Multilingual Jointly Trained Acoustic and Written Word Embeddings

Yushi Hu, Shane Settle, Karen Livescu

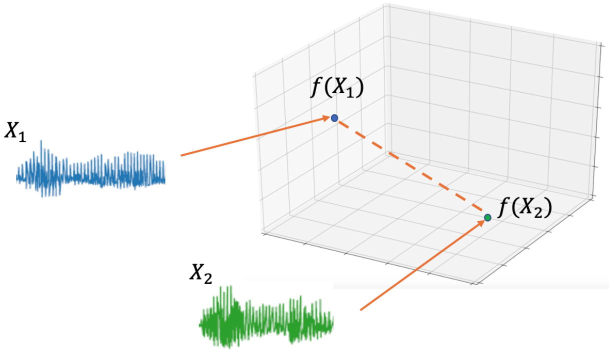
Interspeech 2020

THE UNIVERSITY OF CHICAGO

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

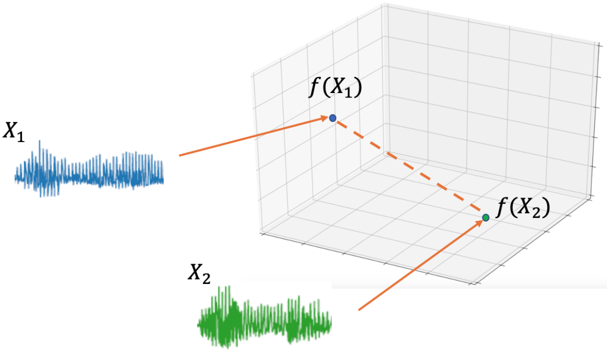
Acoustic word embeddings (AWE)

- An acoustic word embedding (AWE) model f maps a variable-length spoken word segment to a vector.
- AWEs can improve query-by-example search [Settle+ 2017], spoken term discovery [Kamper+ 2016]



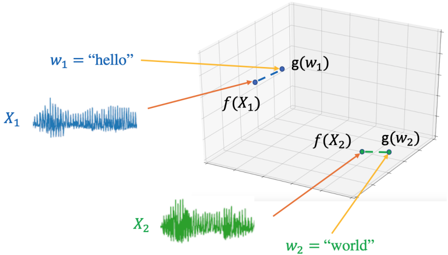
What makes a good acoustic word embedding?

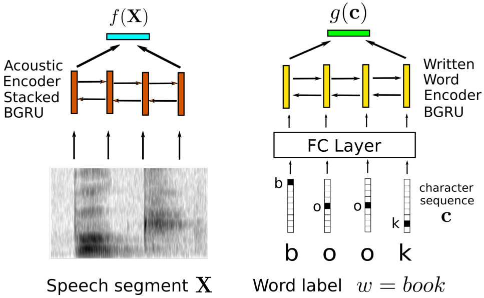
- **Same-word** signals should have similar vectors: factor out speaker, acoustic environment, ...
- Signals from **different words** should be embedded farther apart



Acoustically grounded word embeddings (AGWE)

Given an (acoustic, written) word pair ($\mathbf{X}$, w), jointly train **AWE** function $f(\cdot)$ and **AGWE** function $g(\cdot)$ to learn mappings into a shared space [He+ 2017]



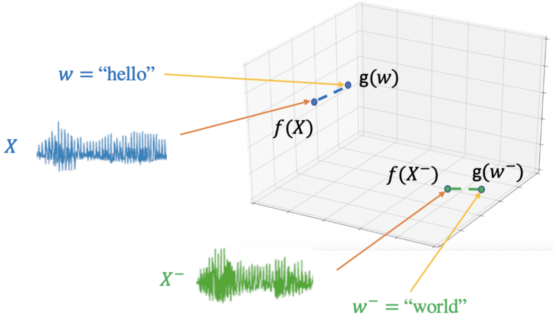


Jointly trained acoustic and written word embeddings

- Contrastive loss [He+ 2017] (we use a modified form)

