuring The Impact of Individual Domain Factors in Self-supervised Pre-training," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Workshop*, 2023.

[28] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating The Reliability of Acoustic Speech Embeddings," *Interspeech*, 2020.

[29] O. Klejch, E. Wallington, and P. Bell, "Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR," in *Interspeech*, 2022.

[30] A. Thual, C. Dancette, J. Karadayi, J. Benjumea, and E. Dupoux, "A K-nearest Neighbours Approach to Unsupervised Spoken Term Discovery," in *Spoken Language Technology Workshop (SLT)*, 2018.

[31] J. Johnson, M. Douze, and H. Jégou, "Billion-scale Similarity Search with GPUs," in *Transactions on Big Data*, 2019.

[32] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-delay Neural Networks," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989.

[33] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-trained Neural Networks for ASR based on Lattice-free MMI," in *Interspeech*, 2016.

[34] T. Schultz, "GlobalPhone: a Multilingual Speech and Text Database Developed at Karlsruhe University," in *Conference on Spoken Language Processing*, 2002.

[35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *pre-print*, 2016.