Keyword Spotting from Continuous Speech using DTW and CNN

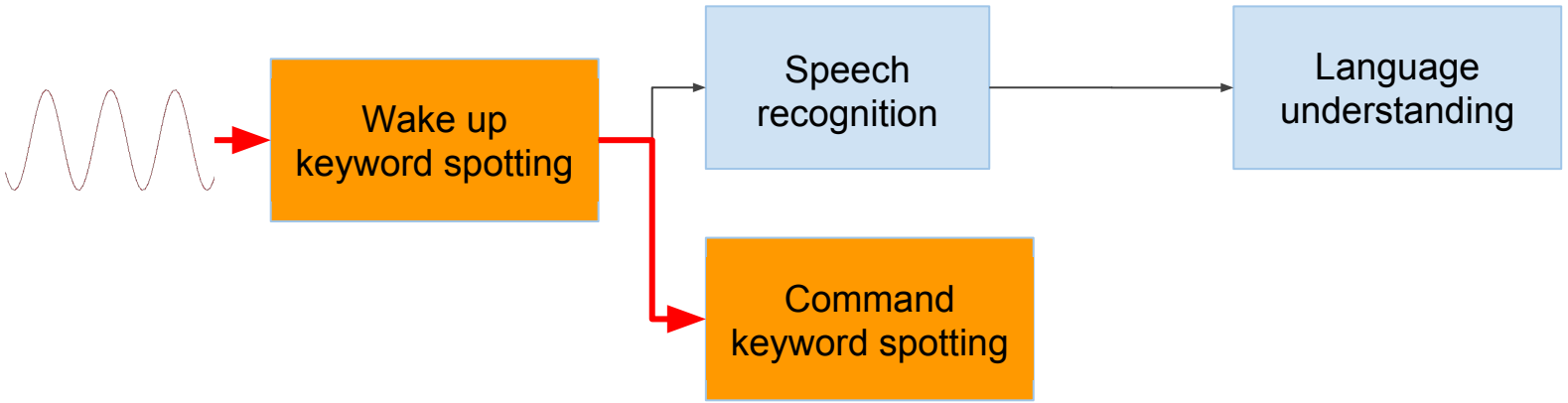
Graduate: Erika-Timea ALBERT
Supervisors: Prof. dr. eng. Rodica POTOLOEA
Prof. dr. eng, Mihaela DINSOREANU
Conf. dr. eng. Camelia LEMNARU

Problem Formulation

- **DOMAIN**: Natural Language Understanting/ Spoken Language Understanding
- **PROBLEM**: Accurately extract information from spoken language in the context of **Home Assistance**

Possible Strategies

1. Speech → text → meaning
2. **Speech** → **meaning (keyword)**



Related Work

- Conventional **Automatic Speech Recognition** methods require:
  - **large amount** of language-specific **annotated audio** data
- **Experiment with KALDI:**
  - speech recognition tool written in C++
  - needs data annotated at phone level
  - simple exmaple integrated
    - **3 speakers**
    - **numbers 0-9**
    - **3 x 10 audio / number**
    - **WER - 20%**

```
eight ey t
five f ay v
four f ao r
nine n ay n
one hh w ah n
one w ah n
seven s eh v ah n
six s ih k s
three th r iy
two t uw
zero z ih r ow
zero z iy r ow
```

KALDI: Issues

- Models have to be retrained for every language
- Large amount of data needed
- Annotate data precisely
  - time consuming
  - requires expertise

Between the circumstances above ASR methods are very effective

Proposed Solution (Unsupervised Approach)

- Generate feature vector by extracting MFCC values from the audio signal, training a GMM model on the resulting vectors and using that model to generate posteriorgrams (as feature vectors)
- Compute the DTW warping matrix resulted by matching the **keyword** posteriorgram to the **utterance** posteriorgram
- Convert it into a grayscale image and train a CNN classifier

Proposed Solution (Pipeline)