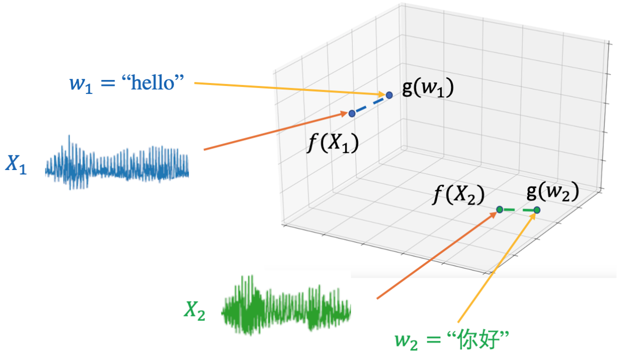
$$max \left\{0, m + d_{cos}(f(X), g(w)) - \min_{w^- \neq w} d_{cos}(f(X), g(w^-))\right\}$$

- Can improve whole-word speech recognition via pre-training [Settle+ 2019]



Multilingual jointly trained acoustic and written word embeddings

- **Goal:** Extend the application of AWEs/AGWEs to many languages
- **Approach:** Map spoken word signals and written words from multiple languages to embeddings in a shared space

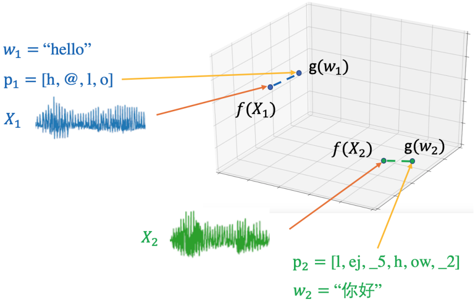


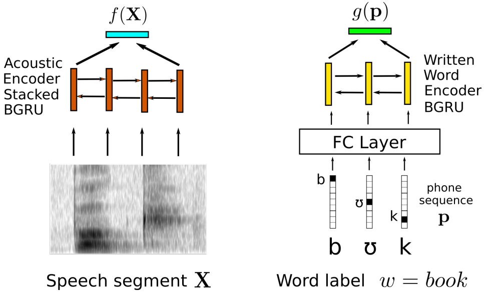
Problem: Prior work on English takes character sequences as the input to g. Our multilingual models need to deal with widely differing written systems.

Using phones as input

Phone sequence as input to the AGWE model g

- Cross-lingual information sharing
- Ability to embed words from unseen languages



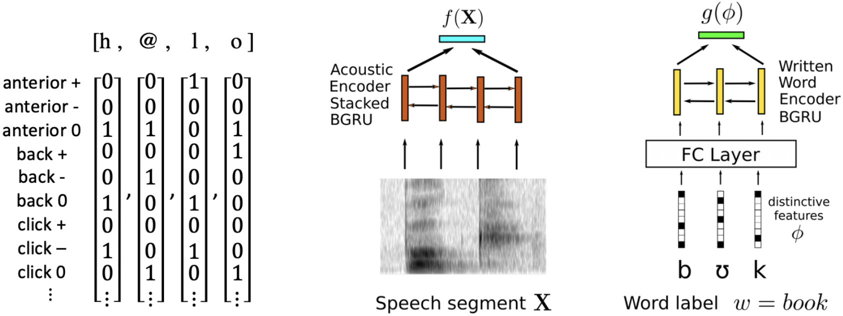


Using distinctive features as input

- 60% of phones in our 255-phone set appear in only one of the 12 languages. Unseen phones are not learned.
- Using distinctive features as input allows almost 100% coverage.

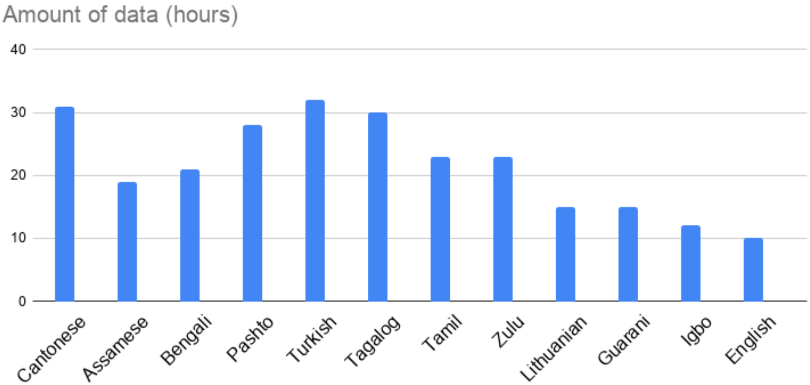
Features of "a"

ATR -
anterior 0
approximant +
back –
click 0
consonantal -
...



