INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Self-Supervised Acoustic Word Embedding Learning via Correspondence Transformer Encoder

Jingru Lin[1], *Xianghu Yue*[1], *Junyi Ao*[2], *Haizhou Li*[1,2]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

{jingrulin, xianghu.yue}@u.nus.edu, junyiao1@link.cuhk.edu.cn, haizhouli@cuhk.edu.cn

Abstract

Acoustic word embeddings (AWEs) aims to map a variable-length speech segment into a fixed-dimensional representation. High-quality AWEs should be invariant to variations, such as duration, pitch and speaker. In this paper, we introduce a novel self-supervised method to learn robust AWEs from a large-scale unlabelled speech corpus. Our model, named Correspondence Transformer Encoder (CTE), employs a teacher-student learning framework. We train the model based on the idea that different realisations of the same word should be close in the underlying embedding space. Specifically, we feed the teacher and student encoder with different acoustic instances of the same word and pre-train the model with a word-level loss. Our experiments show that the embeddings extracted from the proposed CTE model are robust to speech variations, e.g. speakers and domains. Additionally, when evaluated on Xitsonga, a low-resource cross-lingual setting, the CTE model achieves new state-of-the-art performance.

Index Terms: acoustic word embedding, self-supervised learning, low-resource

1. Introduction

Mapping arbitrary-length words into fixed-dimensional vector representations is very useful for many speech processing tasks [1], such as query-by-example [2, 3] and speech recognition [4, 5]. Once word segments are represented as fixed-dimensional vectors, they can be compared through simple cosine or Euclidean distance efficiently, or directly applied to downstream tasks' classifiers [6, 7, 8].

The earliest works try to embed variable-length acoustic words into fixed-dimensional vectors using simple heuristic approaches such as down-sampling [9]. However, these vector representations may not be able to precisely describe the structure of the audio segments for the reason that the speech instances of the same word will never be identical due to the variations in duration, pitch, speakers' accent and gender, etc. This makes learning acoustic word embeddings (AWEs) more challenging than textual word embeddings [10, 11]. The former should generate similar representations from different realisations of the same acoustic word despite the variations, while the latter only needs to describe one unique character sequence for each word.

On the other hand, deep learning, which seeks to learn from the data, has been more successful in describing the acoustic word structures against the variations [12]. Some earlier works that applied deep learning to learn AWEs relied on using the true word identity [13, 5, 14]. Nonetheless, learning from a large amount of annotated data is not only expensive but also contradictory to the way that human infants first acquire languages where little supervision is needed [15]. This drives the research for AWEs towards unsupervised learning where AWEs are learned from unannotated speech data. In unsupervised settings, the boundaries of words in a speech sentence and the identity of the acoustic words are unknown. Therefore, the study of AWEs in unsupervised settings has practical implications for zero- or low-resource settings where transcribed data is unavailable or very scarce.

Generally, in the unsupervised setting, only word pairs ('same' or 'different' to indicate if a word pair has the same or different identity) or word boundary information are needed as a form of weak supervision. For example, Kamper *et al.* applies convolution neural networks in a Siamese setting [16], where the Siamese networks learn to maximise/minimise the distance between words of different/same types; other works [17, 18] also use word pairs information but with different model architectures. Chung *et al.* segments speech sentences into acoustic words based on word boundary information and implements an autoencoder model, which encodes the extracted acoustic words input into fixed-length representations, and then reconstructs this word input out from the representations [1]. Although these works make use of weak supervision during training, advances in zero-resource technology makes it possible to obtain the word pairs and boundary information from speech corpus in unsupervised ways [19, 20]. To further improve the acoustic embeddings, some researchers exploit the use of the words' character sequences [21, 22]. However, the use of character sequences violates the purpose of unsupervised learning and hinder the applications to the low- or zero-resource setting as the transcribed data is not easily available in these settings.

In this work, we extend the research on unsupervised AWEs. In contrast to the above approaches where complex models are trained in attempts to model as much information from the limited target resources as possible, here we want to learn a general representation of acoustic words from a larger resource. When applied to low- or zero-resource settings, the models should work as a robust feature extractor and efficiently adapt to new resources. For this purpose, we propose Correspondence Transformer Encoder (CTE), a novel self-supervised technique to learn general and robust AWEs by leveraging large amounts of unlabelled speech data. Our CTE is built based on the teacher-student learning framework. Specifically, CTE has a student and a teacher encoder that share the same architecture. During pre-training, the student and teacher encoder each takes an acoustic instance of the same word. The student encoder is optimized to minimise the distance between the output representations generated by itself and the teacher encoder, and the teacher encoder is parameterised by an exponentially moving average (EMA) of the student network. In downstream evaluations, the representations generated by the student encoder are used as word embeddings.

