Improving Unsupervised Acoustic Word Embeddings using Speaker and Gender Information

Lisa van Staden, Herman Kamper

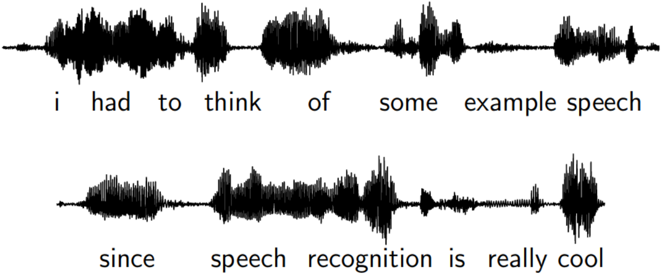
31 January 2020

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

100
1918 · 2018

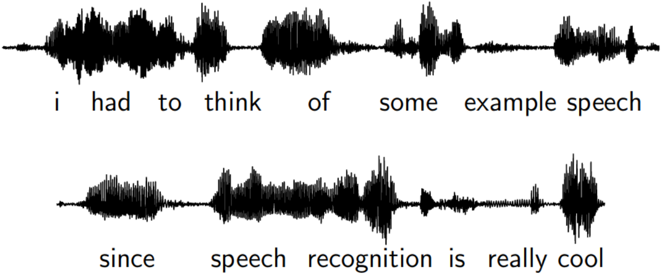
Zero-Resource Speech Processing

Popular methods for speech processing rely on transcribed speech.


i had to think of some example speech
since speech recognition is really cool

Zero-Resource Speech Processing

Popular methods for speech processing rely on transcribed speech.


i had to think of some example speech
since speech recognition is really cool

Obtaining transcriptions is expensive and not always possible.

Tasks in Zero-Resource Processing

We don’t always need to predict text labels:

Tasks in Zero-Resource Processing

We don’t always need to predict text labels:

- Query-by-Example Search: search speech using speech.

Tasks in Zero-Resource Processing

We don’t always need to predict text labels:

- Query-by-Example Search: search speech using speech.
- Unsupervised Term Discovery: Discover repeating patterns in speech.

Speech Segment Comparison

These tasks require comparing speech segments.

The conventional method is Dynamic Time Warping.

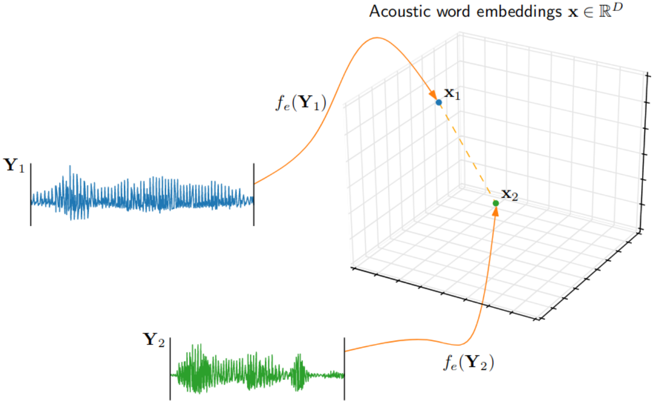
Speech Segment Comparison

These tasks require comparing speech segments.

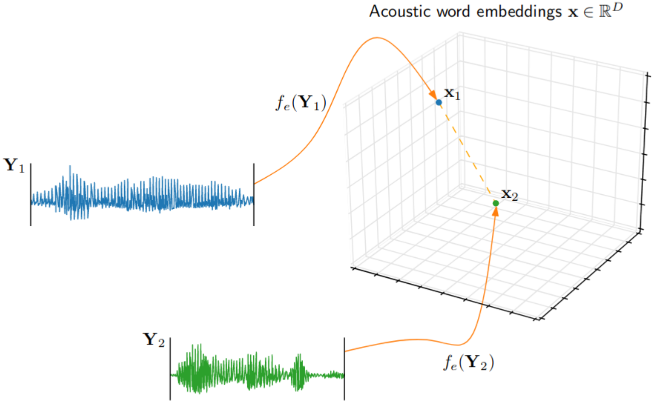
The conventional method is Dynamic Time Warping.

- Computationally expensive.