Languages used in experiments

11 Babel languages + Switchboard English



Experimental setup

Data

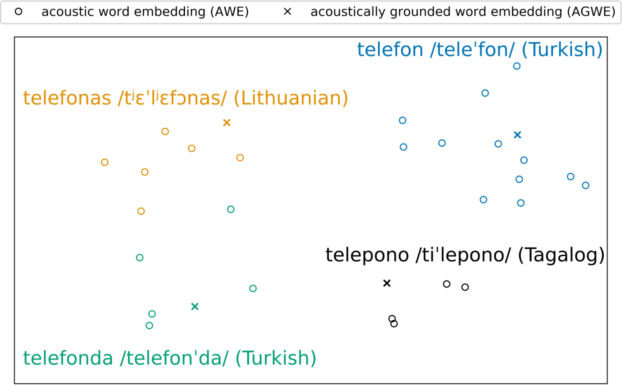
- 11 Babel languages + Switchboard English
- X-SAMPA phones
- Distinctive features from PHOIBLE database
- 36d standard log-Mel spectral features + 3d pitch features

Model

- Acoustic view: 4-BiGRU (512d) $\rightarrow$ 1024d embedding
- Written view: 64d phone/feature emb $\rightarrow$ 1-BiGRU (512d) $\rightarrow$ 1024d embedding

Visualization of learned embeddings

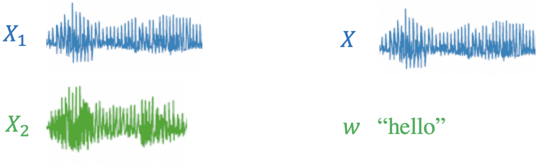
t-SNE visualization of learned acoustic word embeddings (AWE) and acoustically grounded word embeddings (AGWE)



Evaluation

- Tasks: **acoustic word discrimination** and **cross-view word discrimination**
- Compute the cosine distance between embedding vectors and consider a pair a match if its distance falls below a threshold.
- Metric: average precision (AP)

Acoustic word discrimination | **Cross-view word discrimination**



Do X_1 and X_2 correspond to the same word?

Is X an example of w ?

Comparison with prior work on English

Test set average precision (AP) on English word discrimination tasks

- Improves over prior work
- Phone sequence input improves over character-based input

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**

Comparison with prior work on English

Test set average precision (AP) on English word discrimination tasks

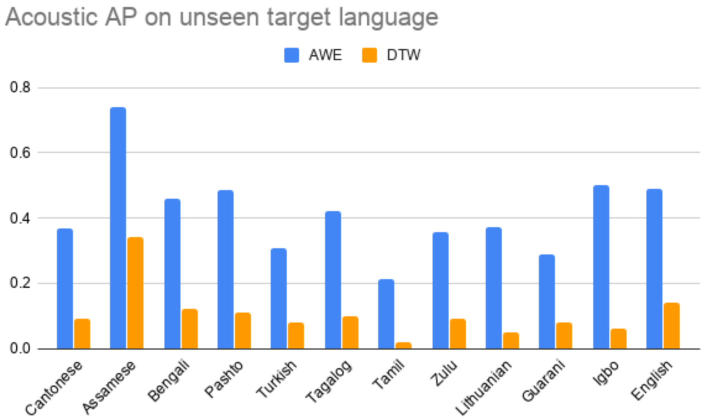
- Improves over prior work
- Phone sequence input improves over the character-based input representation
- Acoustic AP plateaus by around 10 hours of training data
- Phone-based and feature-based input get similar results on English

