# Multilingual Jointly Trained Acoustic and Written Word Embeddings

**Yushi Hu**
University of Chicago
hys98@uchicago.edu

**Shane Settle, Karen Livescu**
TTI-Chicago
{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 pretrained 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 while varying the amount of target language training data, and find significant benefits to the proposed multilingual approach.

## 1 Introduction

Acoustic word embeddings (AWEs) are vector representations of spoken word segments of arbitrary duration (Levin et al., 2013). AWEs are an attractive tool in tasks involving reasoning about whole word segments, as they provide a compact representation that can be used to efficiently measure similarity between segments. For example, AWEs have been used to speed up and improve query-by-example search (Levin et al., 2015; Settle et al., 2017; Yuan et al., 2018), unsupervised segmentation and spoken term discovery (Kamper et al., 2016), spoken term detection (Audhkhasi et al., 2017), and whole-word speech recognition (Bengio and Heigold, 2014; Settle et al., 2019).

Many approaches have been explored for constructing and learning AWEs, including template-based techniques (Levin et al., 2013) and neural network-based models (Settle and Livescu, 2016; Kamper, 2018), but prior work in this area largely focuses on English. In this work, we study the learning of AWEs/AGWEs for multiple low-resource languages. Recent related work (Kamper et al., 2020) has begun to explore multilingual AWEs, specifically for zero-resource languages. 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.[1]

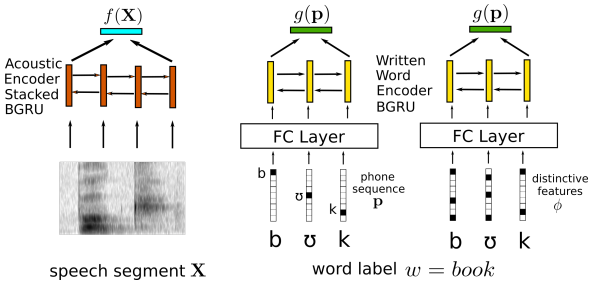


Figure 1: Acoustic word embedding (AWE) model $f$ and two acoustically grounded word embedding (AGWE) models $g$, corresponding to either phone or distinctive feature sequence input.

## 2 Embedding models

An 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.

Our approach is based on prior work on AWE/AGWE learning using a multi-view contrastive loss (He et al., 2017; Settle et al., 2019),

[1] This abstract is based on Hu et al. (2020). Our code, phone set, and feature set can be found at github.com/Yushi-Hu/Multilingual-AWE

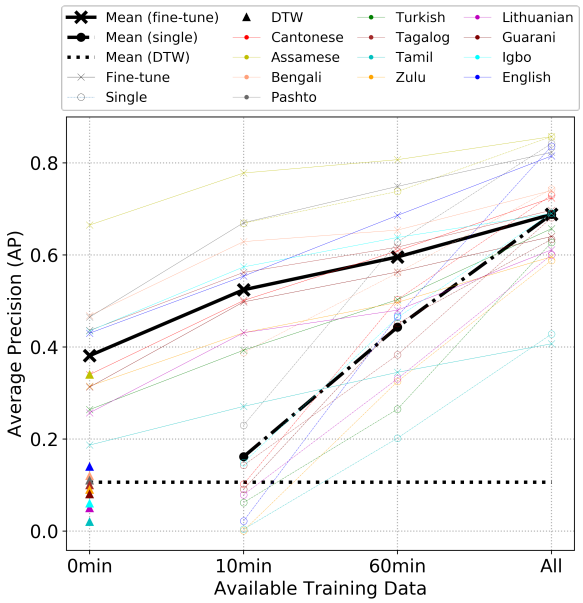


Figure 2: Test acoustic AP for models trained with varying amounts of target language data, supervised with distinctive features.

which jointly learns models for an acoustic view ($f$) and a written view ($g$) (Figure 1). The input to the acoustic embedding model $f$ is a variable-length spoken word segment. In prior work (He et al., 2017), the input to $g$ is a character sequence, but for multilingual training, we use phonetic sequences. However, approximately 60% of phones appear in only one of the 12 languages, so we also investigate using distinctive features (DFs), such as manner and place features, rather than phones.

