Multilingual acoustic word embedding models for processing zero-resource languages

ICASSP 2020

Herman Kamper[1], Yevgen Matusevych[2], Sharon Goldwater[2]

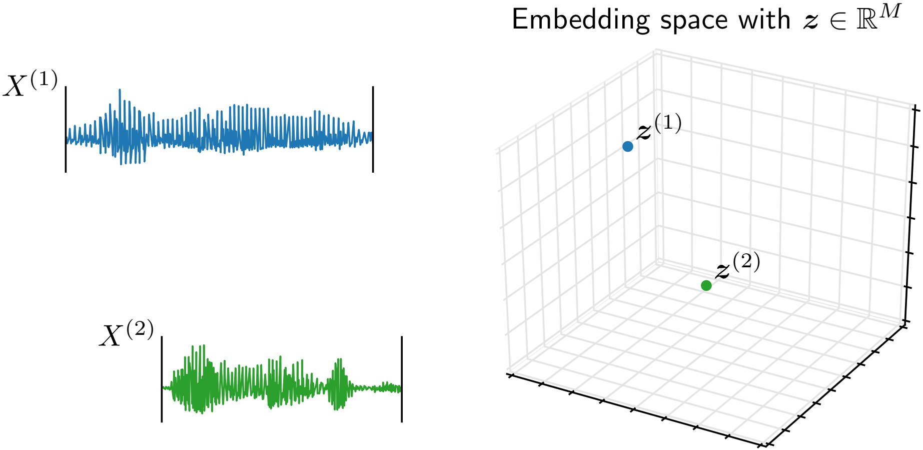
[1]Stellenbosch University, South Africa
[2]University of Edinburgh, UK

http://www.kamperh.com/

Background: Why acoustic word embeddings?

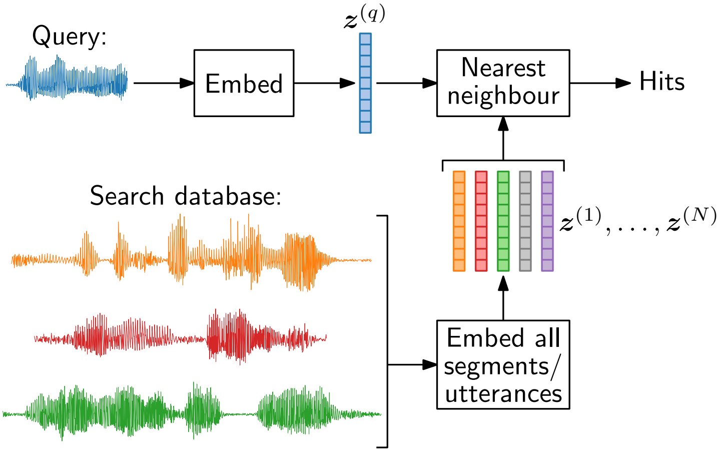
- Current speech recognition methods require large labelled data sets
- *Zero-resource speech processing* aims to develop methods that can discover linguistic structure from unlabelled speech [Dunbar et al., ASRU'17]
- Example applications: Unsupervised term discovery, query-by-example
- **Problem:** Need to compare speech segments of variable duration

Acoustic word embeddings



Embedding space with $\boldsymbol{z} \in \mathbb{R}^M$

Example application: Query-by-example search


Query:
Embed
$\boldsymbol{z}^{(q)}$
Nearest neighbour
Hits
Search database:
Embed all segments/ utterances
$\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N)}$

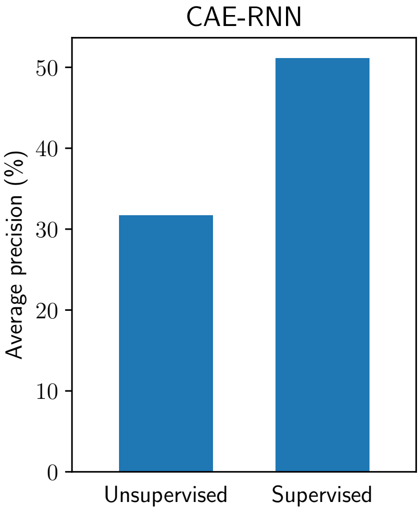
[Levin et al., ICASSP'15]

Supervised and unsupervised acoustic embeddings

- Growing body of work on acoustic word embeddings
- Supervised and unsupervised methods
- Unsupervised methods can be applied in zero-resource settings
- But there is still a large performance gap

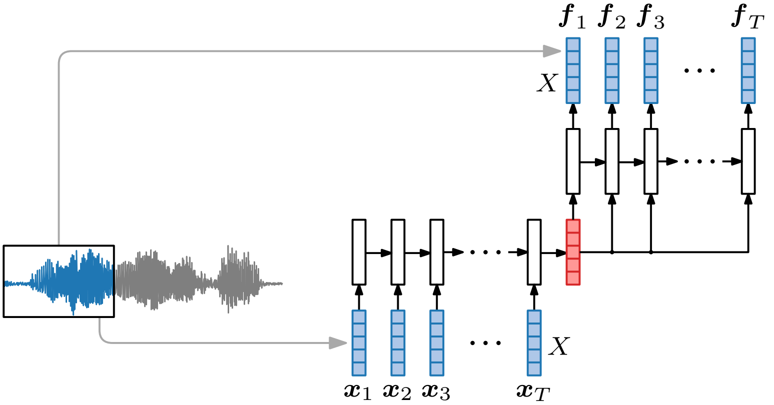
Supervised and unsupervised acoustic embeddings

