ASR-free CNN-DTW keyword spotting using multilingual bottleneck features for almost zero-resource languages

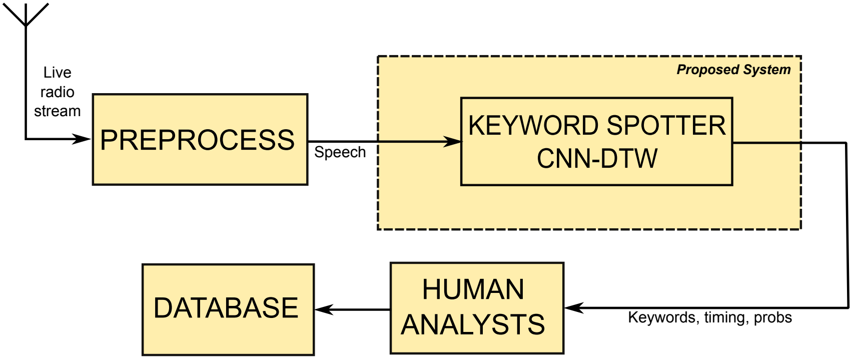
Raghav Menon, Stellenbosch University, South Africa
Herman Kamper, Stellenbosch University, South Africa
Emre Yilmaz, Radbound University & National University of Singapore
John Quinn, UN Global Pulse, Kampala, Uganda
Thomas Niesler, Stellenbosch University, South Africa

August 2018

Introduction

- Social media has become popular for voicing social concerns and views.
- Not true when internet accessibility is poor
- United Nations (UN) survey shows that in Uganda phone-in talk shows are the medium of choice outside metropolitan areas.
- Radio browsing system have been actively supporting UN relief and development programmes by monitoring this medium.
- However these systems are highly dependent on transcribed speech in the target language.
- Radio browsing systems for Acholi and Luganda using approximately 9 hours of data was developed and it took many months to obtain the data.
- We describe a keyword spotting system which relies on only a small number of isolated repetitions of keywords and a large body of untranscribed data.

Radio browsing system



Data

- **In-domain data:** 40 keywords, each spoken twice by 24 South African speakers (12 male, 12 females).
- **Untranscribed data:** 23-hour South African Broadcast News (SABN) corpus.
  - Mix of English newsreader speech, interviews and crossings to reporters broadcast between 1996 and 2006.

	Utterances	**Speech (h)**
Train	5231	7.94
Dev	2988	5.37
Test	5226	10.33
Total	13445	23.64

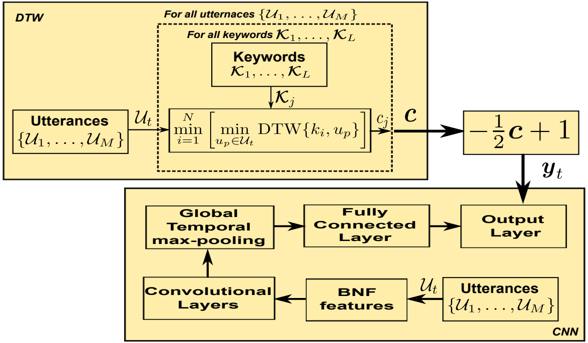
Keyword spotting approaches

- Dynamic time warping (DTW)
  - Good in low resource setting but prohibitively slow as it requires repeated alignment
  - Isolated words are slid one at a time over the search audio with a 3 frame skip.
  - Normalized per frame cosine cost.
  - Presence or absence of keyword determined using appropriate threshold.
- Convolutional neural network (CNN) classifier
  - The CNN was trained as a end-to-end classifier with each keyword example.
  - CNN consists of 3 convolutional layers with max pooling followed by 3 dense layers.
  - Input size restricted to 60 frames.
  - Presence or absence of keyword based on appropriate threshold.

DTW and CNN are baselines.

Keyword spotting approaches

- CNN-DTW keyword spotting
  - CNN-DTW keyword spotting approach uses DTW to generate training data for CNN.
  - Scores calculated between the small set of isolated keywords and a much larger untranscribed dataset which are subsequently used as targets to train a CNN.



- MFCC, bottleneck and autoencoder features considered.

