UNIVERSITEIT iYUNIVESITHI STELLENBOSCH UNIVERSITY

100 1918 · 2018

forward together · saam vorentoe · masiye phambili

Automatic speech recognition and keyword spotting in under-resourced languages

Digital Signal Processing Group, E&E Engineering

21 February 2020

GLOBAL PULSE

DSP group

DSP group: More than speech

http://dsp.sun.ac.za/~trn

- Communication network for wildlife sensors
- Optimised kinetic energy harvesting
- Automatic detection and classification of coughing in audio
- Virtual reality visualisation and analysis of microscopy data
- Sensor network for viticulture
- Interactive document visualisation for the blind

Automatic Language Processing: Then

Automatic Language Processing: Now

Google Voice

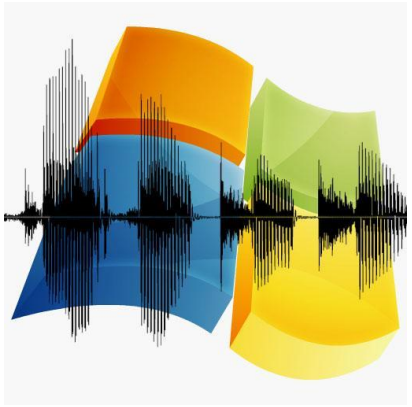
