Acoustic Span Embeddings For Multilingual Query-by-Example Search

Yushi Hu, Shane Settle, Karen Livescu

IEEE SLT 2021

THE UNIVERSITY OF CHICAGO

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Introduction

Query-by-Example speech search (QbE): matching spoken queries to utterances within a search collection

Prior work

Dynamic time warping (DTW) based approach

- Rely on the quality of the frame representations (phone posterior, bottleneck features, ...)
- O(NM) time complexity. N and M are segment lengths.
- Need special modification to work on approximate query matches. Best systems on benchmark QbE tasks often fuse many systems together.

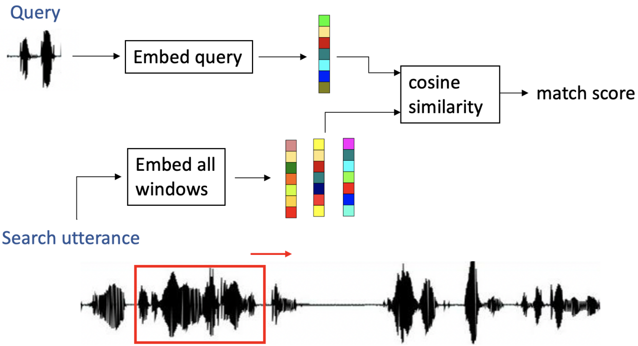
Embedding-based approach

- Improve speed and performance
- Focus on English data and on single-word queries

Motivation

Apply embedding-based QbE to more general settings

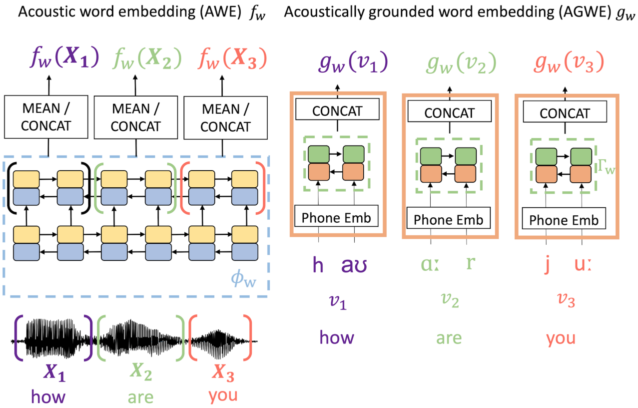
- Arbitrary length queries
- Multiple zero-resource target languages



Contribution 1: Embedding-based QbE on multiple unseen languages

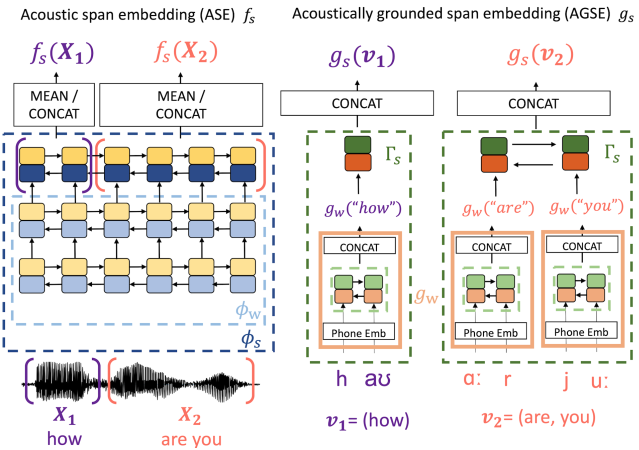
Embedding-based QbE can be effectively applied to multiple unseen languages by using embeddings learned on languages with available data.

- Multilingual jointly trained acoustic and written word embeddings [Hu+ 2020] trained on 12 languages
- We apply this idea to a QbE task with 6 unseen languages



Contribution 2: Acoustic span embeddings (ASE)

- Prior works on embedding-based QbE mainly use acoustic word embeddings and focus on single-word queries
- Queries may contain arbitrary numbers of words
- We extend the idea of acoustic word embeddings to multi-word spans



Evaluation result: Our QbE system is fast, accurate, and simple

QUESST 2015 QbE search task

- 6 low-resource languages
- Challenging acoustic conditions
- Exact and approximate match query settings

Our approach

- Outperforms all prior work on this benchmark
- Much faster than (naive) DTW-based search
- Single ASE model works well in both settings

Method	**# systems**	$minC_{nxe} \downarrow$
BNF+DTW [11]	36	0.778
BNF+DTW [26]	66	0.757
Best prior fusion [7]	4	0.723
ASE(mean)	1	0.706
ASE(mean + concat)	2	**0.670**

More details

Multilingual embedding-based QbE

- Acoustic word embedding (AWE)
- Acoustic span embedding (ASE)
- Search component

