INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multilingual Jointly Trained Acoustic and Written Word Embeddings

Yushi Hu[1], *Shane Settle*[2], *Karen Livescu*[2]

[1]University of Chicago
[2]TTI-Chicago

hys98@uchicago.edu, {settle.shane,klivescu}@ttic.edu

Abstract

Acoustic word embeddings (AWEs) are vector representations of spoken word segments. AWEs can be learned jointly with embeddings of character sequences, to generate phonetically meaningful embeddings of written words, or acoustically grounded word embeddings (AGWEs). Such embeddings have been used to improve speech retrieval, recognition, and spoken term discovery. In this work, we extend this idea to multiple low-resource languages. We jointly train an AWE model and an AGWE model, using phonetically transcribed data from multiple languages. The pre-trained models can then be used for unseen zero-resource languages, or fine-tuned on data from low-resource languages. We also investigate distinctive features, as an alternative to phone labels, to better share cross-lingual information. We test our models on word discrimination tasks for twelve languages. When trained on eleven languages and tested on the remaining unseen language, our model outperforms traditional unsupervised approaches like dynamic time warping. After fine-tuning the pre-trained models on one hour or even ten minutes of data from a new language, performance is typically much better than training on only the target-language data. We also find that phonetic supervision improves performance over character sequences, and that distinctive feature supervision is helpful in handling unseen phones in the target language.

Index Terms: acoustic word embeddings, multilingual, low-resource, zero-resource, distinctive features.

1. Introduction

Acoustic word embeddings (AWEs) are vector representations of spoken word segments of arbitrary duration [1]. AWEs are an attractive tool in tasks involving reasoning about whole word segments, as they provide a compact representation of spoken words and can be used to efficiently measure similarity between segments. For example, AWEs have been used to speed up and improve query-by-example search [2–4], where the distance computation traditionally done with dynamic time warping or subword methods is replaced by vector distances between embeddings. AWEs have also been used for unsupervised segmentation and spoken term discovery [5], where the use of embeddings can eliminate the need to explicitly model subword units.

A variety of approaches have been explored for constructing and learning AWEs, including template-based techniques [1], discriminative neural network models [6, 7], and unsupervised autoencoder-based models [8–12]. Despite the flexibility of fully unsupervised methods, supervised AWE training can be highly data-efficient, providing large performance improvements over unsupervised methods on downstream tasks with just 100 minutes of labelled audio [3].

Some tasks involve measuring similarity between spoken and written words. For such purposes it can be useful to learn both embeddings of spoken words and embeddings of written words that represent the word's phonetic content, which we refer to as *acoustically grounded word embeddings* (AGWEs). For example, AWEs and AGWEs have been used together for spoken term detection [10], where a written query word can be compared to speech segments directly via vector similarity between their embeddings; and to whole-word speech recognition, where such embeddings have been used either for rescoring [14] or to initialize an end-to-end neural model and improve recognition of rare and out-of-vocabulary words [15]. AWEs and AGWEs can be learned jointly [14, 16] to embed spoken and written words in the same vector space (Figure 1).

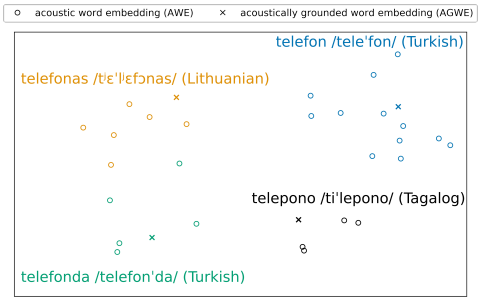

Figure 1: *Examples of jointly trained acoustic and written word embeddings, for the Turkish word "telefon" and several of its nearest neighbors, visualized with t-SNE [13]. The acoustic word embeddings (AWEs) are clustered around the corresponding acoustically grounded written word embeddings (AGWEs).*

Prior work in this area has largely focused on English. In this work, we study the learning of AWEs/AGWEs for multiple languages, in particular low-resource languages. Recent related work [9] has begun to explore multilingual AWEs, specifically for zero-resource languages, including unsupervised approaches trained on a zero-resource language of interest and supervised models trained on multiple additional languages and applied to a zero-resource language. Our work complements this prior work by exploring, in addition to the zero-resource regime, a number of low-resource settings, and the trade-off between performance and data availability. In addition, we learn not only AWEs but also AGWEs, thus widening the range of tasks to which our models apply. One product of our work is a set of written word embeddings for multiple languages in the same phonetically meaningful embedding space.