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**

Performance on unseen target language

Acoustic AP results for distinctive feature-based models on 12 languages

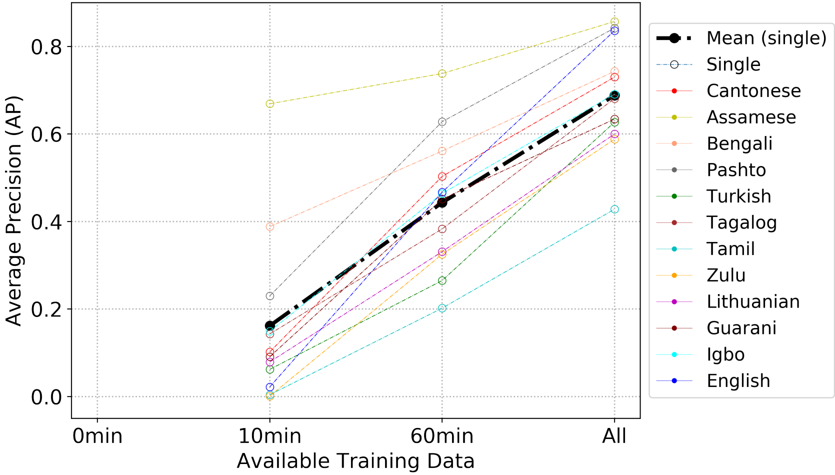
- Train on 11 non-target languages, then test on the unseen target language
- Zero-resource setting
- Our approach significantly outperfoms the unsupervised DTW baselines



Training on varying amounts of monolingual training data

Acoustic AP results for distinctive feature-based models on 12 languages

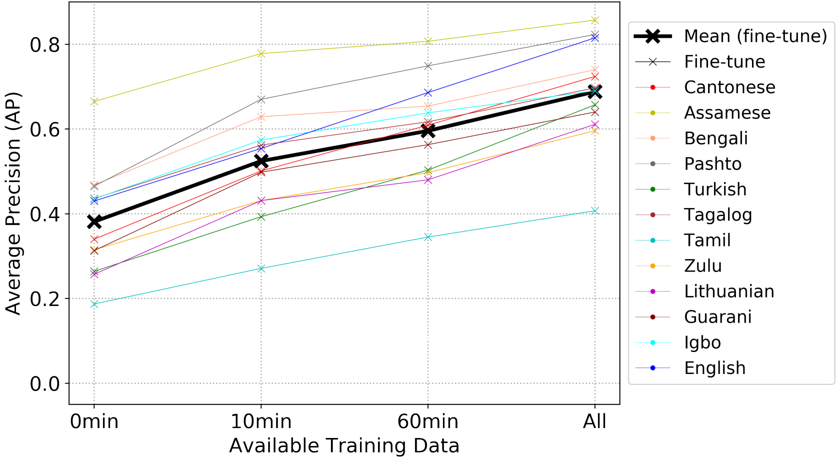
- Train and test on the target language



Multilingual pre-training + target language fine-tuning

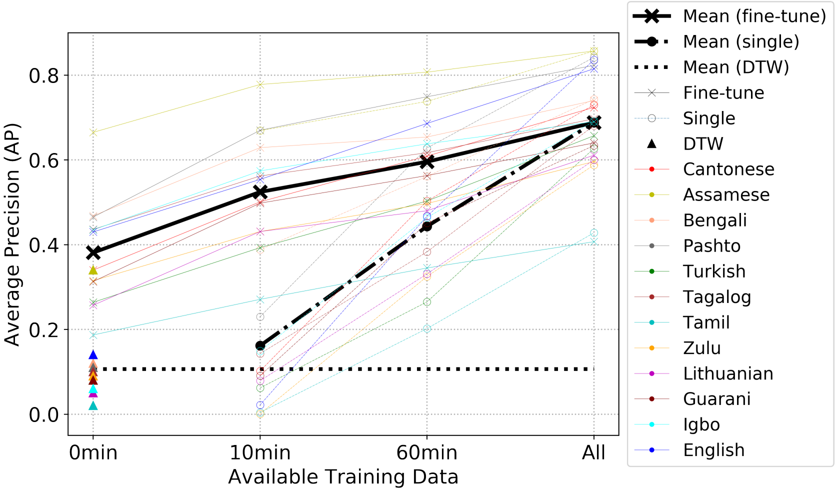
Acoustic AP results for distinctive feature-based models on 12 languages