- Growing body of work on acoustic word embeddings
- Supervised and unsupervised methods
- Unsupervised methods can be applied in zero-resource settings
- But there is still a large performance gap



[Kamper, ICASSP'19]

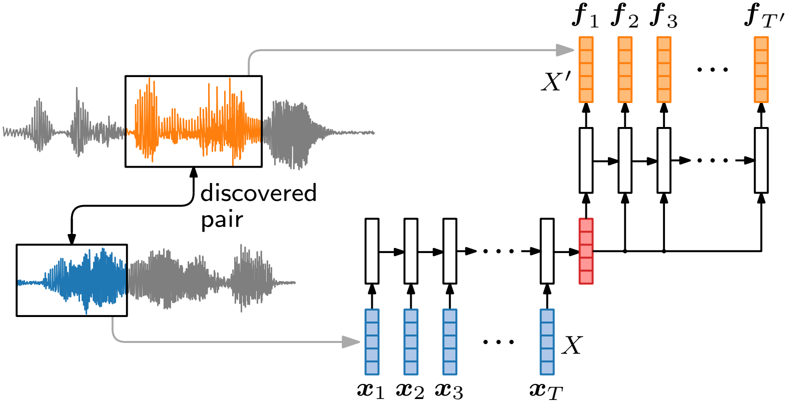
Unsupervised monolingual acoustic word embeddings



$\boldsymbol{f}_1$ $\boldsymbol{f}_2$ $\boldsymbol{f}_3$ $\boldsymbol{f}_T$
X
X
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_T$

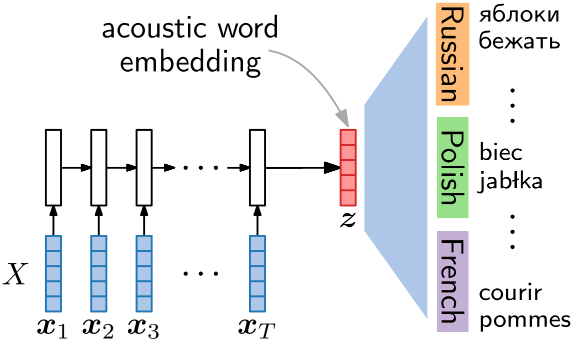
[Chung et al., Interspeech'16; Kamper, ICASSP'19]

Unsupervised monolingual acoustic word embeddings


$\boldsymbol{f}_1$ $\boldsymbol{f}_2$ $\boldsymbol{f}_3$ $\boldsymbol{f}_{T'}$
X'
discovered pair
X
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_T$

[Chung et al., Interspeech'16; Kamper, ICASSP'19]

Supervised multilingual acoustic word embeddings



Experimental setup

- **Training data:** Six well-resourced languages
  Czech (CS), French (FR), Polish (PL), Portuguese (PT), Russian (RU), Thai (TH)

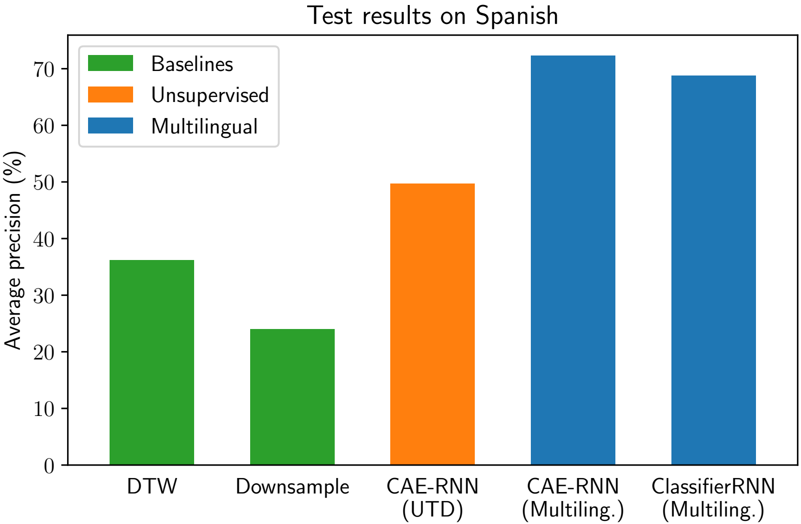
- **Test data:** Six languages treated as zero-resource
  Spanish (ES), Hausa (HA), Croatian (HR), Swedish (SV), Turkish (TR), Mandarin (ZH)

