Unsupervised acoustic sub-word unit detection for Query-by-Example Spoken Term Detection

Marijn Huijbregts, Mitchell McLaren and David van Leeuwen
Radboud University Nijmegen

marijn.huijbregts@let.ru.nl

Introduction

ASR of interviews with war veterans

- Supervised adaptation of AM
- Word error rates still very high (63%)

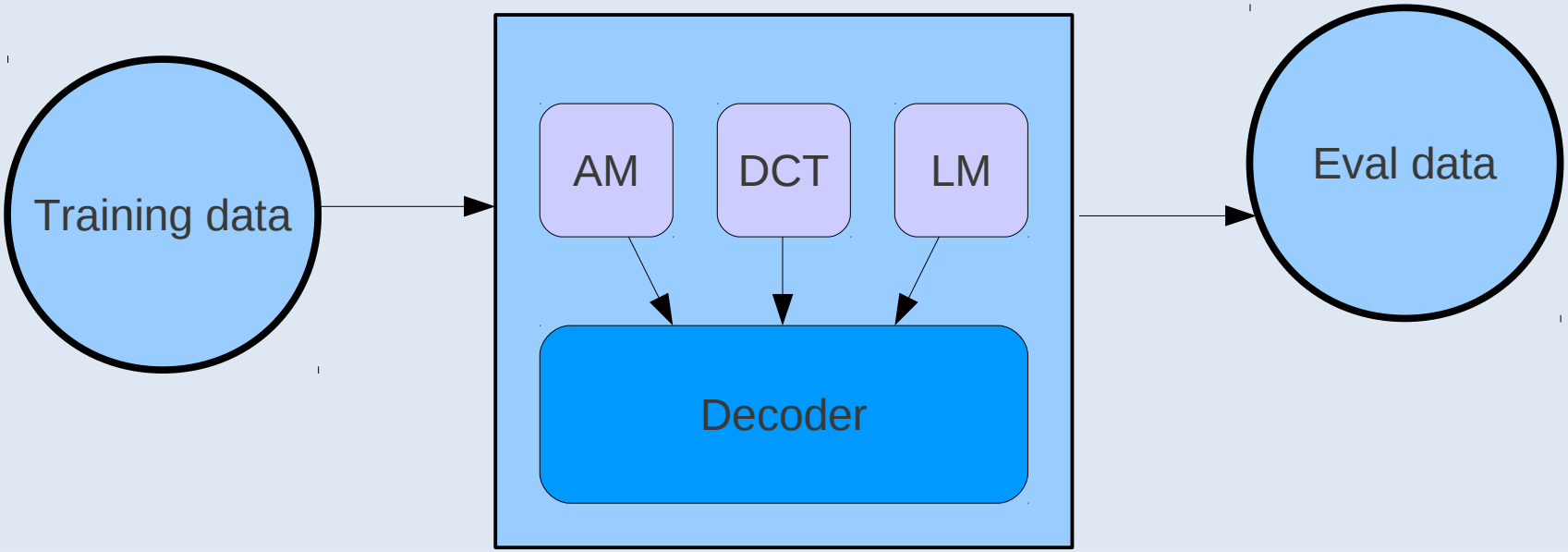
Veteraneninstituut

marijn.huijbregts@let.ru.nl

Radboud Universiteit Nijmegen

Introduction

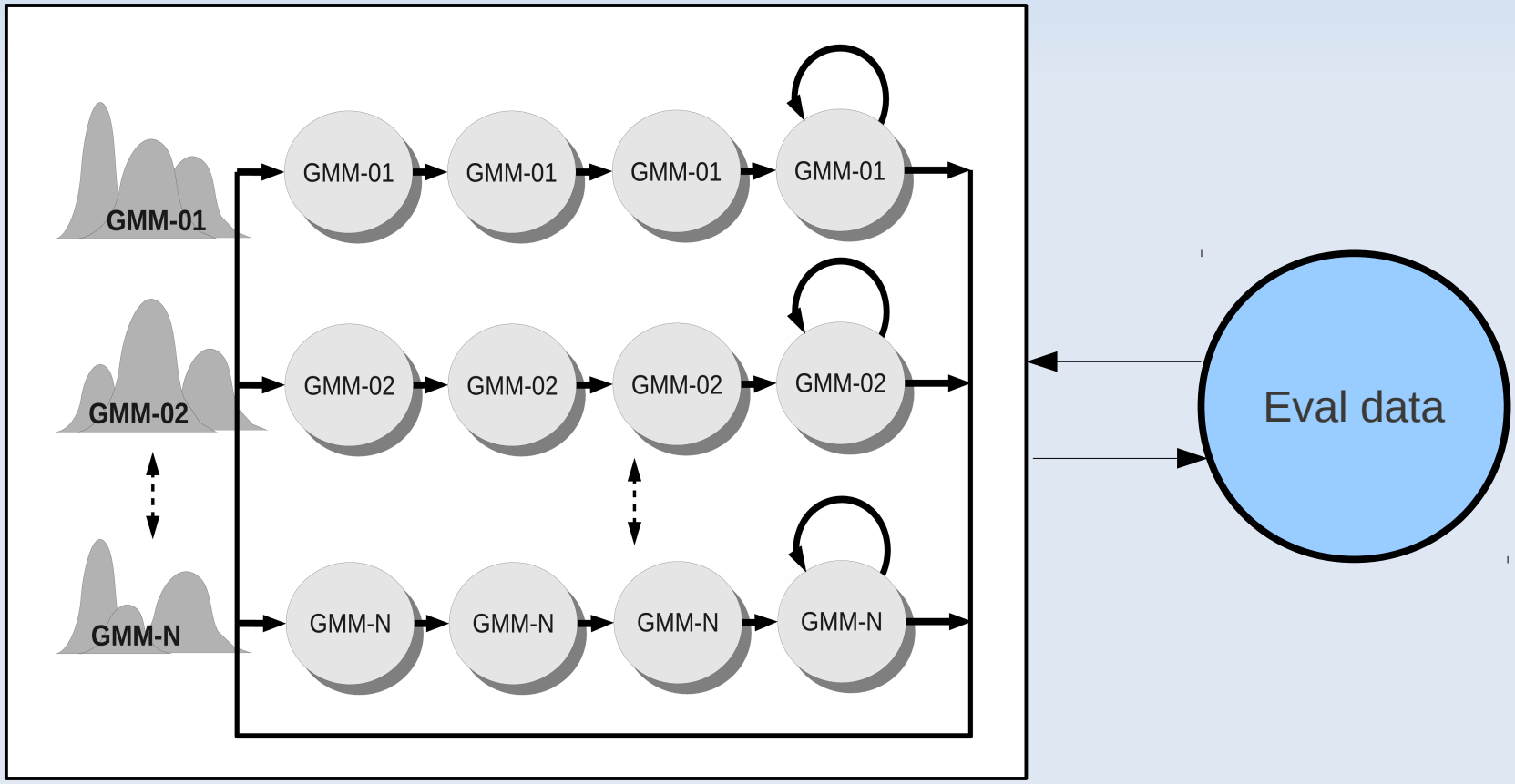
The acoustic train/eval data mismatch



Introduction

The acoustic train/eval data mismatch