NUANCE

DRAGON NaturallySpeaking

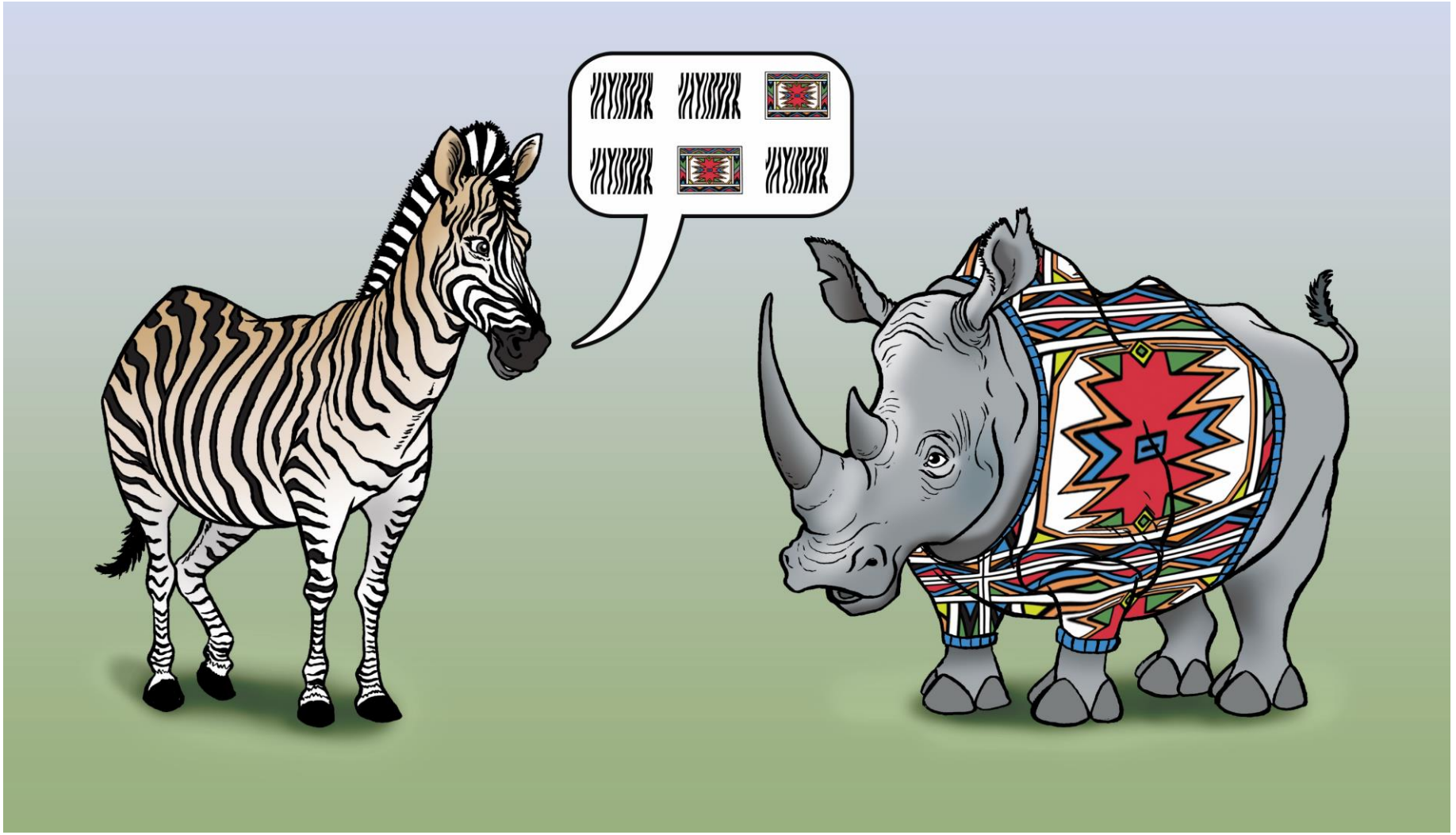
IBM

SAMSUNG

amazon alexa

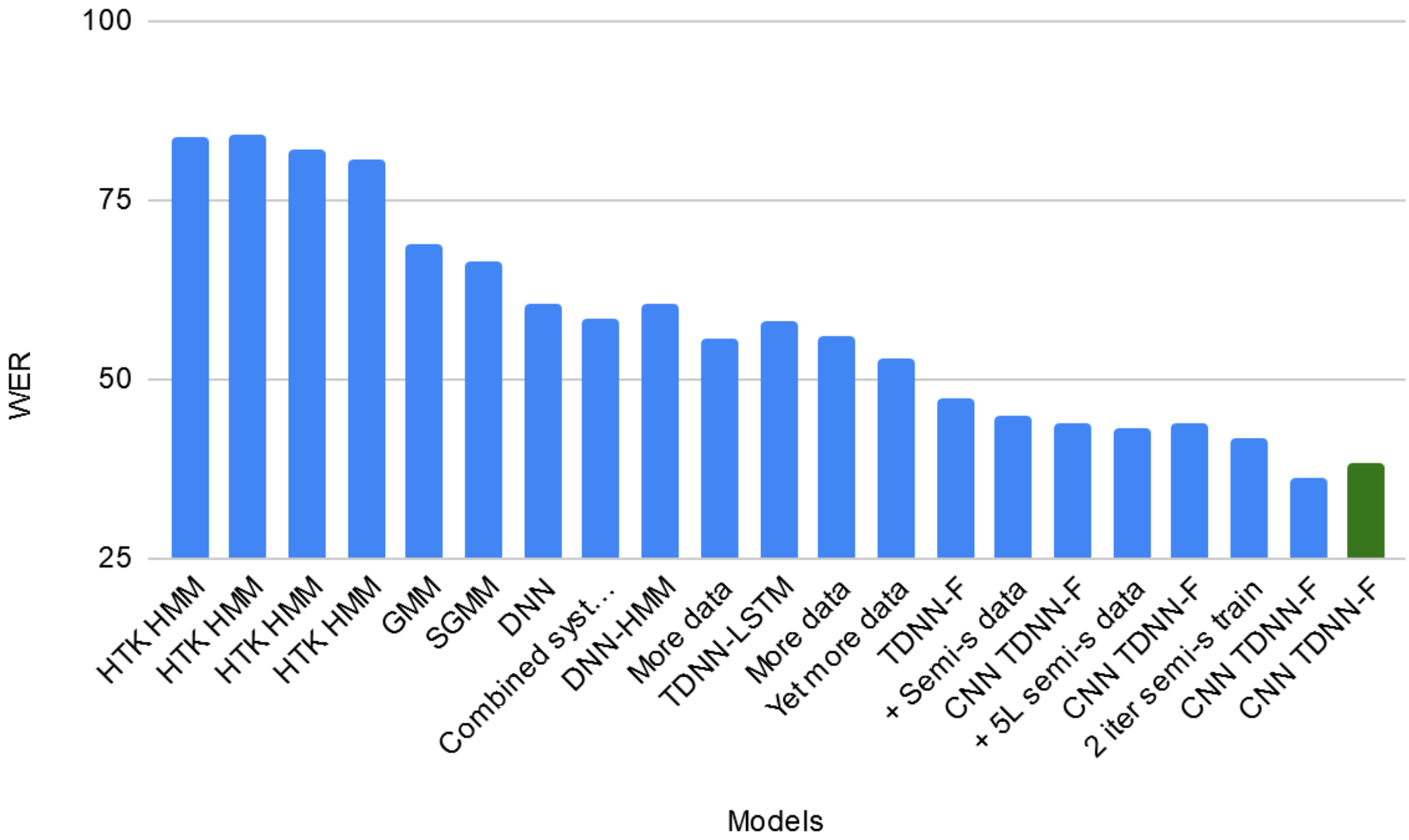


Language usage in South Africa



Multilingual corpus of code-switched South African speech

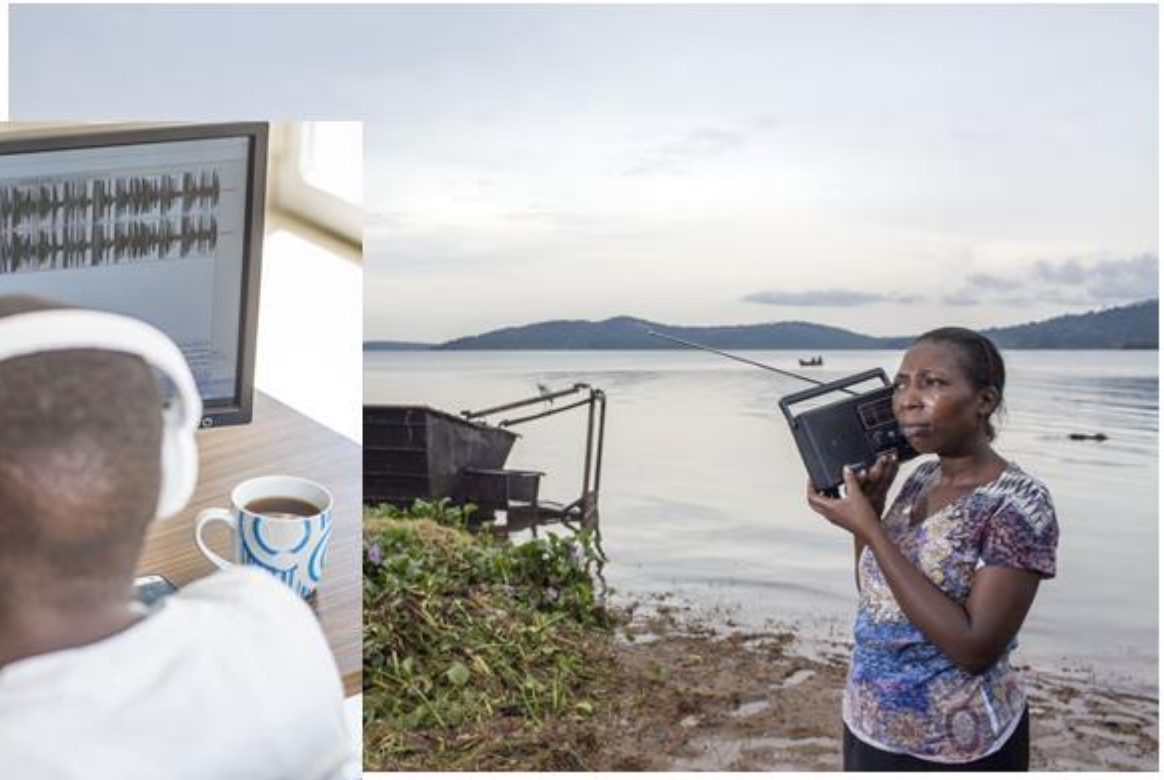
English – isiZulu CS speech



UN project

When Old Technology Meets New:

How UN Global Pulse is Using Radio and AI to Leave No Voice Behind



Target Languages

Speech data

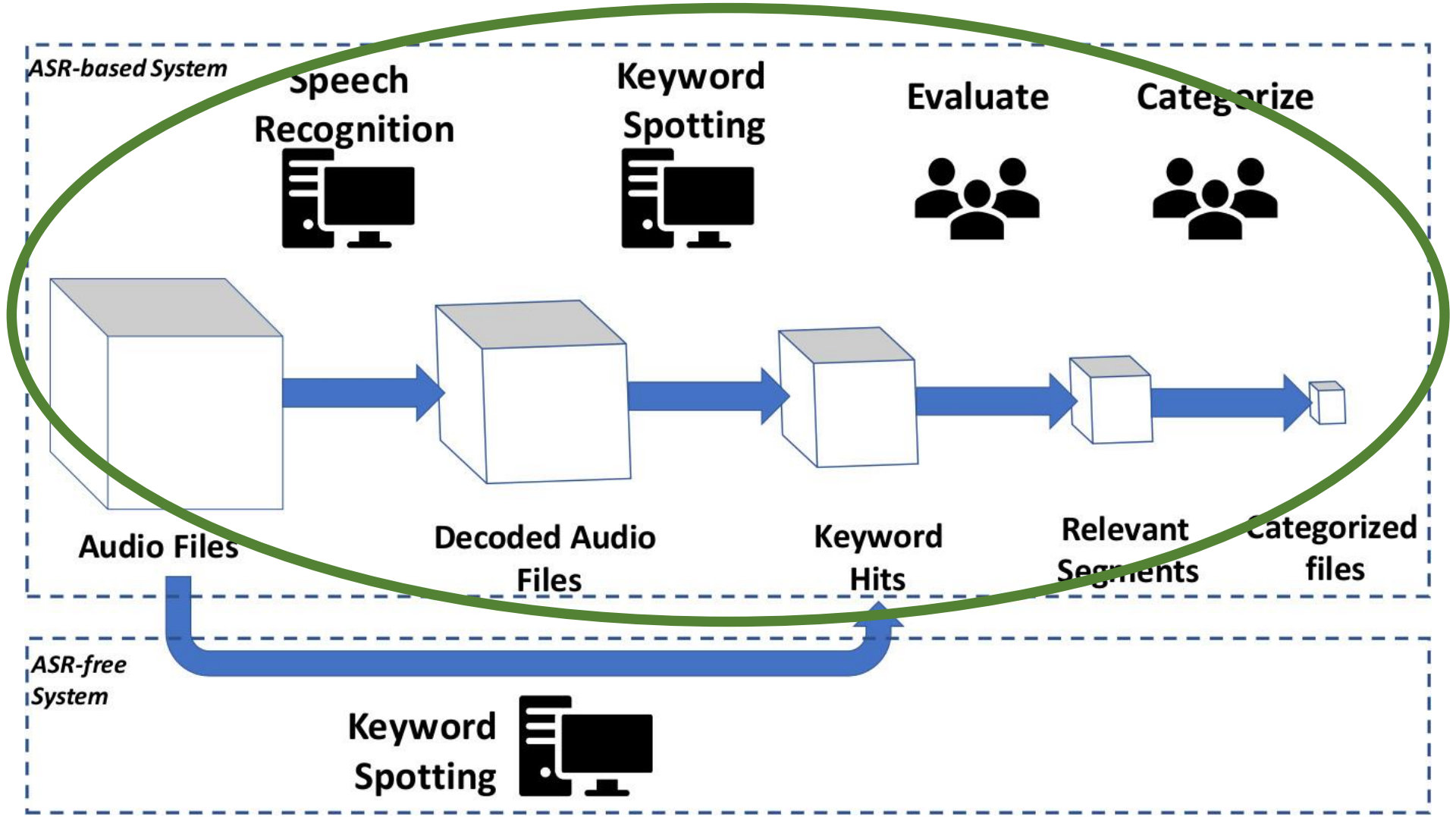
- Ugandan English (6h), Luganda (9h), Acholi (9h, 12min)
- Somali (1.6 h)
- UE was augmented with SAE data (20h)

Text data

- 109 million SAE words
- 1 million Luganda words (online newspaper)
- Transcriptions of the audio data