## 3 Experimental Setup

We use conversational data covering 12 languages from Switchboard (Godfrey et al., 1992) and the IARPA Babel project (Babel). We consider the following settings: **single** (train/test on the target language), **unseen** (train on non-target languages, test on the unseen target language), and **fine-tune** (train on non-target languages, fine-tune/test on the target language). Training data is varied for **single** and **fine-tune** experiments among 10min, 60min, and "all" (full target language training set).

To evaluate our models, we use task-agnostic evaluation approaches similar to prior work (Carlin et al., 2011; He et al., 2017), including acoustic word discrimination and cross-view word discrimination. Results are reported as average precision (AP) where "acoustic AP" and "cross-view AP" describe acoustic and cross-view word discrimination performance, respectively.

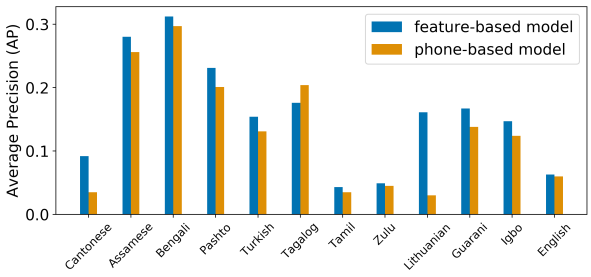


Figure 3: Test set cross-view AP in the **unseen** setting.

## 4 Results

Figure 2 gives our main acoustic AP results for distinctive feature-based models across the 12 languages in the three training settings. 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. On average, our **unseen** models significantly outperform the unsupervised DTW baselines—confirming results of other recent work in the zero-resource setting (Kamper et al., 2020)—as well as the **single**-10min models, and perform similarly to the **single**-60min models.

Figure 3 shows that cross-view AP of **unseen** models typically benefits from using distinctive features over phones. 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.

## 5 Conclusion

Multilingual pre-training improves the quality of of acoustic and acoustically grounded word embeddings when we have only a small amount of (or no) labeled training data for the target language, and phonological feature-based training allows for better transfer to languages with rare phones. In ongoing work we are applying the learned embeddings to improve multilingual query-by-example search.

## References


Kartik Audhkhasi, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, and Brian Kingsbury. 2017. End-to-end ASR-free keyword search from speech. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1351–1359.

IARPA Babel. 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.

Samy Bengio and Georg Heigold. 2014. Word embeddings for speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*.

Michael A Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. 2011. Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*.

Wanjia He, Weiran Wang, and Karen Livescu. 2017. Multi-view recurrent neural acoustic word embeddings. In *"Proc. Int. Conf. on Learning Representations (ICLR)"*.

Yushi Hu, Shane Settle, and Karen Livescu. 2020. Multilingual jointly trained acoustic and written word embeddings. In *Proc. Interspeech*.

Herman Kamper. 2018. 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)*.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2020. Multilingual acoustic word embedding models for processing zero-resource languages. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*.

Keith Levin, Katherine Henry, Aren Jansen, and Karen Livescu. 2013. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Keith Levin, Aren Jansen, and Benjamin Van Durme. 2015. Segmental acoustic indexing for zero resource keyword search. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*.

Shane Settle, Kartik Audhkhasi, Karen Livescu, and Michael Picheny. 2019. Acoustically grounded word embeddings for improved acoustics-to-word speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*.

Shane Settle, Keith Levin, Herman Kamper, and Karen Livescu. 2017. Query-by-example search with discriminative neural acoustic word embeddings. In *Proc. Interspeech*.

Shane Settle and Karen Livescu. 2016. Discriminative acoustic word embeddings: Recurrent neural network-based approaches. In *Proc. IEEE Workshop on Spoken Language Technology (SLT)*.

Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, and Haizhou Li. 2018. Learning acoustic word embeddings with temporal context for query-by-example speech search. In *Proc. Interspeech*.