10.21437/Interspeech.2023-359

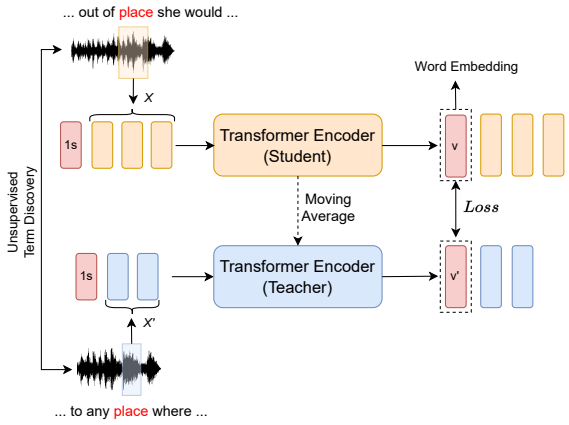

Figure 1: *The overall model architecture of CTE. The teacher encoder and student encoder are fed with different acoustic instances of the same word. The teacher is parameterised by an exponentially moving average of the student weights, while the student predicts the average representation of top K layers from the teacher.*

We verify the effectiveness of our model on both in-, cross-domain, and cross-lingual settings. Experimental results show that our proposed CTE learns more robust word representations while discarding other irrelevant information that is sensitive to variations in speech. To the best of our knowledge, this is the first work that utilises a transformer architecture to learn acoustic word embeddings.

2. Methodology

The goal of our AWE models is to learn a function f such that $f(x)$ can map variable-length acoustic word segment x into a fixed-dimensional word embedding v. It is desired that different acoustic segments representing the same word should be mapped to close proximity in the embeddings space, regardless of their variations. Below we formally present the CTE model.

2.1. Model architecture

The overall architecture of the proposed correspondence transformer encoder CTE is shown in Fig. 1. Our motivation comes from the premise that different acoustic realisations of the same word should be close in the underlying embedding space.

For CTE, a single training input consists of a pair of audio segments that have the same word identity. Given the pair of audio segments, we first extract the 80-dimensional log Mel-filterbank, denoted as (X, X'), where $X = [x_1, \ldots, x_{t_1}]$ and $X' = [x'_1, \ldots, x'_{t_2}]$ with lengths t_1 and t_2, respectively. We pass the acoustic features X to the student encoder and X' to the teacher encoder. The student and teacher encoders are both transformer encoders, and the teacher parameters are an exponentially moving average of the student weights.

To obtain fixed-dimensional word-level representations, taking inspiration from BERT, we add a random vector at the first timestep of the input features and use the first-timestep vector from the representation as the final word embeddings. In practise, the first-timestep random vector can be initialised with all ones or zeros. Here, we only use ones to denote. The representations obtained from X and X' are hence given as:

$$[v, h] = Encoder_{student}([\mathbf{1}, X]) \tag{1}$$

$$[v', h'] = Encoder_{teacher}([\mathbf{1}, X']) \tag{2}$$

where $Encoder_{student}$ and $Encoder_{teacher}$ are the student and teacher encoder respectively, $\mathbf{1}$ is the all-ones vector added to the first timestep of the input features, v and v' are the first-timestep vectors of the encoded representations which are regarded as the respective acoustic embeddings of the input word pair X and X'.

2.2. Self-supervised learning task

Given (X, X'), the model is trained to learn the word embedding v from X that can predict the word embedding v' from X'. This self-supervised learning target v' is constructed based on the first timestep of the output representations from the top K blocks of the teacher encoder. We first apply a layer normalization to each block, and then average the top K blocks to construct v':

$$v' = \frac{1}{K} \sum_{l=L-K+1}^{L} v_l \tag{3}$$

where v_l is the first-timestep vector from l-th layer in the teacher encoder and L is the total number of layers in the encoder. The vector v' is to be predicted by the student encoder. Hence, CTE is optimized by minimising the cosine distance between the vector representations v and v'. The training loss $\mathcal{L}$ is given by:

$$\mathcal{L} = 1 - cos(v, v') = 1 - \frac{v \cdot v'}{||v||_2 \cdot ||v'||_2} \tag{4}$$

In CTE, the student and teacher encoders share the same architecture, but differ in the way they are parameterised. The training loss $\mathcal{L}$ updates the parameters θ of the student encoder while the teacher encoder is parameterised by an exponentially moving average (EMA) of the student parameters. This means, given a target decay rate τ, after each training step, the teacher encoder parameters ξ are given by:

$$\xi = \tau\xi + (1 - \tau)\theta \tag{5}$$

2.3. Word embeddings

In the inference stage, the teacher encoder is discarded and only the student encoder is needed to produce the word embeddings. Given an audio segment, the student encoder takes in its 80-dimensional log Mel-filterbank features, together with an appended ones token at the first timestep, and generates corresponding representations. The first timestep vector of the encoded representations servers as the final acoustic word embedding. This is labeled as v in Fig. 1.