- Adressed in our diarization module

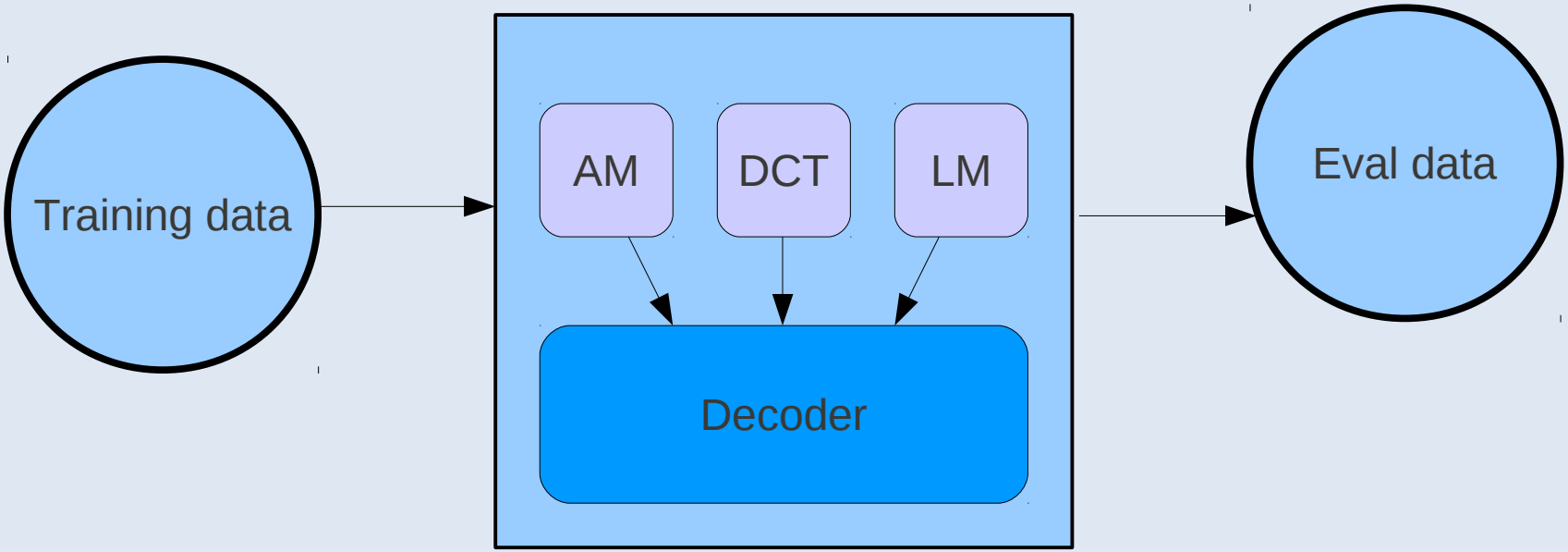


marijn.huijbregts@let.ru.nl

Introduction

The acoustic train/eval data mismatch

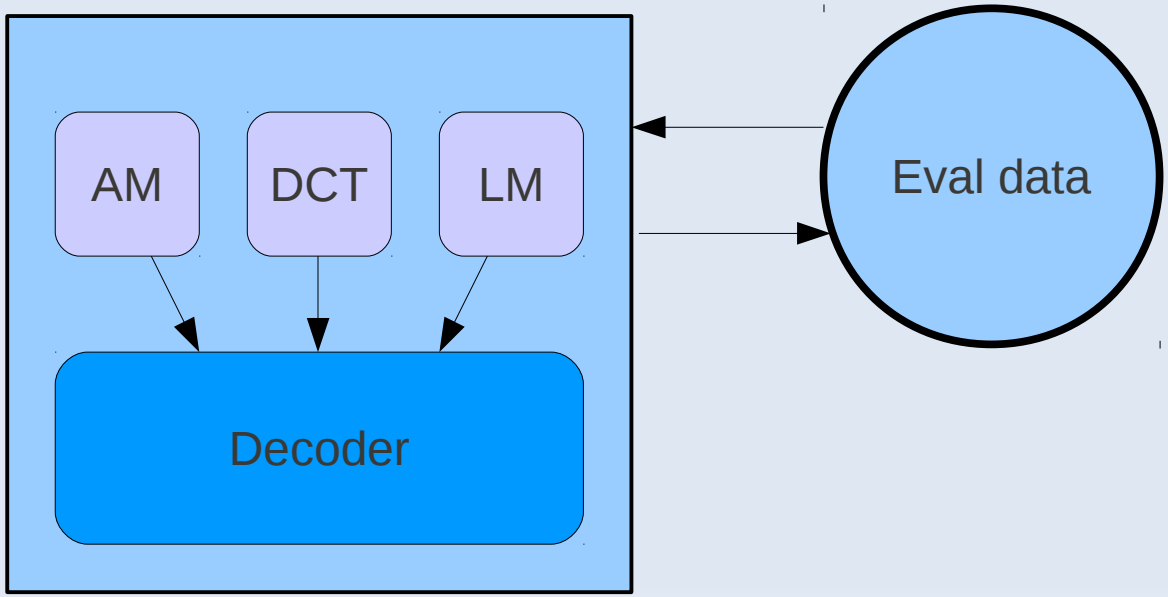
- Adressed in our diarization module
- Can we do the same for ASR?