Experimental setup

- Embedding model
- QbE system

Evaluation

- QUESST 2015 QbE task
- Evaluation metrics

Results

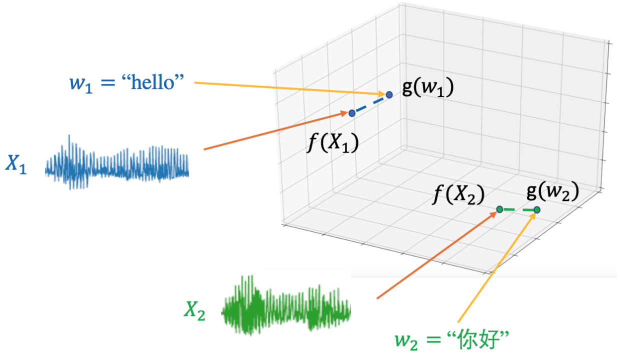
- Comparison with prior work
- Query sub-tasks
- Run time

Conclusion

Multilingual jointly trained acoustic and written word embeddings

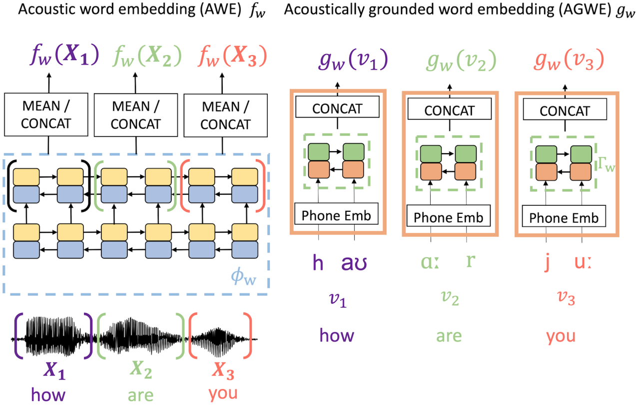
Map spoken word signals and written words from multiple languages to embeddings in a shared space [Hu+ 2020]

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



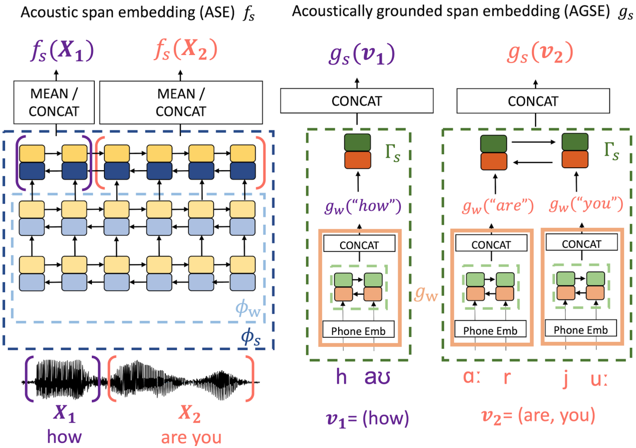
Contextual acoustic word embeddings (AWE)

- **Approach**: jointly train *AWE* function $f_w(\cdot)$ and *AGWE* function $g_w(\cdot)$
- **Architecture**: BiGRU encoder + pooling function
- **Extension**: embed word segments in context
  - Improve QbE performance
  - Help efficiently embed the search collection



Acoustic span embeddings (ASE)

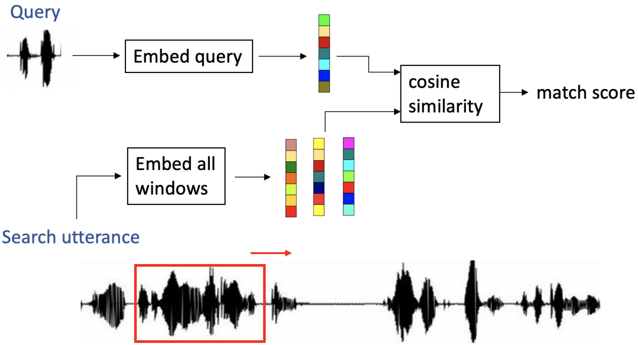
- **Goal**: better model spans of multiple words in queries and search utterances
- **Changed training objective**: contrastive loss over multi-word spans, instead of single word
- Trained on 12 languages. We use only the acoustic-view model f_s in QbE system



Embedding-based QbE system

Given a pre-trained embedding model

- Build an index of utterances in the search collection by embedding all possible segments (sliding window with several window sizes)
- Given an audio query, embed the query and compute a detection score for each utterance by the cosine similarity between the embedding vectors



Experimental setup

Embedding model

Training data

- 11 Babel languages + Switchboard English
- X-SAMPA phones
- 36d standard log-Mel spectral features + 3d pitch features
- SpecAugment

Model

- Acoustic view: 6-BiGRU (256d) $\rightarrow$ 512d embedding
- Written view: 1-BiGRU (256d) $\rightarrow$ 1-BiGRU (256d) $\rightarrow$ 512d embedding

QbE system

- Window sizes $\{12, 15, 18, \ldots, 30, 36, 42, 48, \ldots, 120\}$
- For query (length l_q), compare with all windowed segments with length between $\frac{2}{3}l_q$ and $\frac{4}{3}l_q$.