Pronunciation rules : Phonetic experts

ASR-free CNN-DTW keyword spotting



Acoustic modelling

Acoustic models: data perturbation

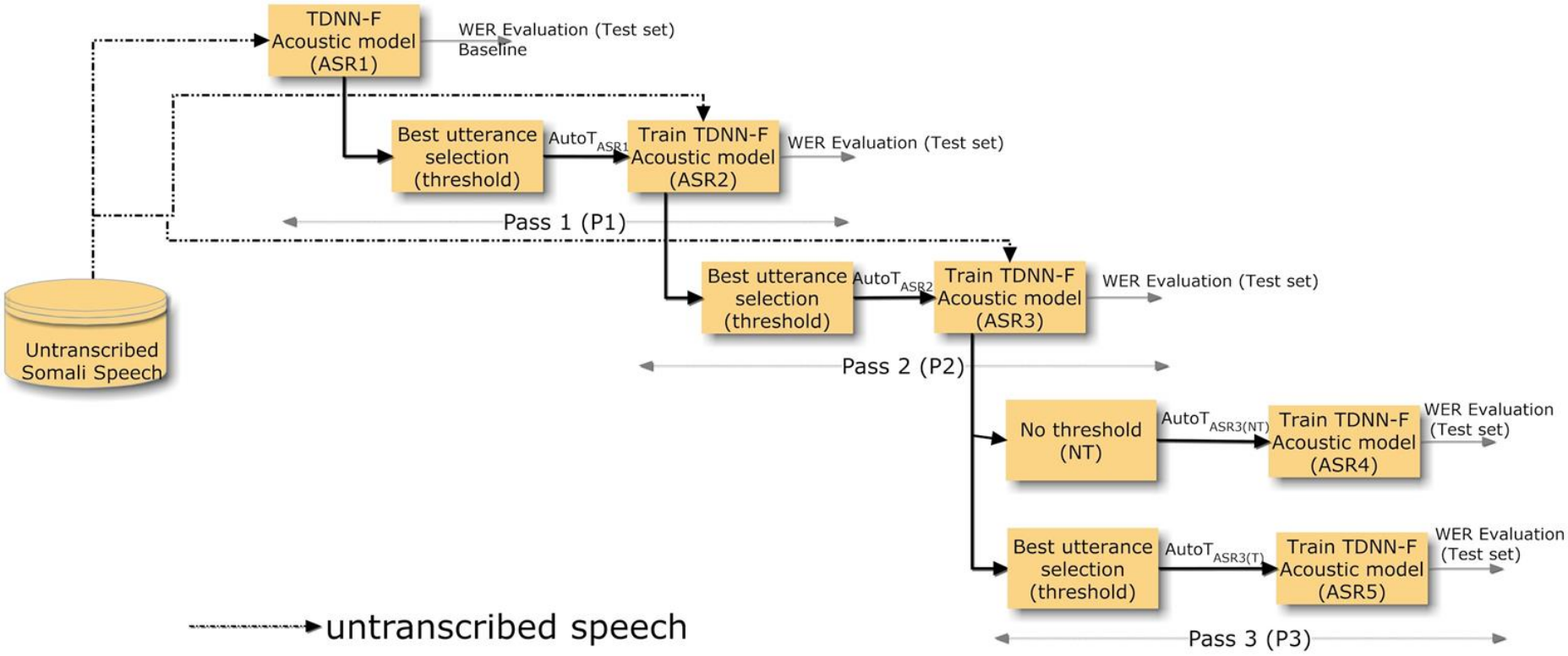
- Convolutional Neural Networks (CNNs)
- Time-Delay Neural Networks (TDNNs)
- Bi-directional Long Short-Term Memory NN (BLSTMs)

