New Technologies for the Information Society
Research Center
University of West Bohemia in Pilsen

Convolutional Neural Network in the Task of Speaker Change Detection

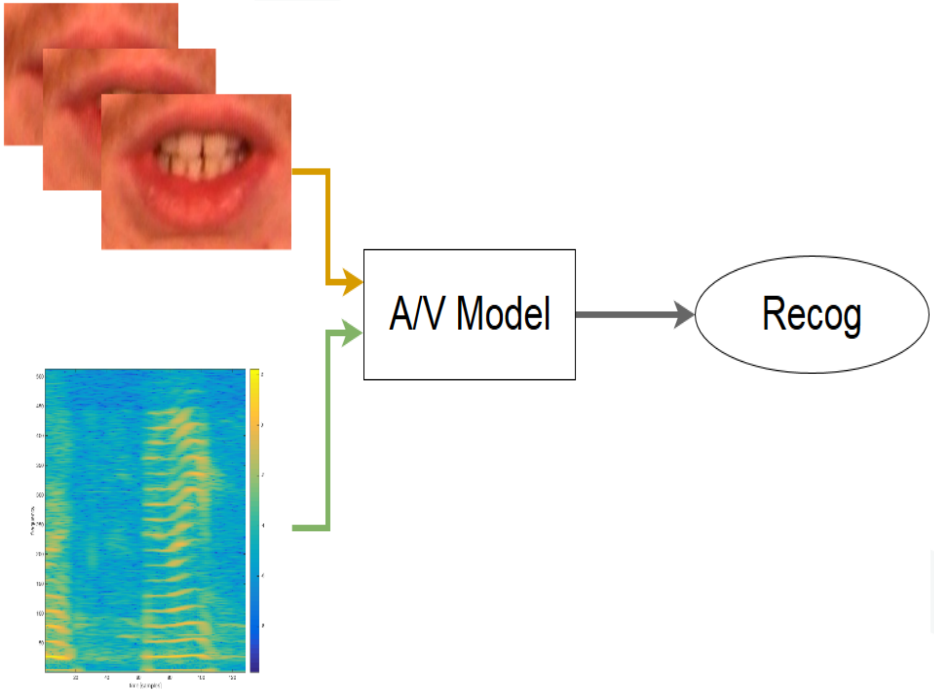
Ing. Marek Hrúz Ph.D.
Ing. Marie Kunešová

Motivation

- Overall goal: Audio-visual model
- Such model will use both modalities for recognition/identification

Motivation

- Overall goal: Audio-visual model
- Such model will use both modalities for recognition/identification

