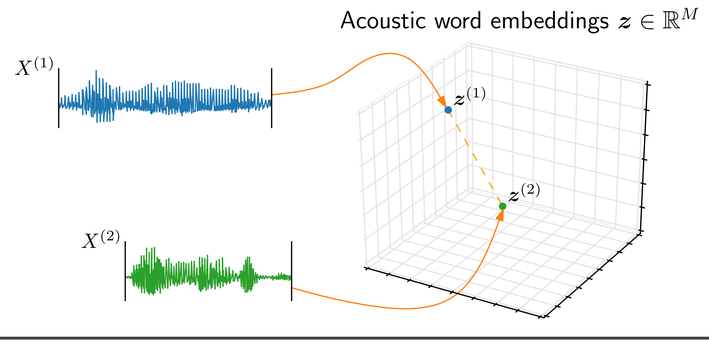
Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models

Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

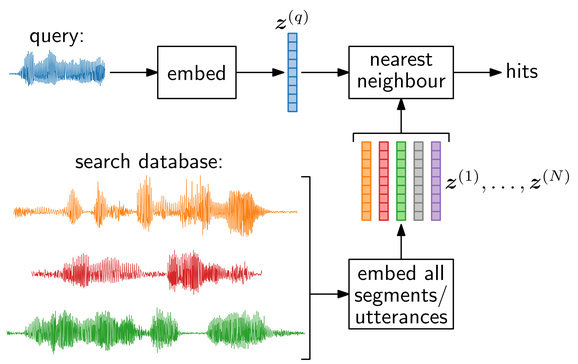
Background

- Current speech recognition methods require large labelled data sets.
- *Zero-resource speech processing* aims to develop methods that can discover linguistic structure directly from unlabelled speech.
- **Problem:** Need to compare speech segments of variable duration.
- Dynamic time warping (DTW) is one option: Can be slow.
- *Acoustic word embedding* methods map variable-length segments into fixed-dimensional space to enable efficient comparisons:

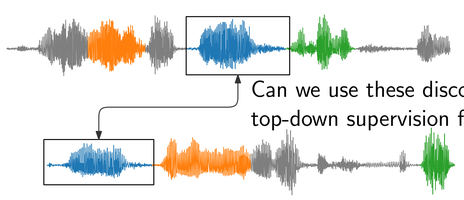


Example application: Query-by-example

Find utterances in a speech collection containing a given spoken query:



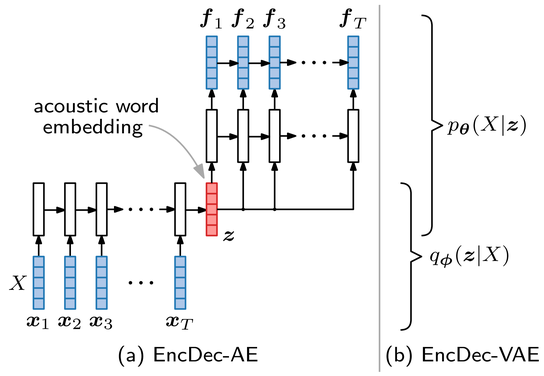
Unsupervised term discovery (UTD)



Discover recurring patterns in unlabelled speech.

Can we use these discovered word pairs as weak top-down supervision for an embedding model?

Encoder-decoder autoencoder (EncDec-AE) & variational autoencoder (EncDec-VAE)

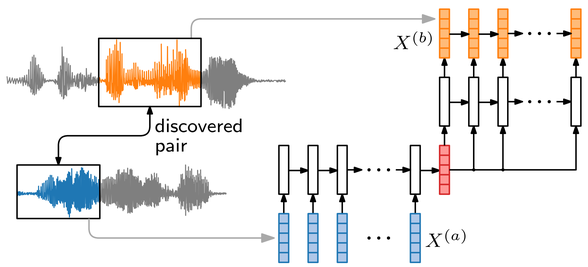


(a) EncDec-AE loss: $\ell(X) = \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{f}_t(X)||^2$

(b) EncDec-VAE loss:

$$\ell(X) = \sum_{t=1}^{T} \left[\frac{1}{2\sigma^2} ||\boldsymbol{x}_t - \boldsymbol{f}_t(\boldsymbol{z}')||^2 + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|X)||p(\boldsymbol{z})\right) \right]$$

Encoder-decoder correspondence autoencoder (EncDec-CAE)



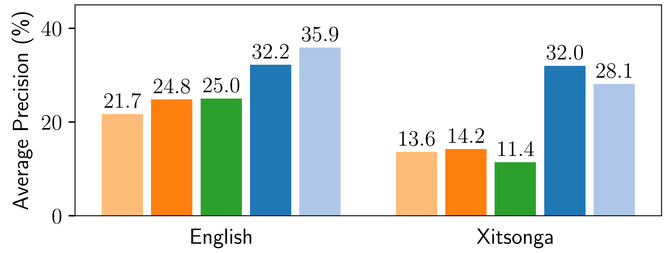
EncDec-CAE loss: $\ell(X^{(a)}, X^{(b)}) = \sum_{t=1}^{T_b} ||\boldsymbol{x}_t^{(b)} - \boldsymbol{f}_t(X^{(a)})||^2$

Experimental details

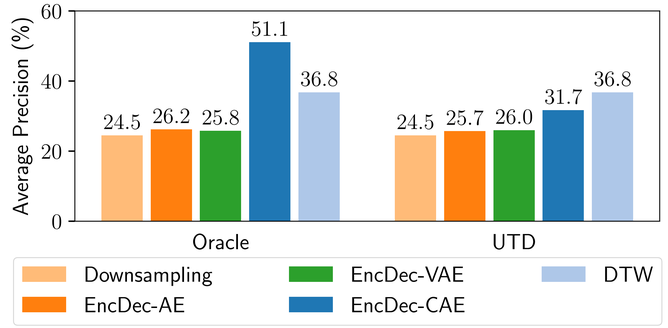
- **Data:** English (Buckeye) and Xitsonga (NCHLT).
- **Evaluation:** Same-different isolated word discrimination.
- **Embeddings:** $M = 130$ embedding dimensions for all models.
- **Upper and lower bound approaches:**
  - Downsampling: 10 equally spaced MFCCs flattened.
  - DTW: Use alignment cost between full test segments.

Same-different word discrimination results

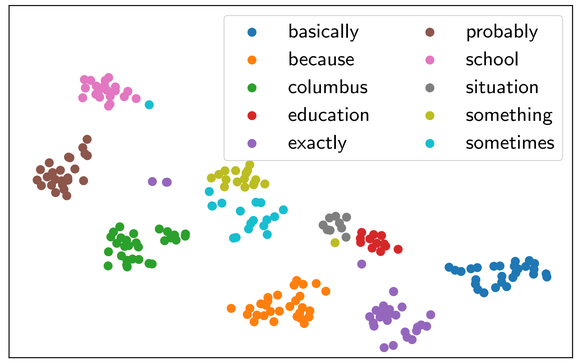
Test results when using segments discovered using UTD for training:



English development results comparing ground truth (oracle) and UTD:



t-SNE visualisation of EncDec-CAE embeds



Conclusions and future work

- Old and new acoustic word embedding methods were evaluated.
- EncDec-CAE outperforms other approaches on two languages.
- **Future:** Use acoustic word embeddings from EncDec-CAE in downstream zero-resource speech technology.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

https://github.com/kamperh/recipe_bucktsong_awe