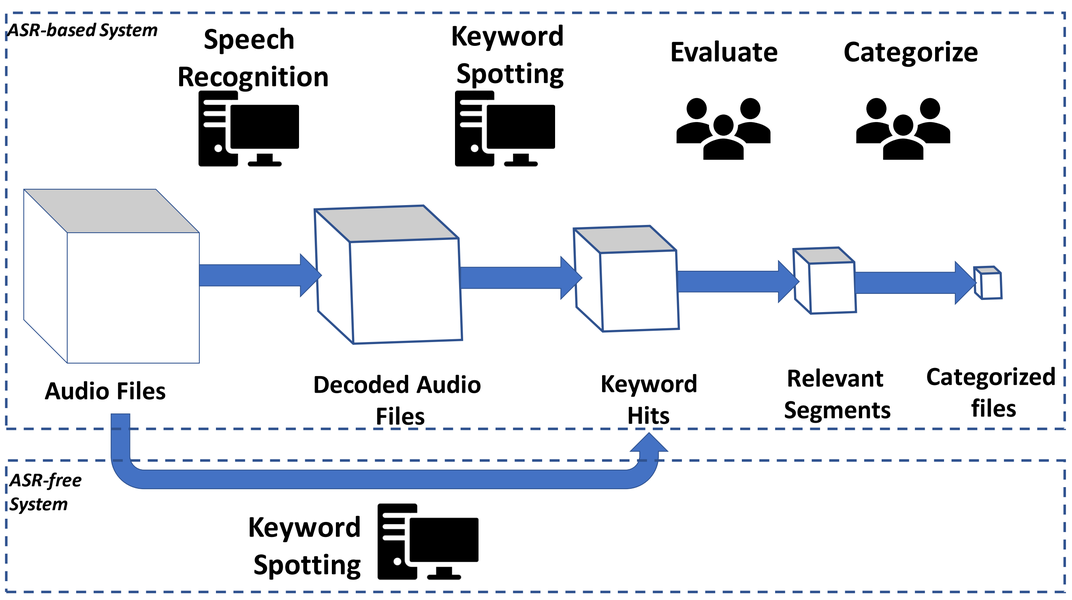
Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders

Raghav Menon[1], Herman Kamper[1], Ewald van der Westhuizen[1], John Quinn[2], Thomas Niesler[1]
[1]Stellenbosch University, South Africa; [2]UN Global Pulse, Kampala, Uganda; [2]University of Edinburgh, UK

Summary

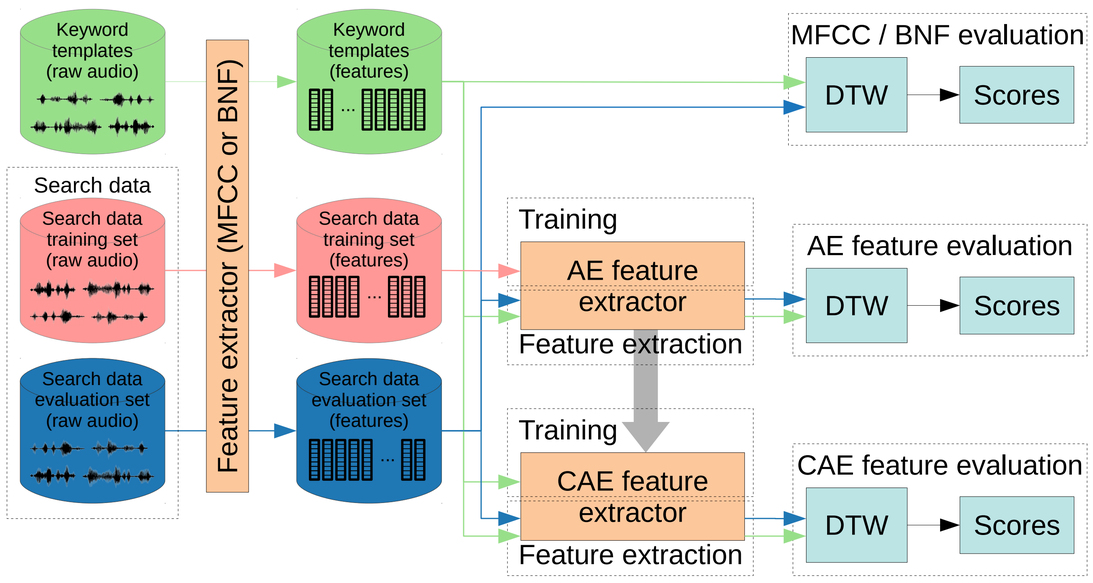
1. Features for dynamic time warping (DTW) in an almost zero-resource setting for a keyword spotting (KWS) application are compared.
2. The keyword spotting systems aid the United Nations (UN) humanitarian relief efforts in parts of Africa with severely under-resourced languages.
3. The objective is to identify acoustic features that provide acceptable KWS performance in such environments.
4. A small, independently compiled set of isolated keywords is the only supervised resource.
5. Multilingual bottleneck features (BNFs) from well-resourced out-of-domain languages and correspondence autoencoder (CAE) features are evaluated.
6. BNFs and CAE features achieve modest (> 2%) performance improvements over baseline MFCCs.
7. BNFs as input to the CAE result in notable (> 11%) performance improvements over MFCCs for two evaluated languages, English and Luganda.
8. Integrating BNFs with the CAE allows both large out-of-domain and sparse in-domain resources to be exploited for improved ASR-free keyword spotting.