Introduction

The acoustic train/eval data mismatch

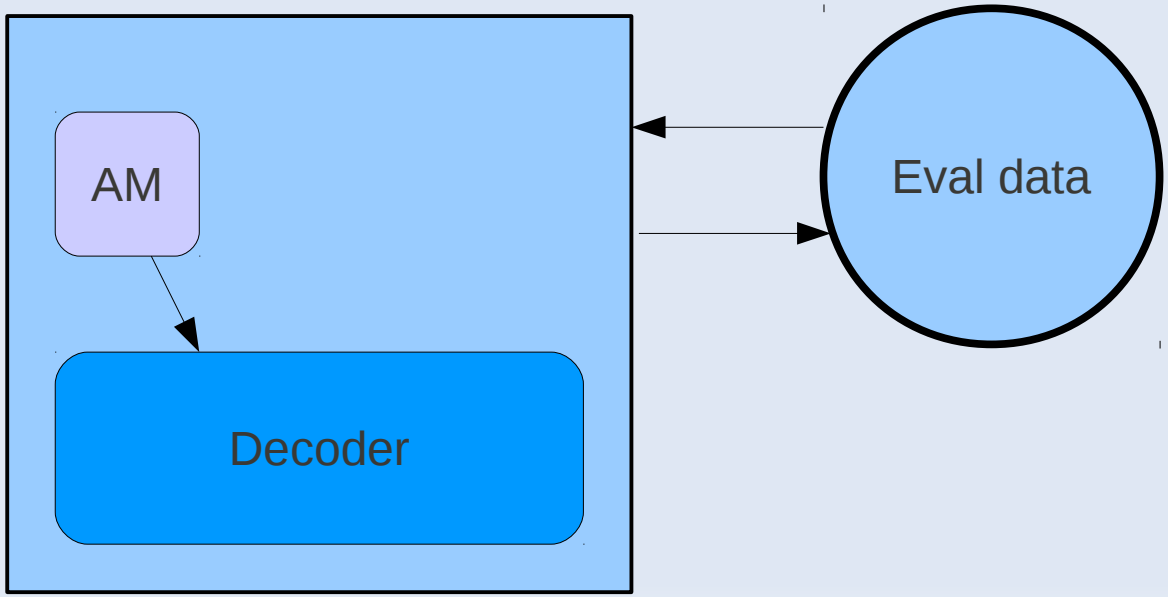
- Adressed in our diarization module
- Can we do the same for ASR?



Introduction

The acoustic train/eval data mismatch

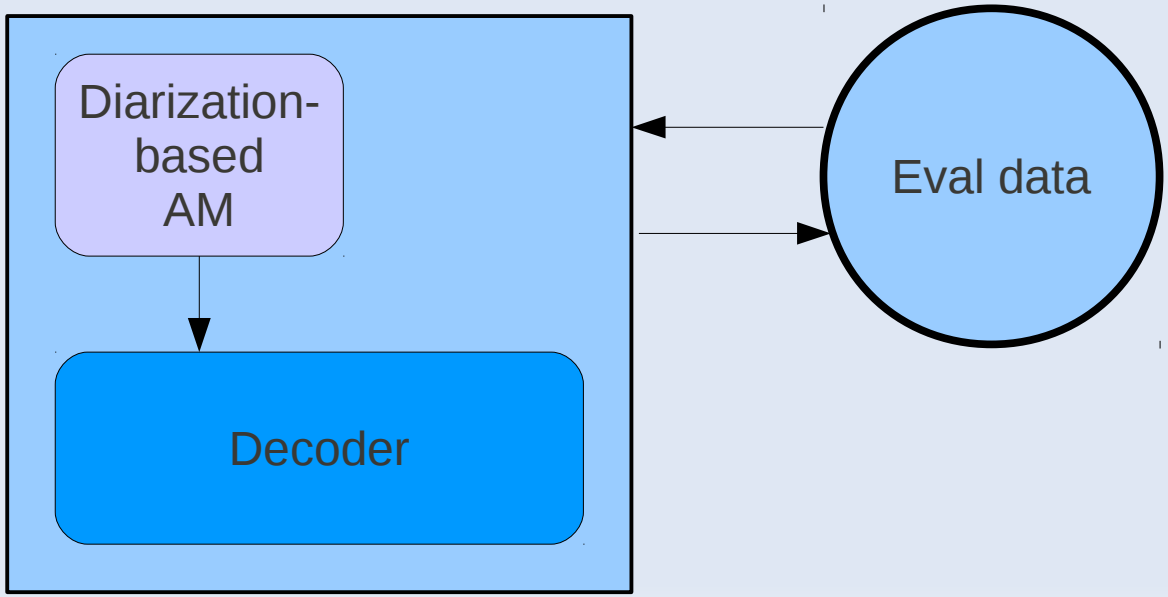
- Adressed in our diarization module
- Can we do the same for AM in ASR?