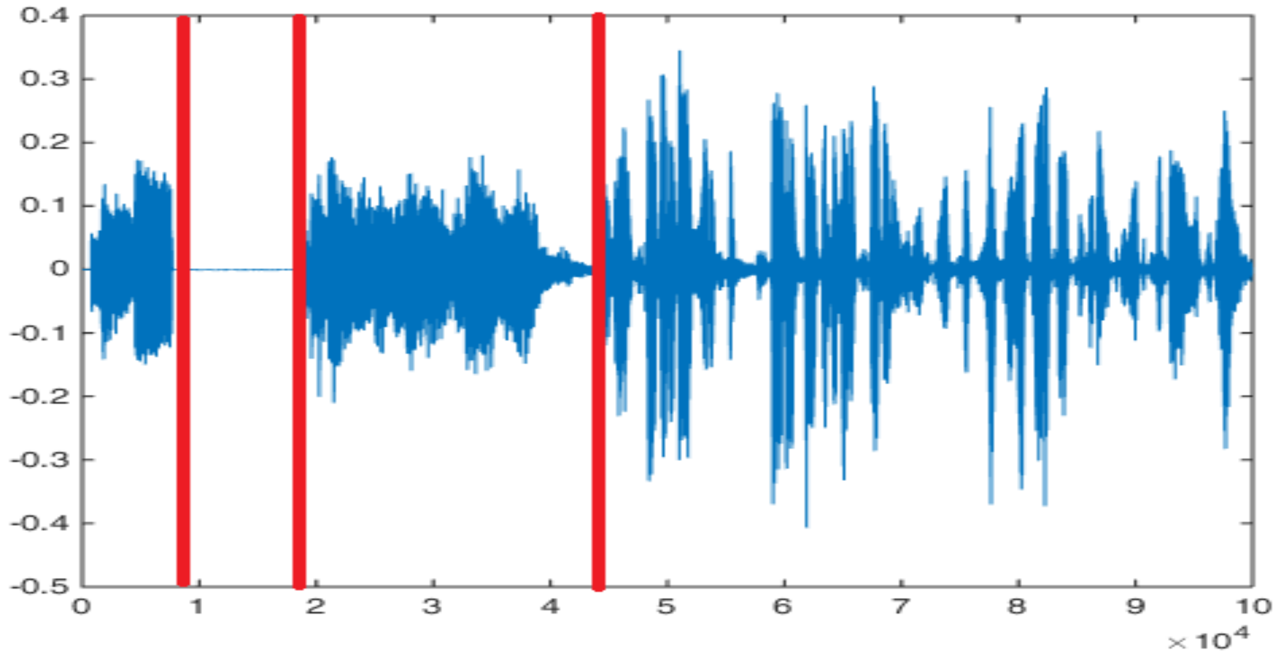


Motivation

- Overall goal: Audio-visual model
- Such model will use both modalities for recognition/identification
- Generally, there can be more modalities
- For the purpose of Human Computer Interfaces:
  - Facial expression
  - Body movement, hand gestures
  - Prosodic analysis of speech

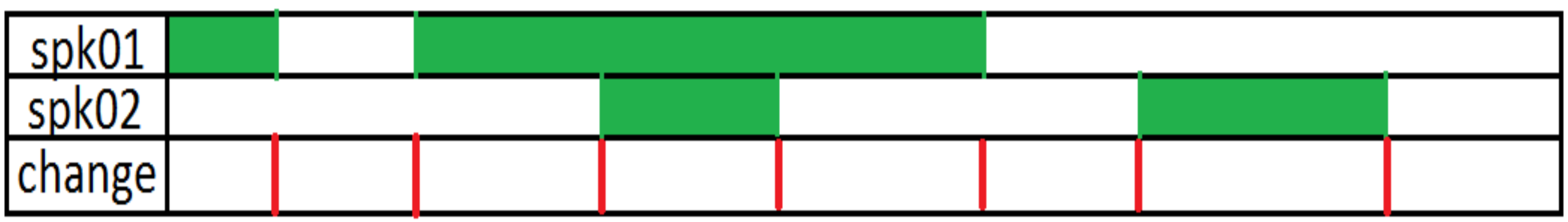
Speaker change detection

- The role of SCD in the big scope is to find segments of A/V data where there is only one speaker present
- SCD can be done on both modalities



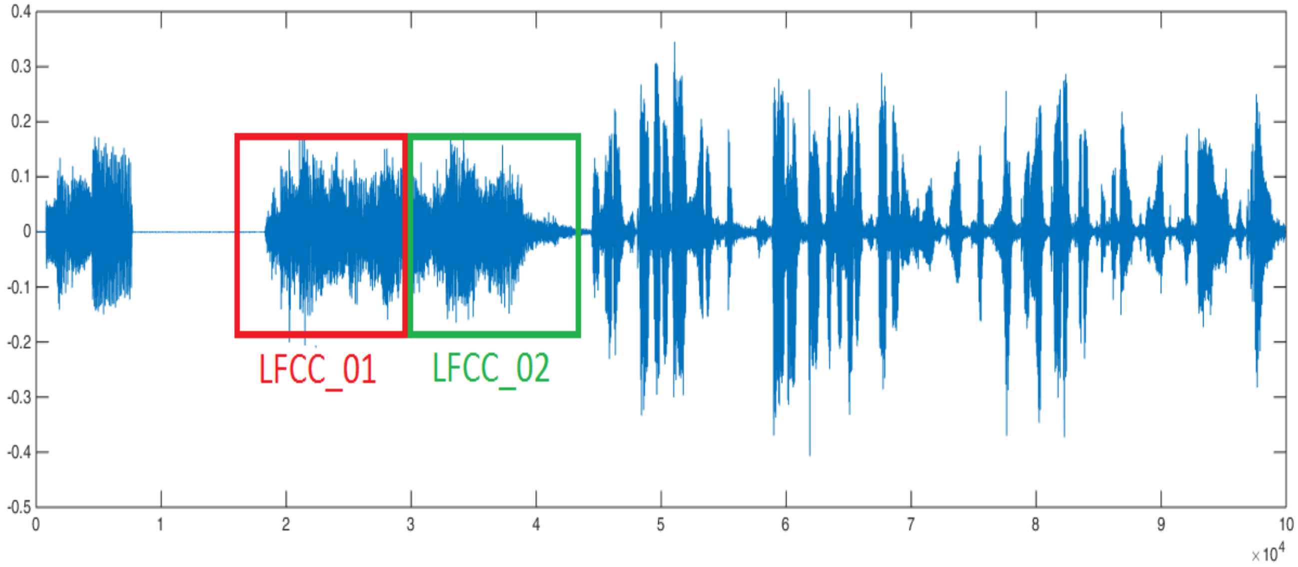
Types of speaker change

- Every time the audio source changes a change occurs
- Spk1-Spk2; Spk1-SIL; Spk1-{Spk1+Spk2}