- **Evaluation:** Same-different isolated word discrimination

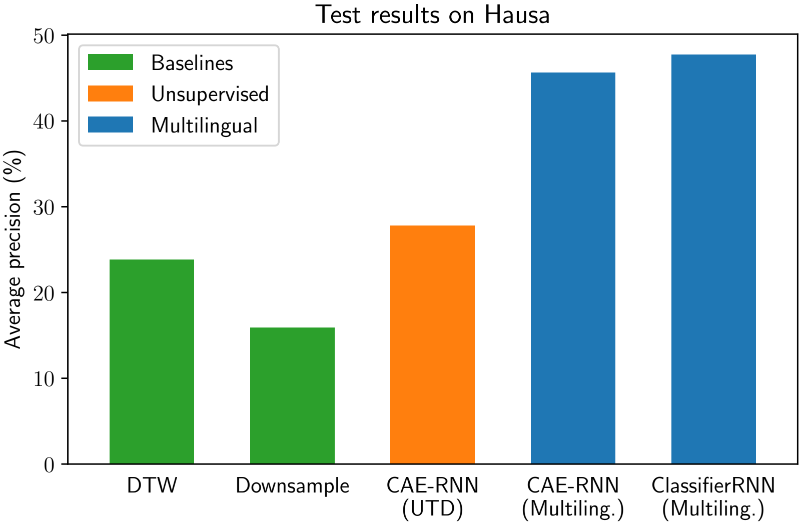
- **Embeddings:** $M = 130$ for all models

- **Baselines:**
  - — Downsampling: 10 equally spaced MFCCs flattened
  - — Dynamic time warping (DTW) alignment cost between test segments

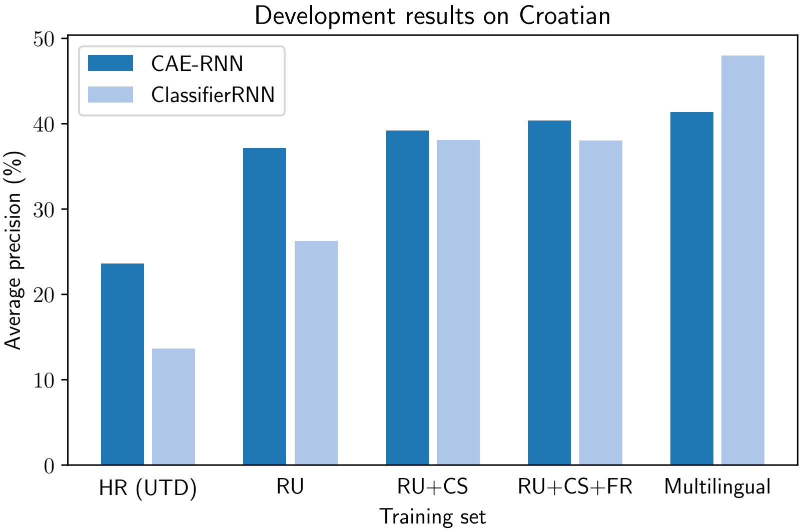
1. Is multilingual supervised > monolingual unsupervised?



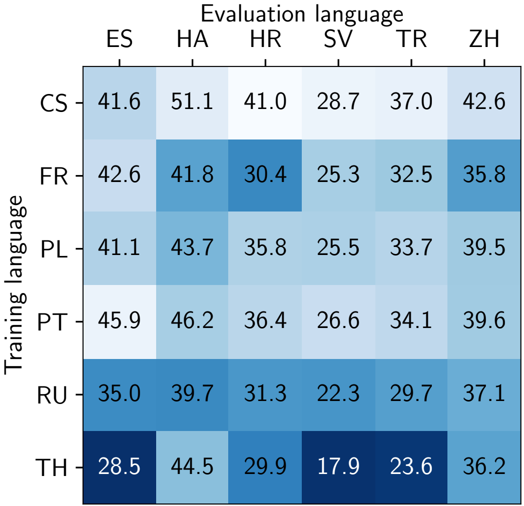
1. Is multilingual supervised > monolingual unsupervised?



2. Does training on more languages help?



3. Is the choice of training language important?



Evaluation language

Training language	ES	HA	HR	SV	TR	ZH
CS	41.6	51.1	41.0	28.7	37.0	42.6
FR	42.6	41.8	30.4	25.3	32.5	35.8
PL	41.1	43.7	35.8	25.5	33.7	39.5
PT	45.9	46.2	36.4	26.6	34.1	39.6
RU	35.0	39.7	31.3	22.3	29.7	37.1
TH	28.5	44.5	29.9	17.9	23.6	36.2

Conclusions and future work

Conclusions:

- Proposed to train a supervised multilingual acoustic word embedding model on well-resourced languages and then apply to zero-resource languages
- Multilingual CAE-RNN and ClassifierRNN consistently outperform unsupervised models trained on zero-resource languages

Conclusions and future work

Conclusions:

- Proposed to train a supervised multilingual acoustic word embedding model on well-resourced languages and then apply to zero-resource languages
- Multilingual CAE-RNN and ClassifierRNN consistently outperform unsupervised models trained on zero-resource languages

Future work:

- Different models both for multilingual and unsupervised training
- Analysis to understand the difference between CAE-RNN and ClassifierRNN
- Does language conditioning help during decoding?

https://arxiv.org/abs/2002.02109

https://github.com/kamperh/globalphone_awe