- Train on 11 non-target languages, then fine-tune and test on the target language



Benefits of multilingual pre-training

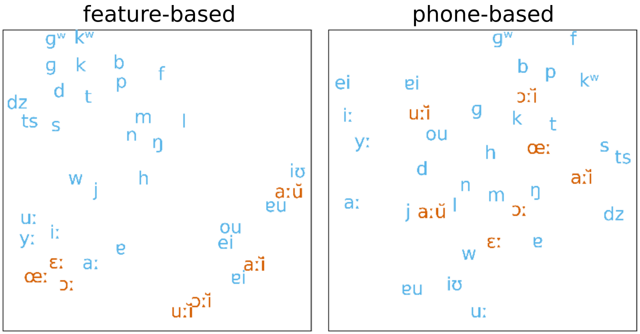
Multilingual pre-training offers clear benefits when resources are limited in the target language



Phonetic vs. distinctive feature supervision

Cantonese phone embeddings taken from the model trained on the other 11 languages

- Feature-based model places Cantonese-specific phones near similar phones.
- Phone-based model is forced to use (random) initial embeddings.

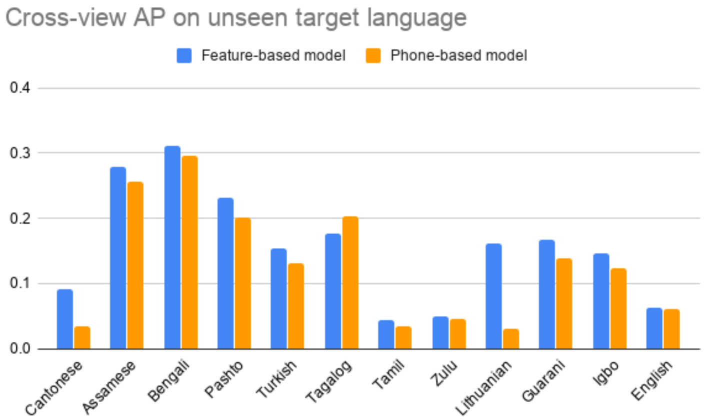


Blue phones appear in other languages; orange phones are unique to Cantonese

Phonetic vs. distinctive feature supervision

Cross-view AP in zero-resource setting (train on 11 non-target languages and test on the unseen target language)

- Models benefit from using distinctive features over phones



Related work

H. Kamper, Y. Matusevych, and S. Goldwater, "Multilingual acoustic word embedding models for processing zero-resource languages," in ICASSP 2020.

- We add new results for varying amount of data.
- We learn not only AWE but also AGWE, thus widening the range of tasks to which our models apply.

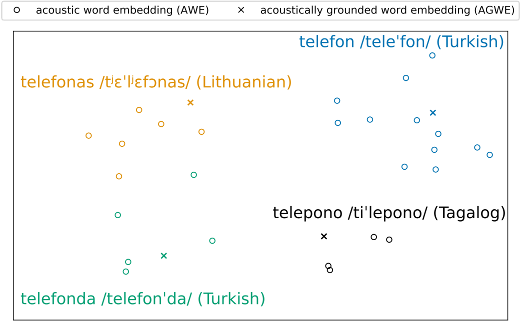
A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," arXiv:2006.13979, 2020.

- Unsupervised cross-lingual pre-training also improves frame representations.

Conclusion and future work

An approach for jointly learning acoustic and written word embeddings for low-resource languages, trained on multiple languages

- Multilingual pre-training offers clear benefits.
- Distinctive features improve cross-lingual transfer.



New work: Our multilingual AWEs work well in query-by-example search.

Future work: Application to keyword search and multilingual ASR.