Introduction

The acoustic train/eval data mismatch

- Adressed in our diarization module
- Can we do the same for AM in ASR?
- Can we use our diarization system for that?



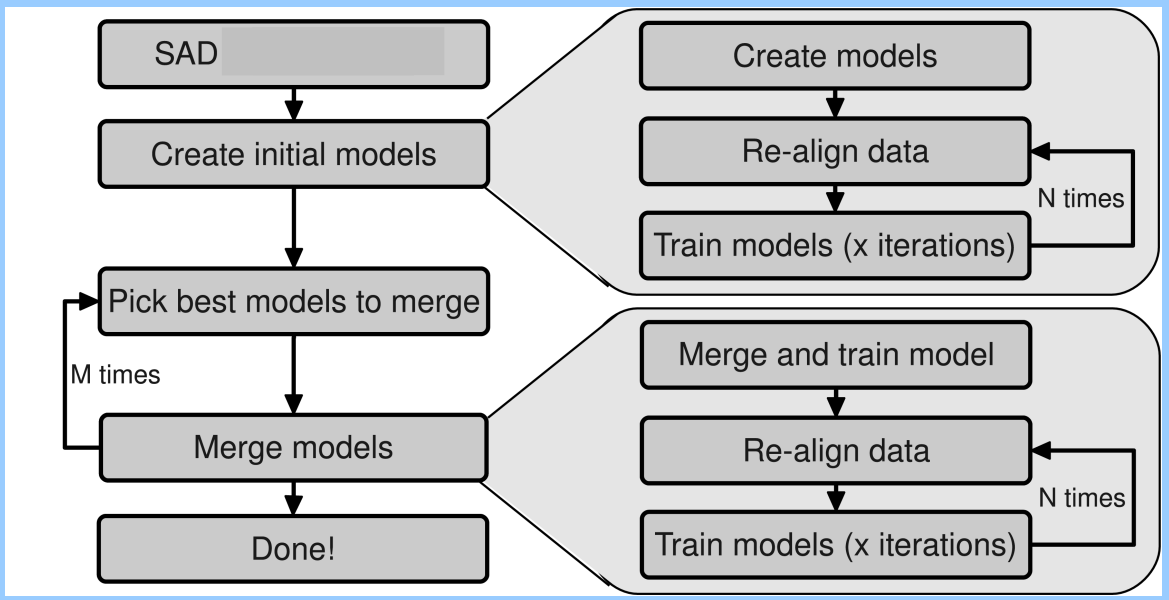
The goal

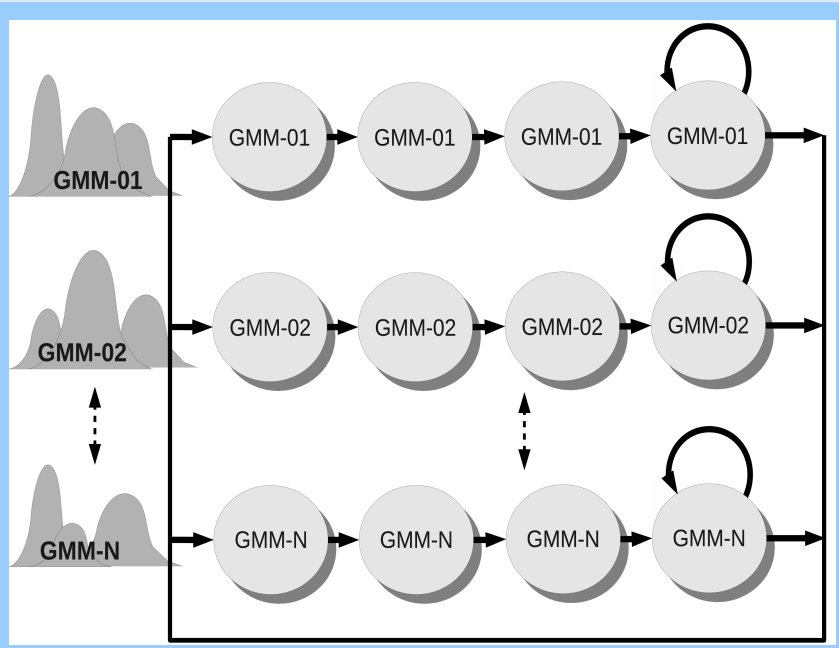
- Automatically segment and cluster an audio recording of a single speaker into *sub-word units.*
- So that these sub-word units can be used to perform ASR.
- First step: these units can be used to perform query-by-example spoken term detection.