QUESST 2015 query-by-example search task

6 languages: Albanian, Czech, Mandarin, Portuguese, Romanian, and Slovak

Size: 18 hours search collection. 445 development queries and 447 test queries.

Three types of queries:

- T1: exact match
- T2: allowing word reordering and lexical variations
- T3: like T2, but conversational queries in context

Acoustic condition: artificially added noise and reverberation

Evaluation metrics

Normalized cross entropy (Cnxe)

- Ratio between the cross entropy of the QbE system output scores and random scoring
- Ranges from 0 to 1. The smaller, the better

Term weighted value (TWV)

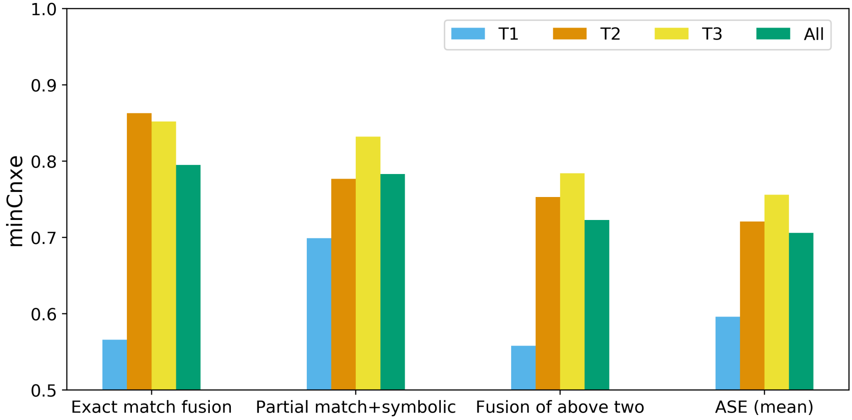
- Computed by miss rate and false alarm rate: $1 - (P_{miss}(\theta) + \beta P_{fa}(\theta))$
- Ranges from $-\beta$ to 1. The bigger, the better

Results: Comparison with prior work

Table 1. QUESST 2015 performance on dev and eval sets measured by $minC_{nxe}$ and $maxTWV$. Training languages are separated into in- and out-of-domain. All SAD systems are based on phone recognizers.

Method	systems	languages in	languages out	labeled data[4] hours	SAD	Augmentation	$minC_{nxe} \downarrow$ / $maxTWV \uparrow$ dev	eval
Top prior results								
BNF+DTW [11]	36	2	4	384+	Yes	noise	0.778 / 0.234	0.787 / 0.206
BNF+DTW [26]	66	2	15	643+	Yes	noise + reverb	0.757 / 0.286	0.747 / 0.274
Exact match fusion [7]	2	0	2	423	Yes	noise + reverb	0.795 / 0.256	
Partial match + symbolic [7]	2	0	1	260	Yes	noise + reverb	0.783 / 0.231	
Fusion of above two [7]	4	0	2	423	Yes	noise + reverb	0.723 / 0.320	
Our systems								
AWE (concat)	1	0	12	664			0.845 / 0.084	
AWE (mean)	1	0	12	664			0.803 / 0.101	
AWE (mean)	1	0	12	664		SpecAugment	0.782 / 0.135	
ASE (concat)	1	0	12	664			0.753 / 0.193	
ASE (mean)	1	0	12	664			0.728 / 0.239	
ASE (mean)	1	0	12	664		SpecAugment	**0.706 / 0.255**	**0.692 / 0.246**
ASE (mean+concat)	2	0	12	664		SpecAugment	**0.670 / 0.323**	**0.658 / 0.298**

Dependence on query sub-task



ASE models are better at accommodating lexical variations and word reordering than DTW-based systems without sacrificing too much performance on exact matches

Run time

The average per-query run times of our implementations of ASE-based and DTW-based QbE search. Tested on a single thread of a CPU.

- Naive ASE is much faster than naive DTW
- Both ASE and DTW could be sped up with approximations (future work)

Table 2. Run times on the QUESST 2015 development set.

Method	# of comparisons (per query)	Run-time (s / query)
DTW on Filterbank features	600K	486
DTW on ASE hidden states	600K	847
ASE-based QbE	4M	**5**

Conclusion and future work

A simple embedding-based approach for multilingual query-by-example search

- Outperforms prior work on the QUESST 2015 QbE task, while also being much more efficient.
- Demonstrates that multilingual acoustic word embedding (AWE) models can be effective for query-by-example search on unseen target languages
- Extends embedding-based QbE to multi-word spans using acoustic span embeddings (ASE)

Future work: use both the acoustic and written view embedding models to search by either spoken or written query