Bottleneck and Autoencoder features

- Large annotated speech resources exist for well-resourced languages.
- We investigate whether these resources can be used to improve the performance of our CNN-DTW.
- Bottleneck features
  - 2-language TDNN: A 11-layer 2-language TDNN trained using the FAME and CGN corpora comprising of approximately 887 hrs of Flemish and Dutch data.
  - 10-language TDNN: A 6-layer 10-language TDNN was trained on Globalphone corpus containing 198 hrs of training data.
- Autoencoder features
  - An autoencoder is a neural network used to reconstruct its input.
  - Can be trained when large amounts of unlabelled data available.
  - Like the BNFs, autoencoders can be trained on different languages.
  - We obtain a 7-layer stacked denoising autoencoder by training each layer individually.
  - Languages used were Acholi (160 hrs), Luganda (154 hrs), Lugbara (9.45 hrs), Rutaroo (7.82 hrs) and Somali (18 hrs).

Experimental setup

- Three baseline systems are considered
  - DTW-QbyE - where DTW is performed for each exemplar keyword on each utterance and the resulting scores averaged.
  - DTW-KS - best score over all exemplars of a keyword type is used.
  - CNN - An end-to-end CNN classifier trained only on the isolated keywords.
- CNN-DTW is supervised by the DTW-KS system.
- SABN transcriptions not used for training or validation, but were used to access accuracy.
- Hyper-parameters optimized by minimizing the target loss on the development set.
- Performance is reported in terms of AUC and EER.

Experimental Results

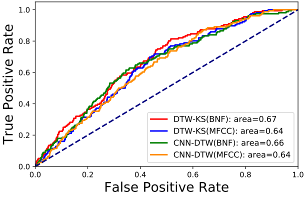
- We consider four feature extractors:
  - Stacked Autoencoder.
  - the 2-language TDNN without speaker normalisation.
  - the 10-language TDNN without speaker normalisation.
  - the 10-language TDNN with speaker normalisation.

Model	**dev**	
	AUC	**EER**
MFCC	0.7556	0.3092
SAE	0.5247	0.4844
TDNN-BNF-2lang	0.7273	0.3356
TDNN-BNF-10lang	0.7725	0.2884
TDNN-BNF-10lang-SPN	**0.7781**	**0.2872**

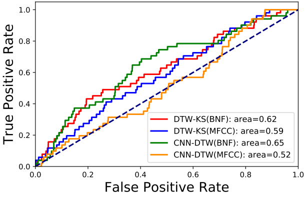
Experimental results

Model	AUC				EER			
	dev		test		dev		test	
	MFCC	BNF	MFCC	BNF	MFCC	BNF	MFCC	BNF
CNN	0.5698	0.5298	0.5448	0.5364	0.4435	0.4813	0.4771	0.4725
DTW-QbyE	0.6639	0.6899	0.6612	0.6873	0.3864	0.3556	0.3885	0.3661
DTW-KS	0.7556	0.7781	0.7515	0.7699	0.3092	0.2872	0.3162	0.3012
CNN-DTW	0.6360	0.7537	0.6285	0.7422	0.4073	0.3058	0.4161	0.3214
CNN-DTW-GNL	0.6443	0.7535	0.6357	0.7518	0.4036	0.3091	0.4092	0.3153

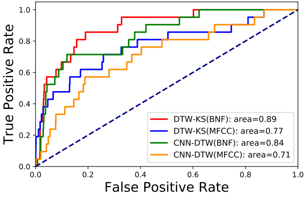
Experimental results



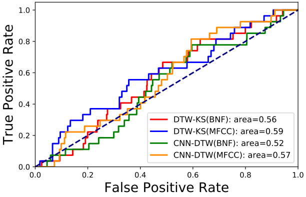
(a) Keyword: Government



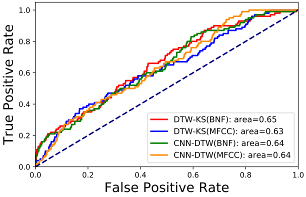
(b) Keyword: Attack



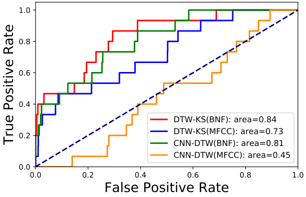
(c) Keyword: HIV



(d) Keyword: Health



(e) Keyword: War



(f) Keyword: Wounded

Conclusion

- We investigated the use of multilingual bottleneck (BNF) and autoencoder features in a CNN-DTW keyword spotter.
- The autoencoder features and BNFs trained on two languages did not improve performance over MFCCs, but BNFs trained on a corpus of 10 languages lead to substantial improvements.
- We conclude that our CNN-DTW approach, which combines the low-resource advantages of DTW with the speed advantages of CNN, benefits from incorporating labelled data from other well-resourced languages through the use of BNFs.