Our approach extends prior work on AWE+AGWE training for English, based on joint embedding training with a multi-view contrastive loss [15, 16]. In contrast with prior work, we use phonetic pronunciations as supervision rather than characters, to avoid the difficulties introduced by multiple written alphabets. In addition, to better model rare or unseen phones, we explore the use of distinctive features as an alternative. We evaluate our embeddings by their performance on word discrimination, a measure that is independent of any particular downstream task.

http://dx.doi.org/10.21437/Interspeech.2020-2828

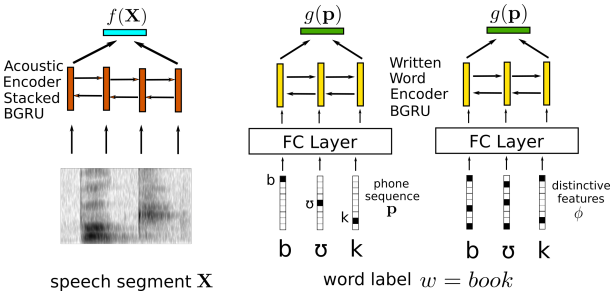

Figure 2: *Acoustic word embedding (AWE) model f and two acoustically grounded word embedding (AGWE) models g, corresponding to either phone or distinctive feature sequence input.*

Our main finding is that, in low-resource settings, pre-training embeddings on multiple languages and fine-tuning on a small amount of target language data produces much higher-quality embeddings than training on target language data alone. In addition, performance is improved by using a pronunciation dictionary to embed the phone sequences of written words, rather than using the character sequence as in prior work. We also find that using distinctive features in place of phones has advantages in settings where there are unseen phones in the target language. We provide results showing how embedding quality changes with increasing training set size, and in particular find that modest data sizes ($\sim$ 10 hrs) are likely sufficient for training AWEs/AGWEs with our approach.

2. Embedding Models

An acoustic word embedding (AWE) model f maps a variable-length spoken segment $\mathbf{X} \in \mathbb{R}^{T \times D}$, where T is the number of acoustic frames and D is the frame feature dimensionality, to an embedding vector $f(\mathbf{X}) \in \mathbb{R}^d$. The goal is to learn f such that segments corresponding to the same word are embedded close together, while segments of differing words are embedded farther apart. We use supervised training and leverage the labeled data available in multiple languages besides the target language, in order to improve the performance on low-resource or zero-resource target languages.

2.1. Embedding learning with phone-based supervision

Our approach is based on prior work on AWE/AGWE learning using a multi-view contrastive loss [15, 16], which jointly learns models for an acoustic view (f) and a written view (g) (Figure 2). The input to the acoustic embedding model f is a variable-length spoken segment corresponding to a word. In prior work [15, 16], the input to g is a character sequence, but for extension to multiple languages with widely differing writing systems, we instead use a phone sequence as input to g; specifically, we use the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) [17]. We obtain the phonetic transcription from the written one using a pronunciation dictionary. Therefore, we do not require manual phonetic transcriptions, but do require a pronunciation dictionary for each language. Once a multilingual model is trained, however, the acoustic model f can be used to embed spoken word segments in an unseen language, regardless of whether or not we have a pronunciation dictionary for it. To further improve cross-lingual transfer, we also explore using sequences of distinctive features in place of phones (Section 2.2).

A training set for a particular language l consists of pairs $\{(\mathbf{X}_i^l, w_i^l)\}_{i=1}^{N^l}$, where $\mathbf{X}_i^l \in \mathbb{R}^{T_i^l \times D}$ is a spoken word segment and $w_i^l \in \mathcal{V}^l$ is its word label. Word labels are used to determine whether two acoustic segments correspond to the same word, and to retrieve the phone sequence $\mathbf{p}_i^l \in \mathcal{P}^{L_i^l}$ for each word w_i^l from a lexicon where L_i^l is the phone sequence length. The acoustic embedding model f is a stacked bidirectional gated recurrent unit (BGRU) network [18].[1] The embedding model g of the written view consists of a learned phonetic embedding layer followed by a single-layer BGRU network. For each view, we concatenate the last time step outputs of the top layer in both directions of the BGRU as the fixed-dimensional embedding. We train f and g jointly to minimize the following objective:

$$\sum_{l \in \mathcal{L}} \sum_{i=1}^{N^l} \left[m + d(f(\mathbf{X}_i^l), g(\mathbf{p}_i^l)) - \min_{w^\star \in \mathcal{V}^l / w_i^l} d(f(\mathbf{X}_i^l), g(\mathbf{p}^\star)) \right]_+ + \left[m + d(g(\mathbf{p}_i^l), f(\mathbf{X}_i^l)) - \min_{\substack{\mathbf{X}^\star \in \\ \{\mathbf{X} | w \in \mathcal{V}^l / w_i^l\}}} d(g(\mathbf{p}_i^l), f(\mathbf{X}^\star)) \right]_+ \quad (1)$$

where m is a margin hyperparameter, d denotes cosine distance $d(u, v) = 1 - \frac{u \cdot v}{\|u\| \|v\|}$, $\mathcal{L}$ is the set of training languages, N^l is the number of training pairs for language $l \in \mathcal{L}$, and $\mathbf{p}^\star$ is the phonetic pronunciation of word $w^\star$. Intuitively, Equation 1 encourages embedding acoustic segments close to embeddings of their corresponding word labels and far from embeddings of other word labels, while implicitly separating acoustic segment embeddings of dissimilar words. For efficiency, rather than finding the most offensive example across the dataset (i.e. $w^\star \in \mathcal{V}^l / w_i$), we use the root mean squared distance over the K most offending examples within the mini-batch (i.e., $\sqrt{\frac{1}{K} \sum_{k=1}^{K} d(u, v)^2}$ where u, v represent the arguments of d above). Each mini-batch contains data from only one language.

2.2. Distinctive feature-based supervision

To make the best use of multilingual data, we would like to share as much information as possible across languages. The X-SAMPA phone set improves over written characters, but is far from perfect, as approximately 60% of phones in our 255-phone set appear in only one of the twelve languages used in our experiments. As a result, input embeddings for unseen phones are poorly (or not) learned. To address this issue, we investigate using distinctive features (DFs), such as manner and place features, rather than phones.[2] While many phones are language-specific, all feature values we consider are used in multiple languages with the exception of click features in Zulu and tone features in Cantonese and Lithuanian.

In this approach, we map each phone p_j to a vector of its distinctive features $\boldsymbol{\phi}_j$ containing 1 in dimensions corresponding to features that are "on" and 0 for features that are "off" (see Figure 2). We then pass this sequence of binary vectors through a linear layer which outputs vectors with the same dimensionality as the phone embeddings in the phone-based model. This step can be viewed as computing phone embeddings as a sum of distinctive feature embeddings. After this embedding step, the rest of the model is identical to the phone-based model.

3. Experimental Setup

We use conversational data from 12 languages: English data (10 hrs) from a subset of Switchboard [19], and 11 languages

[1] While previous related work has used LSTMs [15, 16], in initial experiments on English we found better performance with GRUs.

[2] Our code, phone set, and feature set can be found at `https://github.com/Yushi-Hu/Multilingual-AWE`

from the IARPA Babel project [20]: Cantonese (31 hrs), Assamese (19 hrs), Bengali (21 hrs), Pashto (28 hrs), Turkish (32 hrs), Tagalog (30 hrs), Tamil (23 hrs), Zulu (23 hrs), Lithuanian (15 hrs), Guarani (15 hrs), and Igbo (12 hrs). The development and test sets are about 1-3 hours per language. We use Kaldi [21] (Babel recipe `s5d`) to compute input acoustic features, train HMM/GMM triphone models, and extract word alignments for the Babel languages. The acoustic features are 117-dimensional, consisting of 36-dimensional log-Mel spectra + 3-dimensional (Kaldi default) pitch features. Training, development, and test sets are constructed from non-overlapping conversation sides. Babel training sets include segments with duration 25–500 frames corresponding to words occurring ≥ 3 times in training; development and test sets include segments of 50–500 frames with no word frequency restrictions. English development and test sets are the same as in prior related work [1, 6, 7, 16, 22, 23]. The English training set contains the same conversation sides as in prior work; however, for consistency within multilingual experiments we use word segments satisfying the same duration and frequency restrictions as for the Babel language training sets, which does not exactly match the prior AWE work. For comparison with prior work on English (Table 1), we also train a separate set of models on the same training set as in prior work. For distinctive feature-based experiments, the features for a given phone are retrieved from the PHOIBLE database [24].