3. Experiments

3.1. Training datasets

We use the *train-clean-100* and *train-clean-360* split of the publicly available LibriSpeech dataset [23] for training. The force-alignment transcriptions generated by Montreal Forced Aligner [24] are used to get the word boundaries and pairs. The duration of word segments ranges from 0.5 to 2 seconds. During training, the model takes a word pair which is the different acoustic instances of the same word. The true labels of the words are not used. For a word pair (X, Y), we also include (Y, X) in the training.

Table 1: *Summary for model architecture and dataset used for CTE Small and Base models.*

		Small	Base
Input Feature		80-dim FBANK	80-dim FBANK
Transformer	layer	6	12
	embedding dim.	256	512
	inner FFN dim.	1024	2048
	attention heads	4	8
	average K	4	8
Training Set	LibriSpeech	100 hr	460 hr
No. of Word Pairs		121k	321k
Training Steps		50k	60k

3.2. Models and training details

All the models are trained with fairseq toolkit [25]. We experiment with two model configurations that share the same encoder architecture but differ in model size. Table 1 summarises all the model parameters for the *small* and *base* models, along with the respective data used. The *small* model consists of 6 Transformer encoder layers, model embedding dimension 256, inner feed-forward network dimension 1024 and 4 attention heads, while the *base* model consists of 12 Transformer encoder layers, model embedding dimension 512, inner feed-forward network dimension 2048 and 8 attention heads. The number of blocks that are used to construct the self-supervised learning targets is 4 and 8 respectively for the CTE *small* and the *base* model. We use 80-dimensional log mel filter bank with a frame length of 25 ms and an overlap of 10 ms as the input features. The models are optimized with Adam [26]. For the teacher parameterisation, the target decay rate, τ, is set to 0.999.

3.3. Evaluation tasks

To assess the quality of the vector representations learned by CTE models, we first conduct a series of analyses on the in-domain LibriSpeech dastaset. We use the CTE learned from the training datasets to encode audio segments in the test sets, which are sourced from LibriSpeech dev-clean and test-clean. It is worth noting that the audio segments used in both training and testing come from in-domain datasets but are mutually exclusive and from different speakers.

Next, we measure the intrinsic quality of the word embeddings without being tied to a particular downstream task. The same-different task [27], which involves determining whether a given pair of acoustic segments, each representing a true word, are of the same or different word types, is designed for this purpose. The evaluation metric used is the average precision (AP), which is obtained from the area under ROC curve. We conduct the same-different experiments on two different datasets: a cross-domain English dataset and a cross-lingual Xitsonga dataset. For the English dataset, we use speech from Buckeye corpus of conversational English [28], which contains 6 hours of speech for each of the train, dev and test splits respectively. As for Xitsonga, we use a 2.5-hour portion from NCHLT corpus [29]. For adapting to the different domain and language, we fine-tune the CTE models on the word pairs obtained from respective datasets. The word pairs are obtained by either ground truth transcriptions or the unsupervised term discovery (UTD) system [19], in which the former sets an upper bound and the latter simulates a low-resource setting. We fine-tune the CTE models for a total of 5k steps in each experiment.

Through these experiments, we evaluate both the generalizing ability of CTE models to a new domain, as well as the transferability of CTE representations to a new unseen language. For a fair comparison, the train, validation and test splits follow the common experimental setting in [18]. The AP score is reported for all the experiments.

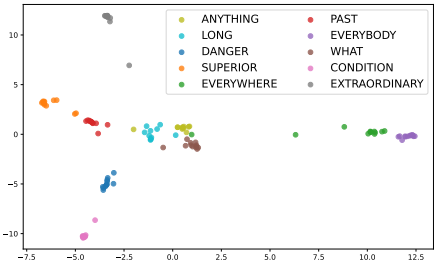

Figure 2: *Word embeddings for 10 randomly selected words.*

4. Results and Analysis

4.1. In-domain analysis

Our CTE model aims to map the variable-length speech segments into a fixed-dimensional representation and puts the same word with different acoustic variations close in the underlying space. To demonstrate this, we visualise the word embeddings extracted directly from CTE *base* model. Fig. 2 shows the embeddings extracted from CTE *base* model for 10 randomly selected words from different speakers in the test set described in Section 3.3. These embeddings are reduced to two dimensions using principal component analysis (PCA). From the plot, we can see that most of the embeddings of the same words are clustered in close proximity to each other, showing that we have obtained discriminative word embeddings for the unseen speakers. This means our CTE models have learned to disentangle unnecessary speaker information and extracted speaker-independent word-level information.