Acoustic Word Embeddings



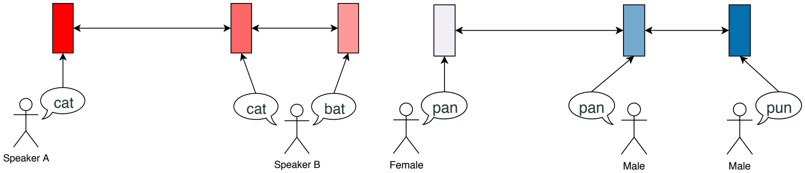
Acoustic Word Embeddings



We want to map speech to these representation without using labels.

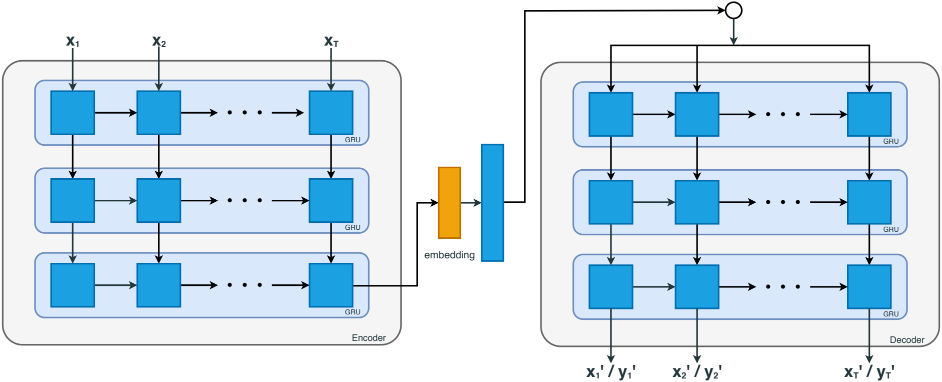
Speaker and Gender Information

Acoustic properties of speech from different speakers/genders differ.

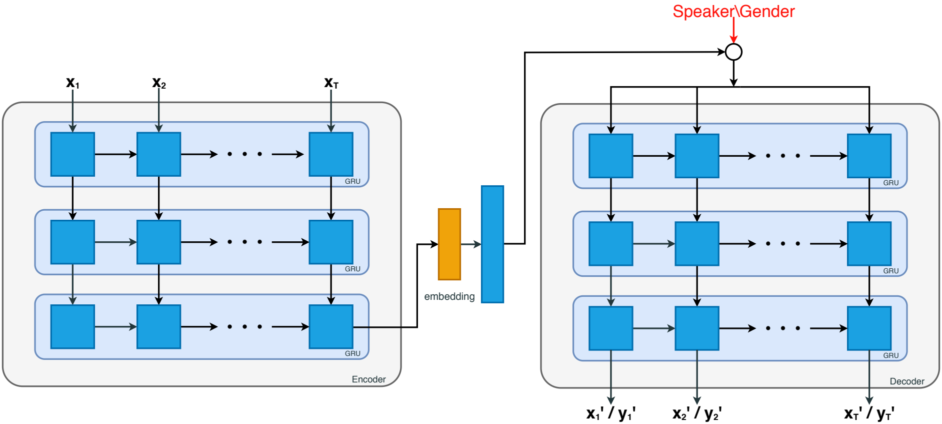


We want embeddings to be robust.