We consider the following experimental settings: **single** (train and test on the target language), **unseen** (train on the 11 non-target languages, and then test on the unseen target language), and **fine-tune** (train on the 11 non-target languages, then fine-tune and test on the target language). We vary the amount of training data for **single** and **fine-tune** experiments among 10min, 60min, and "all" (the entire training set for the target language). The **unseen** setting is a zero-resource setting. To tune hyperparameters in **unseen** experiments, an average evaluation score from the development sets of the 11 training languages is used such that the target language is not seen until test evaluation.

3.1. Evaluation

AWEs and AGWEs can be used for a variety of downstream tasks. Here we use a task-agnostic evaluation approach, similarly to prior work [1,6,7,16,22,25], including two "proxy" tasks: acoustic word discrimination and cross-view word discrimination. Acoustic word discrimination is the task of determining whether a pair of acoustic segments $(\mathbf{X}_i, \mathbf{X}_j)$ correspond to the same word, while cross-view word discrimination is the task of determining whether an acoustic segment and word label $(\mathbf{X}_i, w_j)$ correspond to the same word. In both cases, we compute the cosine distance for each pair of embeddings, and consider a pair a match if its distance falls below a threshold. Results are reported as average precision (AP), i.e. area under the precision-recall curve generated by varying the threshold. We refer to the performance measure as "acoustic AP" for acoustic word discrimination, and "cross-view AP" for the cross-view task . The acoustic word discrimination task was first introduced by [22] as a proxy task for query-by-example search, and has been successfully used (along with the cross-view task) as a tuning criterion when applying AWEs/AGWEs to downstream tasks [3, 4, 15].

3.2. Hyperparameters

Hyperparameters are tuned on the small Switchboard subset from prior work [1, 6, 7, 16, 23]. The same model architecture is used for all languages. The acoustic view model is a 4-layer BGRU (with 0.4 dropout rate between layers), while the written view model consists of an input embedding layer and a 1-layer BGRU. Both recurrent models use 512 hidden units per direction per layer and output 1024-dimensional embeddings. When encoding sequences of either phones or phonetic features in the written view model, the embedding layer maps input representations to 64-dimensional vectors as depicted in Figure 2. There are 255 X-SAMPA phones and 38 distinctive features. Some distinctive features can take on more than 2 values, so we represent each value of each distinctive feature separately, giving 101 learned feature embeddings.

During training, the margin m in Equation 1 is set to 0.4, and negative sampling uses $K = 20$. We perform mini-batch optimization with Adam [26], with batch size 256 and initial learning rate 0.0005. The learning rate is decayed by a factor of 10 if the cross-view AP on the development set(s) fails to improve over 5 epochs. Training stops when the learning rate drops below 10^{-8}. All experiments use the PyTorch toolkit [27].

Table 1: *Test set performance of several embedding approaches on the English acoustic and cross-view word discrimination tasks. The numbers reported are average precision (AP).*

Method	Acoustic	Cross-view
100-minute training set		
MFCCs + DTW [6]	0.21	
CAE + DTW [23]	0.47	
Phone posteriors + DTW [22]	0.50	
Siamese CNN [6]	0.55	
Supervised CAE-RNN [9]	0.58	
Siamese LSTM [7]	0.67	
Multi-view LSTM [16] [3]	0.81	
Our multi-view GRU (chars)	0.81	0.71
Our multi-view GRU (phones)	**0.84**	**0.77**
Our multi-view GRU (features)	**0.83**	**0.76**
10-hour training set		
Our multi-view GRU (phones)	**0.88**	**0.81**
Our multi-view GRU (features)	**0.87**	**0.81**
135-hour training set		
Our multi-view GRU (phones)	**0.89**	**0.86**
Our multi-view GRU (features)	**0.89**	**0.86**

4. Results

4.1. Comparison with prior work on English

A number of previously reported results on acoustic word embeddings have used a particular subset of Switchboard for training [1, 6, 7, 16, 23]. To compare with this work, we also train using this same subset. We find (Table 1) that our models outperform all previous methods, including (our implementation of) CAE-RNN from recent work on multilingual AWEs [9].