Next, we analyse if the embeddings obtained can effectively distinguish words with similar phonetic structures. We compute the average cosine similarity for each pair of acoustic segments against the phoneme sequence edit distance (PSED), shown in Fig. 3. From the plot, it is obvious that word pairs with larger PSED have smaller cosine similarities. Notably, the cosine similarity is largest for PSED = 0, and there is a clear drop from 0.789 to 0.492 as PSED increases from 0 to 1. This clear distinction is useful for applications that require high recall. As PSED increases to 4 or more, the average cosine similarity drops to 0.078 only. The gradual decrease in cosine similarity with increasing PSED also indicates that the word embeddings are able to describe the sequential phoneme structure despite being trained with only word-level loss.

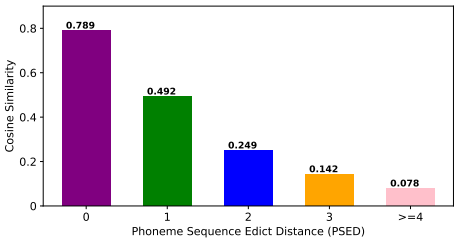

Figure 3: *The average cosine similarity between embeddings for different phoneme sequence edit distance (PSED).*

4.2. Cross-domain analysis

4.2.1. Ground-truth upper bounds

In this section, we evaluate the generalizability of the proposed CTE model across a different domain. Table 2 presents the average precision (AP) score on Buckeye English validation and test sets for different models, with training word pairs obtained from ground-truth transcriptions. As a strong baseline, we include HuBERT [32], a self-supervised learning model that is pre-trained on the 60k hour split from the Libri-light dataset [33]. The word embeddings are obtained by either mean pooling or subsampling the representations from HuBERT Large [31]. Except for HuBERT, other baseline models are trained on the same word pairs from the Buckeye datasets.

As reported in Table 2, the strongest baseline models are HuBERT and MCVAE [30]. Specifically, HuBERT with mean pooling has achieved an AP score of 67.5% and 67.8% on validation and test sets, while MCVAE has achieved an AP score of 58.8% on the validation set. In comparison, CTE models show substantial improvements over all other embedding approaches: CTE *small* achieves 67.0% and 71.7% and CTE *base* achieves 72.3% and 75.5% on the validation and test sets. This shows that the embeddings learned capture acoustic information that is general and transferable to other domains. Moreover, the fact that our CTE models outperform HuBERT, which is pre-trained on a much larger dataset, also implies that our word-level training strategy is more robust to variations and effective in capturing word-level information than the frame-level training strategy that is adopted in HuBERT.

4.2.2. Unsupervised setting

Table 3 presents the results for training using pairs discovered by the UTD system [19], which simulates the low-resource setting. In this setting, the training pairs are discovered by a UTD system, while a small set of ground truth pairs are used as the validation data. The training pairs discovered by the UTD system inevitably suffer from erroneous matches. Therefore, instead of using complex models to learn from potentially inaccurate training pairs, our models can leverage prior knowledge to mitigate the effects of inaccuracy. The experimental results show that the CTE *base* model exhibits promising performance in the low-resource setting, achieving an AP score of 43.2% on validation set and 44.1% on test set.

4.3. Cross-lingual analysis

The above experiments have shown promising results in English even when in-domain data is scarce. However, there are many low-resource languages where a large corpus might not be available. Therefore, our objective here is to investigate the transferability of CTE models trained in English when employed in a different language. In Table 3, we show the performance of word embeddings obtained from the CTE models when applied to Xitsonga, an unseen low-resource language. Among all the models, the CTE *base* model achieves an AP score of 46.7%, which is comparable with HuBERT with a subsampling pooling strategy. However, note that HuBERT with subsampling pooing strategy produces a word embedding with a much larger dimensionality (10240 dimensions), which is less efficient for applications with limited memory.