Acoustic sub-word units

The diarization system:

- Prevents classification on short-term charactarisics:
  - Minimum duration constraint
  - No delta's in its feature vector





Acoustic sub-word units

Unsupervised acoustic sub-word unit detector (UASUD)

Diarization

- Multiple speakers
- Minimum duration 2.5s
- No delta's
- Initial nr clusters varies
- Stop with BIC

UASUD

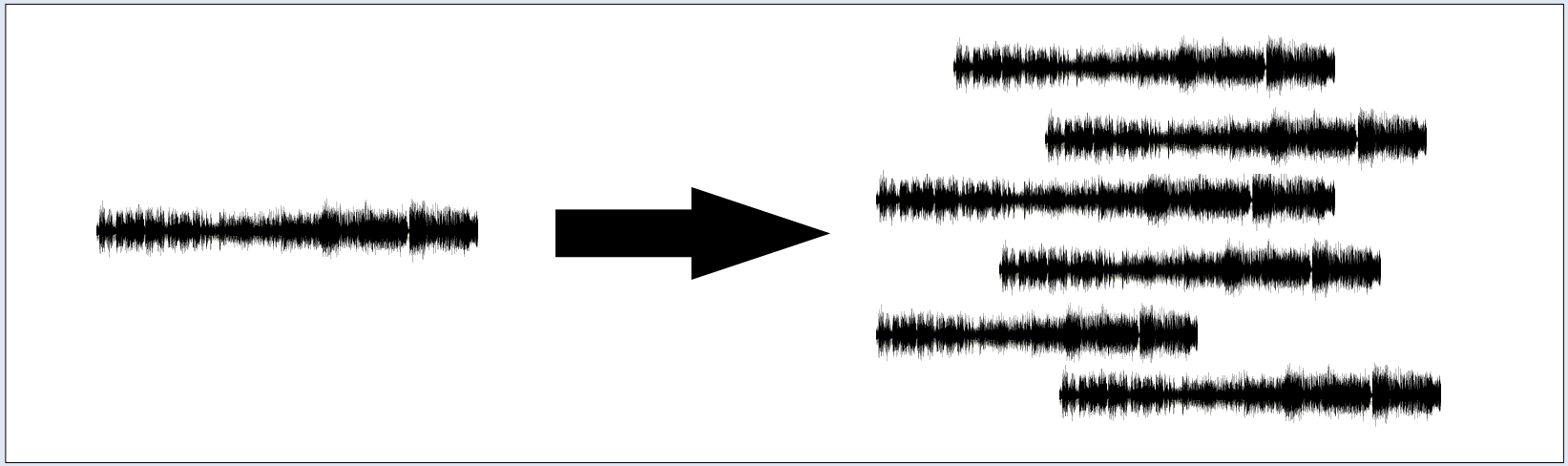
- One speakers
- Min duration 40ms
- Delta's and delta-delta
- 207 initial clusters
- Stop at 57 clusters

marijn.huijbregts@let.ru.nl

Query-by-example

Evaluation on query-by-example spoken term detection

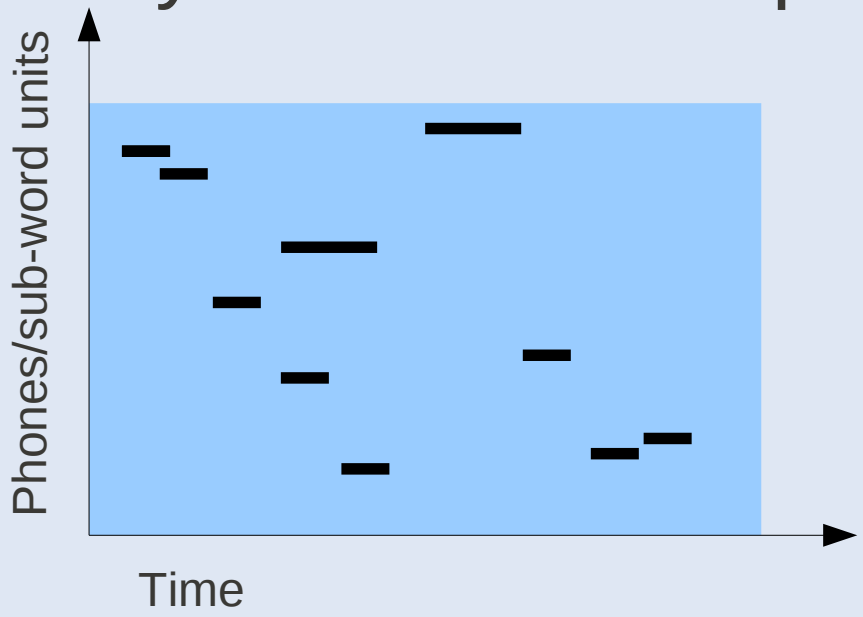
- No DCT or LM available
- Use examples from the audio itself as query
- The system has to obtain all other occurences of the example in the audio