Language models: data augmentation

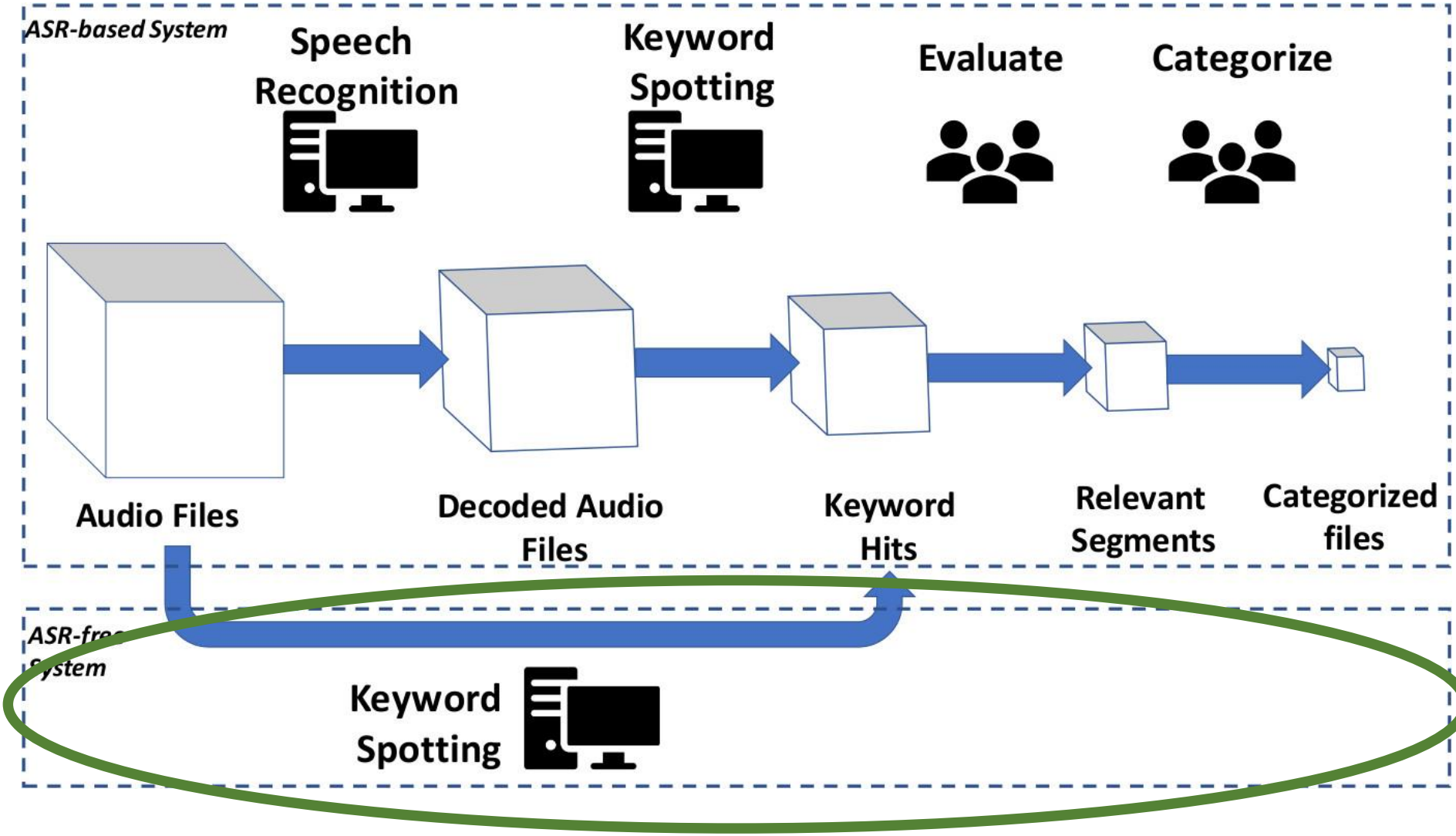
- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory Neural Networks (LSTMs)

Somali speech recognition

Multi-pass semi-supervised training



ASR-free CNN-DTW keyword spotting



ASR-free CNN-DTW keyword spotting

Aim:

- Rapid deployment of keyword spotting systems in new languages

Idea:

- Use Dynamic Time Warping (DTW) as supervision to train Convolutional Neural Networks (CNNs) using small set of isolated keywords
- Recordings of keywords are used as exemplars in DTW template matching, apply to untranscribed speech
- Use DTW scores as targets to train CNN on same unlabelled data
- Very little labelled data is required but large amount of unlabelled data can be leveraged