Interestingly, the CTE models, when evaluated on Xitsonga, outperform itself when evaluated on Buckeye English. This trend can be attributed to the higher pair-wise matching precision in Xitsonga than in Buckeye English. Consequently, this finding suggests that the CTE model is robust to changes in language while still being sensitive to the quality of training pairs, despite the beneficial effects of pre-training in mitigating the impact of inaccurate training pairs as shown in Section 4.2.2.

Table 2: *Average precision on Buckeye English validation set and test set for different models using ground truth word pairs. As we take the statistics from the papers directly, some average precision of the test sets are not reported in their papers.*

Model	Validation	Test
CTE small	67.0	71.7
CTE base	**73.3**	**76.3**
Downsampling [18]	24.5	21.7
DTW alignment [18]	36.8	35.9
ENCDEC-CAE [18]	51.1	-
MCVAE [30]	58.8	-
HuBERT (EN) mean [31]	67.5	67.8
HuBERT (EN) subsample [31]	65.3	64.8

Table 3: *Average precision on the cross-domain Buckeye English validation and test datasets, as well as the cross-lingual Xitsonga test set for different models. The training pairs for both datasets are discovered using the unsupervised term discovery system (UTD).*

Model	English		Xitsonga
	Valid	Test	Test
CTE small	32.9	32.8	31.1
CTE base	**43.2**	**44.1**	**46.7**
Downsampling [18]	24.5	21.7	13.6
DTW alignment [18]	36.8	35.9	28.1
CAE-RNN [34]	-	36.8	41.8
ENCDEC-CAE [18]	31.7	32.2	32.0
MCVAE [30]	37.6	39.5	44.4
HuBERT (EN) mean [31]	-	-	39.0
HuBERT (EN) subsample [31]	-	-	46.0

5. Conclusion

In this paper, we propose a novel self-supervised framework, CTE, to learn robust acoustic word embeddings from the large-scale unlabeled speech corpus and achieves state-of-the-art performances. Our experiments have shown that the word embeddings extracted from the proposed CTE model exhibits different levels of robustness, particularly in their robustness towards the change of speakers, domain and language, which positions it as a competitive model in learning acoustic word embeddings. Future work includes scaling up the model size and training data to increase the model capacity, improving the quality of word pairs obtained from the UTD system and applying the CTE models to more diverse downstream tasks.

6. Acknowledgements

This work is supported by 1) Huawei Noah's Ark Lab; 2) National Natural Science Foundation of China (Grant No. 62271432); 3) Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046)

7. References

[1] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *Interspeech*, pp. 765–769, 2016.

[2] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," *Interspeech*, pp. 2874–2878, 2017.

[3] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Learning acoustic word embeddings with temporal context for query-by-example speech search," *Interspeech*, pp. 97–101, 2018.

[4] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5641–5645.

[5] S. Bengio and G. Heigold, "Word embeddings for speech recognition," 2014.

[6] X. Yang, J. Ye, and X. Wang, "Factorizing knowledge in neural networks," in *European Conference on Computer Vision*, 2022.

[7] J. Ye, S. Liu, and X. Wang, "Partial network cloning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] X. Yang, D. Zhou, S. Liu, J. Ye, and X. Wang, "Deep model reassembly," in *Advances in Neural Information Processing Systems,*, 2022.

[9] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *2013 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2013, pp. 410–415.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] X. Gao, X. Yue, and H. Li, "Self-transriber: Few-shot lyrics transcription with self-training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] 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 *Proc. ICML Workshop Representation Learn*, 2012.

[14] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5236–5240.

[15] Y.-C. Chen, S.-F. Huang, H.-y. Lee, Y.-H. Wang, and C.-H. Shen, "Audio word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1481–1493, 2019.

[16] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4950–4954.

[17] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5818–5822.

[18] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6535–3539.

[19] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2011.

[20] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," *arXiv preprint arXiv:1611.04496*, 2016.

[22] M. Jung, H. Lim, J. Goo, Y. Jung, and H. Kim, "Additional shared decoder on siamese multi-view encoders for learning acoustic word embeddings," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 629–636.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[25] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[28] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[29] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based asr data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[30] P. Peng, H. Kamper, and K. Livescu, "A correspondence variational autoencoder for unsupervised acoustic word embeddings," *arXiv preprint arXiv:2012.02221*, 2020.

[31] R. Sanabria, H. Tang, and S. Goldwater, "Analyzing acoustic word embeddings from pre-trained self-supervised speech models," *arXiv preprint arXiv:2210.16043*, 2022.

[32] 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.

[33] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[34] L. Van Staden and H. Kamper, "A comparison of self-supervised speech representations as input features for unsupervised acoustic word embeddings," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 927–934.