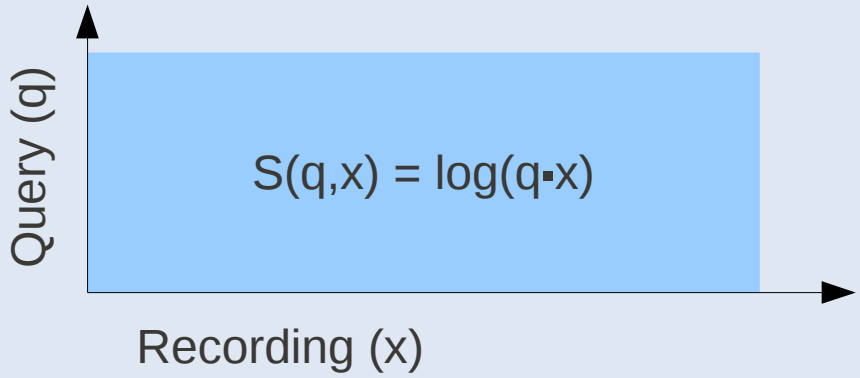


System framework

Hazen, Shen, and White, "Query-by-example spoken term detection using phonetic posteriorgram templates"

- Posteriorgrams
- Similarity matrix (query / recording)
- Dynamic time warping (ratio 2)



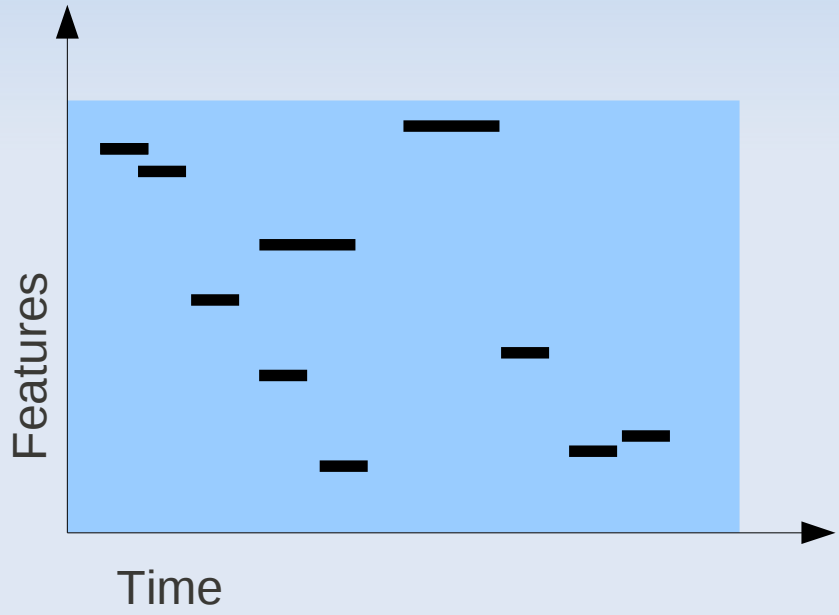


marijn.huijbregts@let.ru.nl

System framework

Four systems:

- UASUD
- Phones
- MFCC
- GMM



Yaodong Zhang and James Glass, “Towards multi-speaker unsupervised speech pattern discovery”

Results

Experiment	MAP on BN	MAP on interviews
MFCC	0.25	0.32
phone	0.27	0.06
GMM	0.27	0.36
UASUD	0.28	0.38

Future work

- Speaker independent sub-word units.
  (similar to cluster linking in large-scale-diarization)
- Automatically find recurring sequences of sub-word units.
- Use the small annotated piece to generate an initial, rough dictionary.