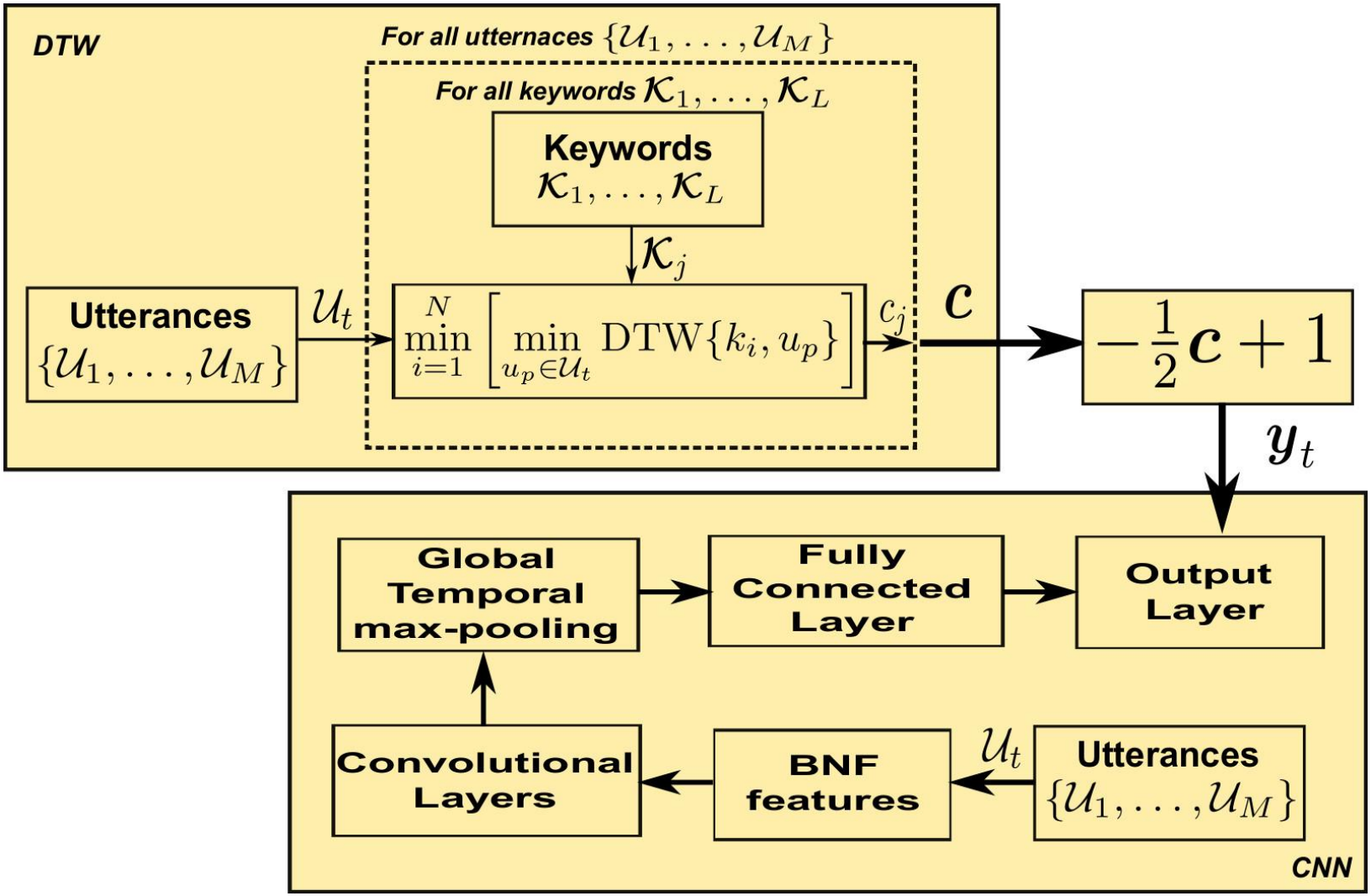
Features for ASR-free keyword spotting

- Query-by-example: search "string" provided as audio
- Use Dynamic Time Warping to match query with utterances in search collection
- Various feature representations investigated, e.g.
  - Multilingual bottleneck features (2 & 10 languages)
  - Stacked autoencoder
  - Correspondence autoencoder
  - Combinations of these