Radio browsing system



- Live audio from phone-in radio talk shows is processed and monitored for keywords.
- Current ASR-based radio browsing systems require large annotated speech resources.
- **Dynamic time-warping** (DTW) keyword spotting systems:
  - are word template-based;
  - can perform in an almost zero-resource setting.

Feature extraction and evaluation



- Various feature extractors are evaluated.
- Autoencoder (AE) and correspondence autoencoder (CAE) extractors are trained on **unlabelled** training data.
- Keyword templates are used to fine-tune the CAE.
- DTW performs template matching on evaluation search data using the keyword templates.
- The presence of a keyword is determined by applying a threshold to the DTW scores.

Data sets

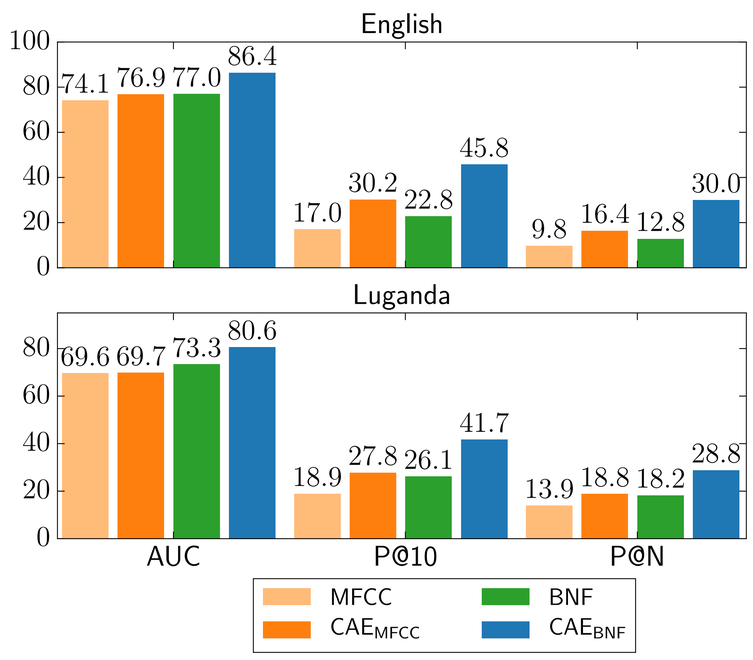
- **Search data** from radio talk show speech.
  - Training data is unlabelled.
  - Only evaluation sets are labelled.

Set	English		Luganda	
	#utts	duration (h)	#utts	duration (h)
Train	5 231	7.94	6 052	5.57
Dev	2 740	5.37	1 786	2.04
Test	5 005	10.33	1 420	1.99
Total	12 976	23.64	9 258	9.60

- **Keywords templates** are the only labelled in-domain data and are used to train the KWS.