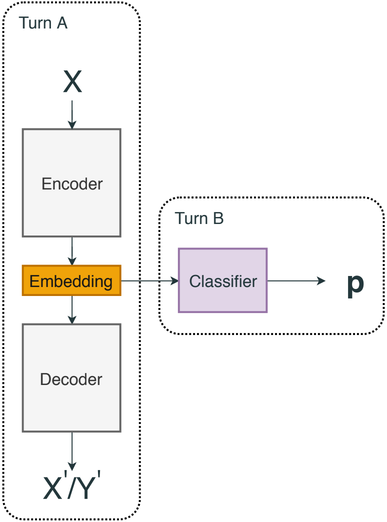
RNN (Correspondence) Autoencoder



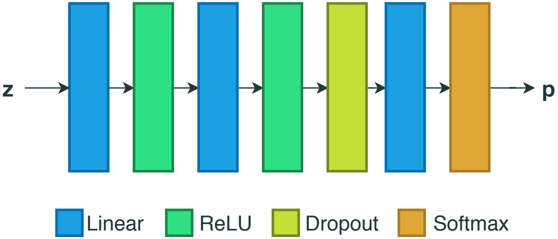
Speaker/Gender Conditioning



Adversarial Training



Speaker/Gender Classifier

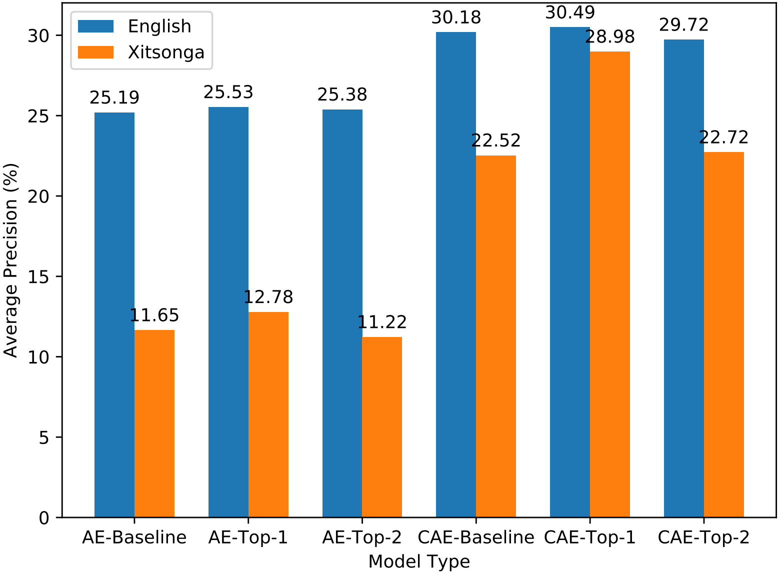


Evaluating Quality of AWEs

Use the same-different task to evaluate AWEs:

- Measure if AWEs are similar given a threshold.
- Calculate area under Precision vs Recall curve.

Results

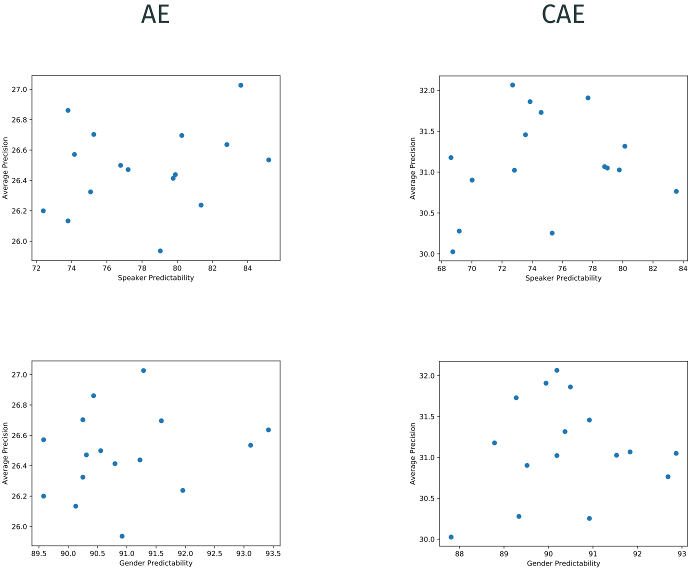

English
Xitsonga
Average Precision (%)
0
5
10
15
20
25
30
25.19
11.65
25.53
12.78
25.38
11.22
30.18
22.52
30.49
28.98
29.72
22.72
AE-Baseline
AE-Top-1
AE-Top-2
CAE-Baseline
CAE-Top-1
CAE-Top-2
Model Type

Evaluate Speaker and Gender Predictability

Analyse if the speaker and gender information has decreased:

- Use speaker/gender classifier model.
- Evaluate accuracy.

Average Precision vs Speaker/Gender Predictability



AE

CAE

Conclusions

- English data shows marginal improvement by incorporating speaker information.

Conclusions

- English data shows marginal improvement by incorporating speaker information.
- Best Xitsonga model shows 22% improvement.

Conclusions

- English data shows marginal improvement by incorporating speaker information.
- Best Xitsonga model shows 22% improvement.
- It’s difficult to remove speaker and gender information.

Conclusions

- English data shows marginal improvement by incorporating speaker information.
- Best Xitsonga model shows 22% improvement.
- It’s difficult to remove speaker and gender information.
- Future work ...