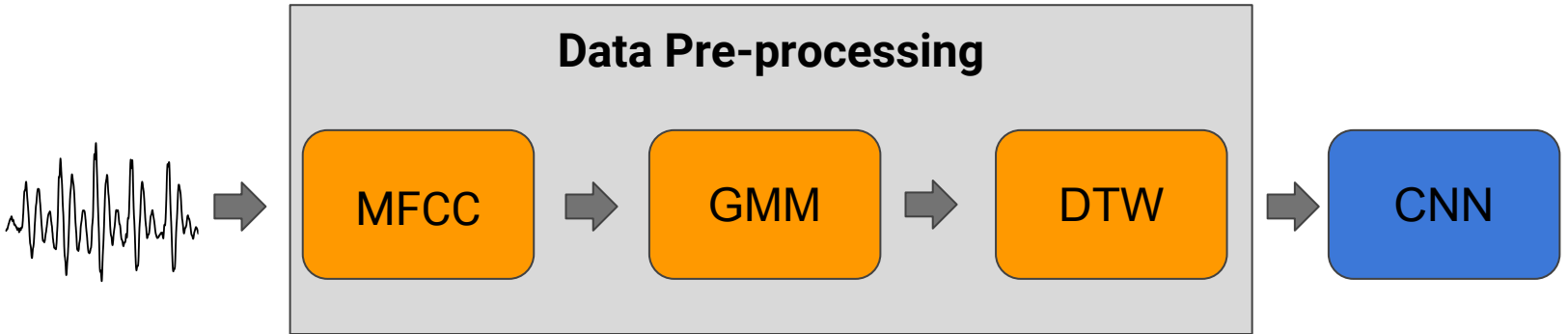
MFCC for short, overlapping frames (25 ms, 10 ms overlap)
- TRAIN + Keywords

13-dim. vectors (first 13 out of 26)
- discard high frequencies
- identify speech components

GMM trained with 50 components
- TRAIN + Keywords

Modified DTW - taking avg cost
Grayscale images scaled to 32x128

CNN trained on 32x32 patches - each labelled as parent image



Data

TIMIT Acoustic - Phonetic Speech Corpus

- Recorded: 1993
- Sample Rate: 16000
- Language: English
- Nr. speakers: 630
- Nr. files: 6300 (10/speaker)
- 8 major American Dialects
- Phonetic and word transcriptions

Keyword	Total nr. occur.	Train	Test	Keyword extracted
artists	14	5	3	6
carry	632	446	166	20
children	25	6	9	10
development	15	3	6	6
house	18	6	6	6
money	25	5	10	10
problem	15	3	6	6
time	37	12	9	16
wash	637	441	176	20
water	640	443	177	20

Desired Result of Pre-processing

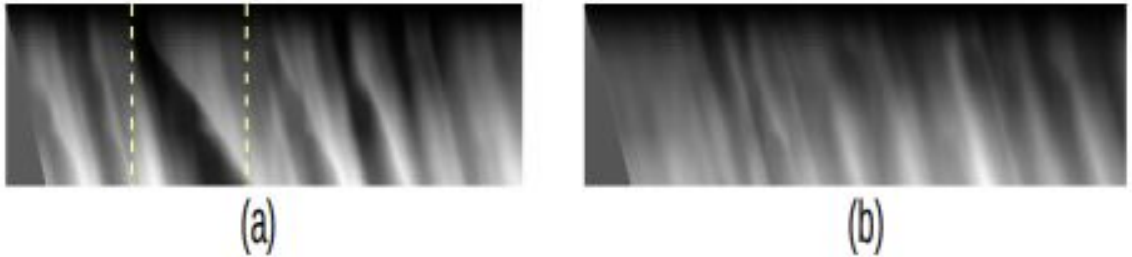


Figure 1: *Warping matrix formed when keyword is (a) present and (b) absent. Highlighted region in (a) corresponds to the region where keyword is present.*

R. Shankar , C.M. Vikram , and S.R.M. Prasanna, "*Spoken Keyword Detection using joint DTW-CNN*"

References

- **[Gan, Henao, 2015]** D. C. L. C. Zhe Gan, Ricardo Henao, “Learning deep sigmoid belief networks with data augmentation,” Journal For Machine Learning, vol. 38, p. 268276, 2015.
- **[Garcia and Gish,2006]** A. Garcia and H. Gish, “Keyword spotting of arbitrary words using minimal speech resources”, in Proc. ICASSP, 123–127, Atlanta, 2006.
- **[Mantena, Achanta, and Prahallad]** G. Mantena, S. Achanta, and K. Prahallad,“Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 5, pp. 946–955, May 2014.
- T. Hazen, W. Shen and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates”, in Proc. ASRU, 2009.
- **[Shankar et al, 2018]** Shankar, Ravi et al. “Spoken Keyword Detection Using Joint DTW-CNN.” Interspeech, 2018
- **[Xu and Sarikaya, 2013]** P. Xu and R. Sarikaya, “Convolutional neural network based triangular CRF for joint intent detection and slot filling,” in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.
- **[S.Y. et al, 2009]** S. Y. et al., The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, Cambridge, 2009.
- **[Zhang and Glass, 2009]** Yaodong Zhang and James Glass, “Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams”, In Proceedings of ASRU, pages 398--403, 2009