Language	# keywords	# speakers	# utterances
English	40	24	1 160
Luganda	18	16	603

Results

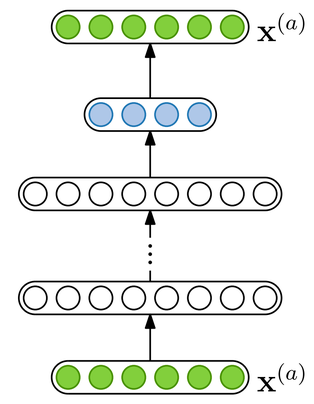


- AUC: Area under the receiver operating characteristic curve.
- P@10: Precision at 10 is the proportion of correct keyword detections among the top 10 hits.
- P@N: Precision at N is the proportion of correct keyword detections among the top N hits.
- In terms of AUC:
  $\mathbf{CAE_{BNF}} > \text{BNF} > \text{CAE}_{\text{MFCC}} > \text{MFCC}$
- Multilingual feature extraction and target language fine-tuning are complimentary.

Neural network feature extraction

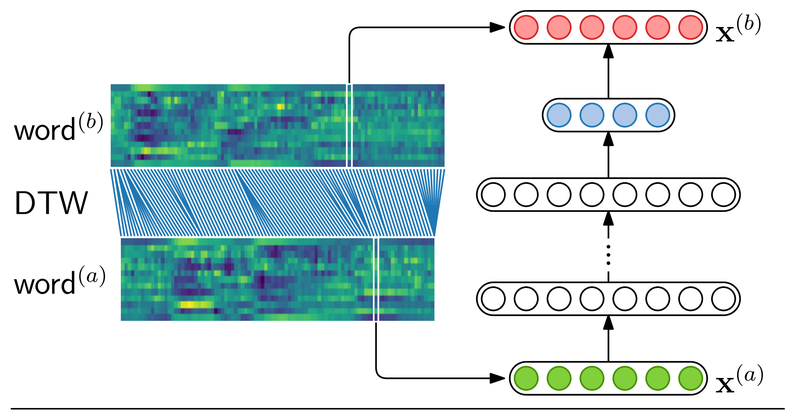
Autoencoder

- The same feature frame is used at the input and output of the network.
- Hence no annotations or labels required for training.



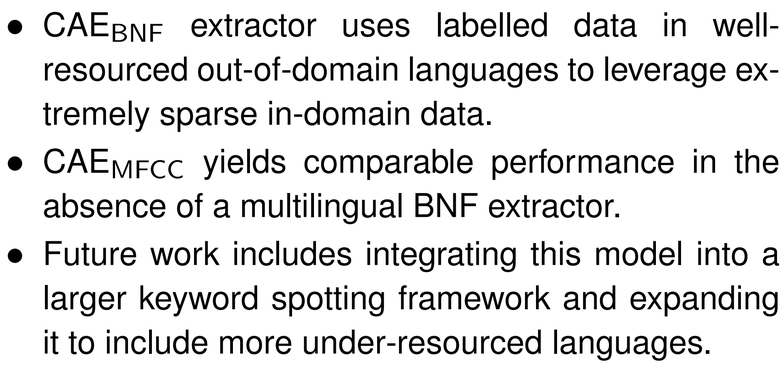
Correspondence autoencoder

- Two different instances of a word aligned by DTW.
- Alignments used to train CAE.
- Factors not common to keyword pairs (speaker; gender; channel) are suppressed, while common factors (word identity) are enhanced.



Multilingual bottleneck feature extractor

- Ten languages from the GlobalPhone corpus are used at the output of the network.



Combinations of features with NNs

- AE with MFCC → AE_{MFCC}
- AE with BNF → AE_{BNF}
- CAE with MFCC → CAE_{MFCC}
- CAE with BNF → CAE_{BNF}

Conclusion

- Keyword templates are the only labelled data.
- Extractor and feature combinations can lead to improved KWS performance.
- CAE_{BNF} yielded the best performance among the evaluated feature types.
- CAE_{BNF} extractor uses labelled data in well-resourced out-of-domain languages to leverage extremely sparse in-domain data.
- CAE_{MFCC} yields comparable performance in the absence of a multilingual BNF extractor.
- Future work includes integrating this model into a larger keyword spotting framework and expanding it to include more under-resourced languages.

Acknowledgements

We thank NVIDIA Corporation and Telkom South Africa for equipment and funding support. H. Kamper is supported by a Google Faculty Award. We thank Enno Hermann for assisting with the BNF extraction.