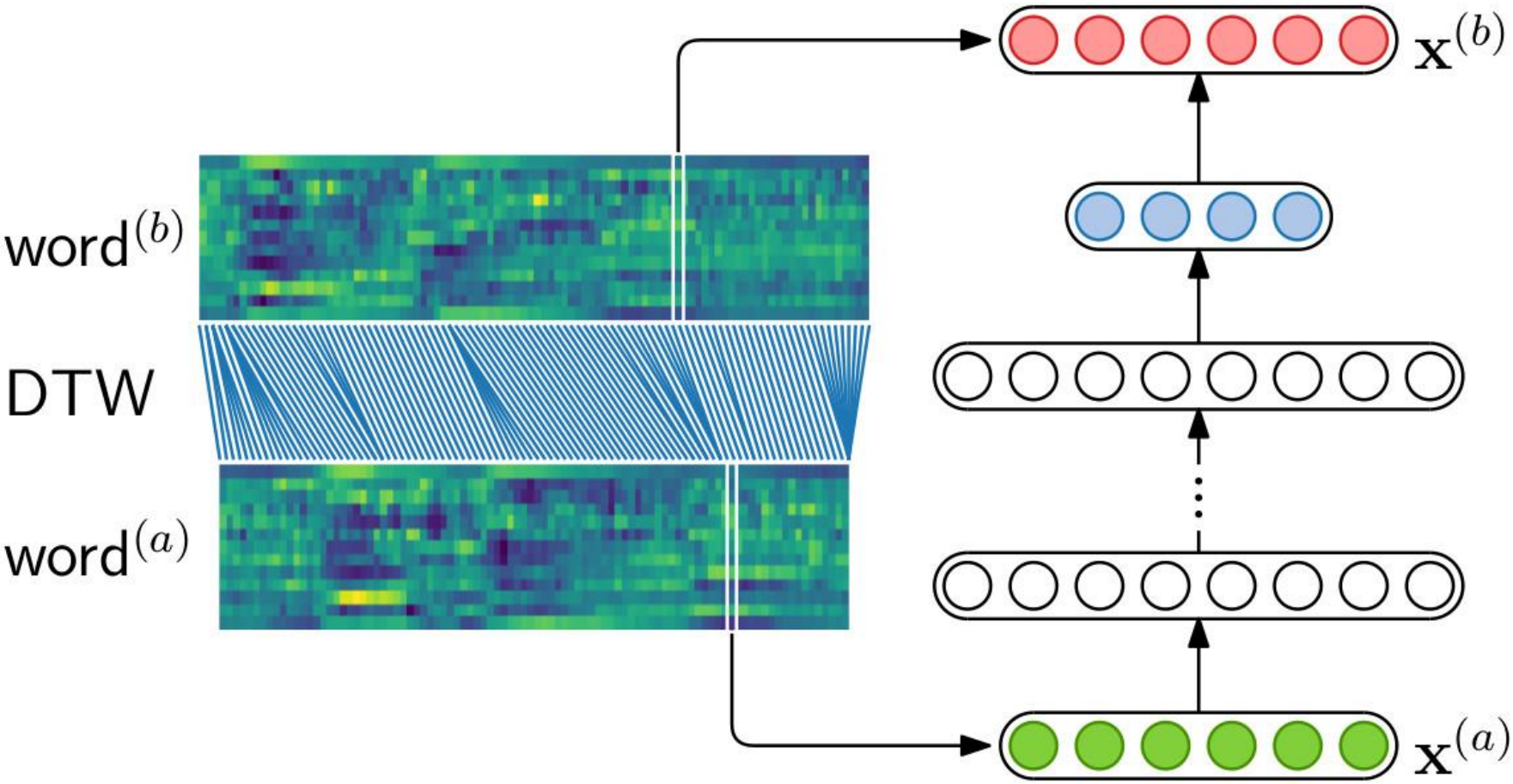
Results

- Multilingual feature extraction combined with target language fine-tuning can be complimentary
- CCN keyword spotting does not match DTW-based system
- BUT outperforms CNN classifier trained only on keywords
- Main advantage of CNN: orders of magnitude faster at runtime than DTW
- Feature extractors trained on well-resourced datasets can improve performance
- Best performance: CAE trained on BNF

CNN DTW


DTW
For all utternaces $\{\mathcal{U}_1, \ldots, \mathcal{U}_M\}$
For all keywords $\mathcal{K}_1, \ldots, \mathcal{K}_L$
Keywords
$\mathcal{K}_1, \ldots, \mathcal{K}_L$
$\mathcal{K}_j$
Utterances
$\{\mathcal{U}_1, \ldots, \mathcal{U}_M\}$
$\mathcal{U}_t$
$\min_{i=1}^{N} \left[\min_{u_p \in \mathcal{U}_t} \mathrm{DTW}\{k_i, u_p\} \right]$
c_j
$\boldsymbol{c}$
$-\frac{1}{2}\boldsymbol{c} + 1$
$\boldsymbol{y}_t$
Global Temporal max-pooling
Fully Connected Layer
Output Layer
Convolutional Layers
BNF features
$\mathcal{U}_t$
Utterances
$\{\mathcal{U}_1, \ldots, \mathcal{U}_M\}$
CNN

Correspondence autoencoder



Keyword spotting examples

Topic	Analyst translation
natural-disaster, food-security	"Elephants that are suspected to have come from South Sudan went and attacked Abalo Kodi village and destroyed food [crops] about 20 acres."
refugees.camps	"I stand with my two legs and say that staying in the camps is very very good [...] those days when people were not in the camps they used to keep money in anthills and under the beds, but after coming out of the camps they have knowledge about banking."
health.service-delivery	"The road here is so bad that the ambulance got stuck in a ditch and could not reach the hospital. People came and had to collect the medicine and carry it on foot to the hospital."
health.malaria-prevention	"People are using mosquito nets in the wrong way, for example scrubbing their bodies, washing dishes, making fences around chicken houses, some even turkey houses or pigsties."

Current work

Mali

- More volatile environment
- Difficult to install transmitters without raising suspicion
- Bambara, Fulani
- Some transcribed data, no text