We also find that representing the written word as a phone sequence improves over the character-based input representation used in prior work. Based on these results, for the remaining experiments we use our multi-view GRU-based models with phonetic representations of written words. Between the two phonetic representations (phone-based and feature-based), results are almost identical, with a slight edge for the phone-based representation. However, for multilingual experiments, we will largely use the feature-based representation as it allows us to embed previously unseen phones, and improves multilingual

[3] [16] calculates cross-view AP differently and is not comparable.

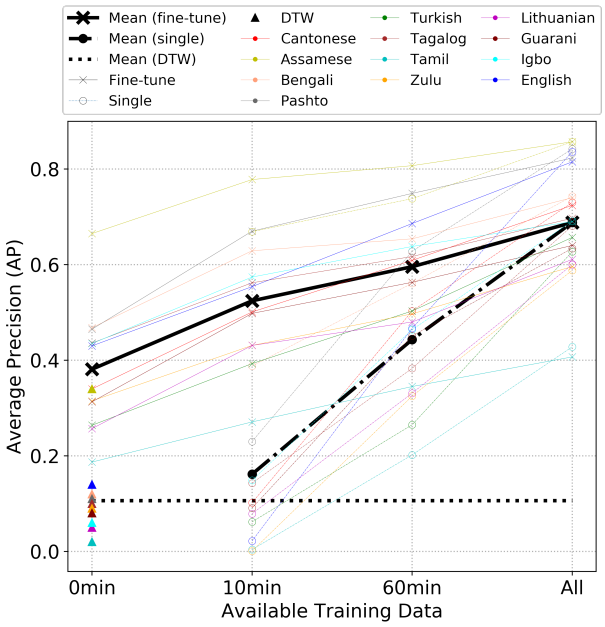

Figure 3: *Test set acoustic AP for models trained with varying amounts of target language data, supervised with distinctive features. Note that the x-axis is not linear. "0 min" refers to the* **unseen** *training setting, i.e. the* **fine-tune** *setting with no target language training data. "All" refers to all training data available for the target language. Average acoustic AP across languages is given by the thick black lines; language-specific results are given by the thin colored lines.*

performance (see Section 4.3).

Finally, we compare results of our models trained on different sizes of training set. We find that acoustic AP plateaus by around a 10-hour training set, with the results of a model trained on only 100 minutes being not far behind.

4.2. Evaluation of multilingual acoustic word embeddings

Figure 3 gives our main acoustic AP results for distinctive feature-based models across the 12 languages in the three training settings. (In terms of acoustic AP, there is little difference between the phone-based and distinctive feature-based models.) These results indicate that, when resources are limited in the target language, multilingual pre-training offers clear benefits. Fine-tuning a multilingual model on 10 minutes of target language data can outperform training on 60 minutes from the target language alone. Furthermore, if 60 minutes of data is available in the target language, multilingual pre-training cuts the performance gap between training on just that 60 minutes alone and the full 10-30 hour training set by more than half. The **unseen** model (equivalent to **fine-tune** with 0 minutes of target language data) is a zero-resource model with respect to the target language since it is trained on the other 11 languages and never sees examples from the target language until test evaluation. On average, our **unseen** models significantly outperform the unsupervised DTW baselines—confirming results of other recent work in the zero-resource setting [9]—as well as the **single**-10min models, and perform similarly to the **single**-60min models.

4.3. Phonetic vs. distinctive feature supervision

While acoustic AP is largely unaffected by the choice of phone vs. feature supervision, in terms of cross-view AP, Figure 5 shows that **unseen** models typically benefit from using distinctive features over phones.

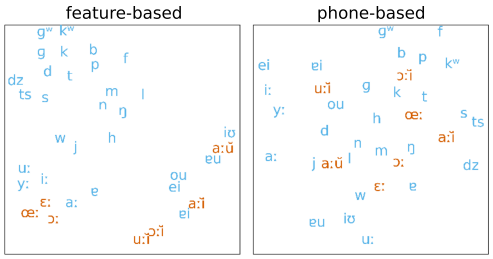

Figure 4: *t-SNE [13] visualizations of Cantonese phone embeddings from* ***unseen*** *models supervised with distinctive features (left) and phones (right). Blue phones appear in other languages; orange phones are unique to Cantonese.*

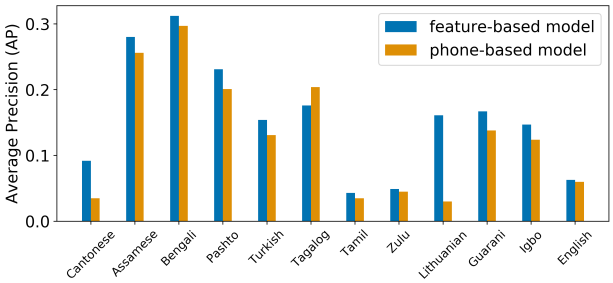

Figure 5: *Test set cross-view AP in the* **unseen** *setting.*

The two languages with the largest improvement from distinctive features are Cantonese and Lithuanian. The Cantonese data includes a large number of diphthongs that are unseen in other languages, so their embeddings cannot be learned in the phone-based model, but the features of those diphthongs are shared with phones in other languages. In the Lithuanian data, vowels are paired with their tones, making these phones unique to Lithuanian and again making it impossible to learn the vowel embeddings from other languages using phone-based supervision. All distinctive features, except for the Lithuanian-specific tone features themselves, are shared with other languages, making it easier for the feature-based model to learn good embeddings.

In addition to generating embeddings of spoken and written words, our written embedding models also include a learned embedding for each phone. Figure 4 visualizes Cantonese phone embeddings taken from a model trained on the other 11 languages. The model trained using distinctive features is able to infer reasonable embeddings for the phones that are unique to Cantonese and unseen in other languages, placing them near similar phones in the embedding space, while the phone-based model is forced to use (random) initial embeddings.

5. Conclusions

We have presented an approach for jointly learning acoustic and written word embeddings for low-resource languages, trained on data from multiple languages. Multilingual pre-training offers significant benefits when we have only a small amount of (or no) labeled training data for the target language. By using distinctive features to encode the pronunciations of written words, we improve cross-lingual transfer by allowing phones unseen during training to share information with similar phones seen in the training set. Future work will apply our learned embeddings to downstream tasks and expand to a larger language set.

6. Acknowledgements

This research was funded by NSF award IIS-1816627, and by an AWS Machine Learning Research Award. We thank Herman Kamper for helpful feedback.

7. References

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

[2] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[3] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *Proc. Interspeech*, 2017.

[4] 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," in *Proc. Interspeech*, 2018.

[5] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[6] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

[8] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[9] H. Kamper, Y. Matusevych, and S. Goldwater, "Multilingual acoustic word embedding models for processing zero-resource languages," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[10] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end ASR-free keyword search from speech," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[11] Y.-A. Chung, C.-C. Wu, C.-H. Shen, and H.-Y. Lee, "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks," in *Proc. Interspeech*, 2016.

[12] N. Holzenberger, M. Du, J. Karadayi, R. Riad, and E. Dupoux, "Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments," in *Proc. Interspeech*, 2018.

[13] L. J. P. van der Maaten and G. E. Hinton, "Visualizing data using *t*-SNE," *Journal of Machine Learing Research (JMLR)*, 2008.

[14] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[15] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[16] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *"Proc. Int. Conf. on Learning Representations (ICLR)"*, 2017.

[17] J. Wells, "Computer-coding the IPA: a proposed extension of SAMPA," 1995.

[18] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[19] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 1992.

[20] "IARPA Babel language pack: IARPA-babel101b-v0.4c, IARPA-babel102b-v0.5a, IARPA-babel103b-v0.4b, IARPA-babel104b-v0.4by, IARPA-babel105b-v0.5, IARPA-babel106-v0.2g, IARPA-babel204b-v1.1b, IARPA-babel206b-v0.1e, IARPA-babel304b-v1.0b, IARPA-babel305b-v1.0c, IARPA-babel306b-v2.0c."

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[22] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. Interspeech*, 2011.

[23] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[24] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[25] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *"Proc. Int. Conf. on Learning Representations (ICLR)"*, 2015.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.