Speaker Change Detection in the Past

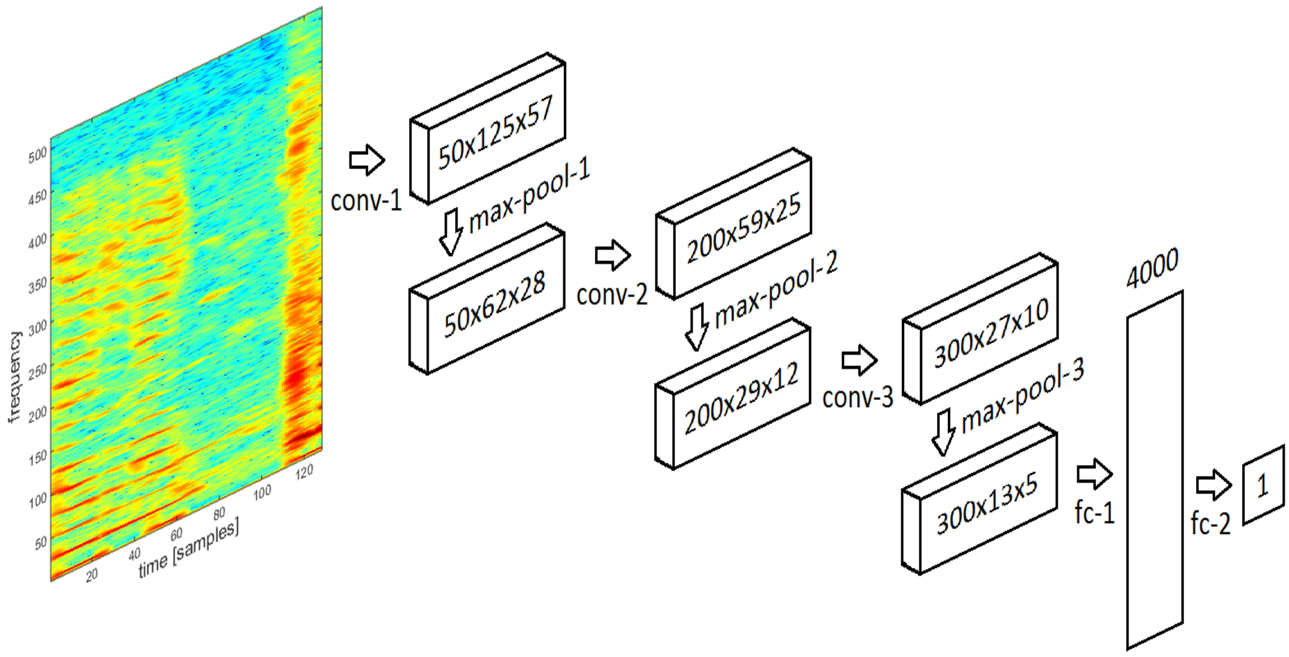
- Most of the past research is based on comparing features extracted from speech segments using a sliding window



- LFCC are modelled as a Gaussian distribution
- The Gaussians are compared via Bayesian Informational Criterion

Convolutional Neural Network

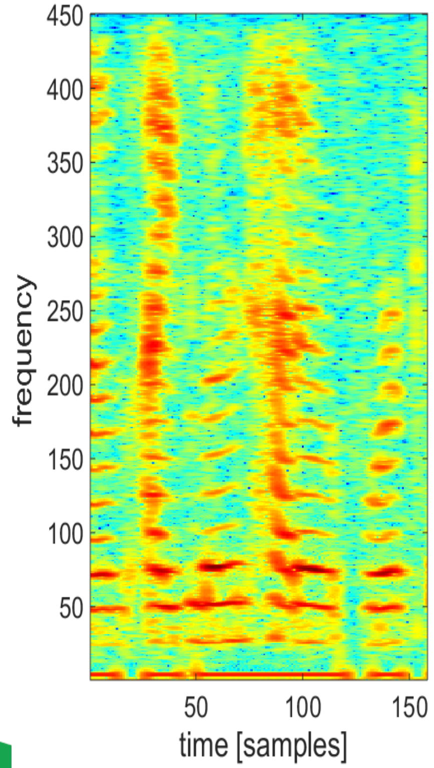
- Because of the success of CNNs in classification and regression we want to test them in the task of SCD



Where is the change?

- The input of the CNN is a spectrogram covering 1.4 seconds of audio

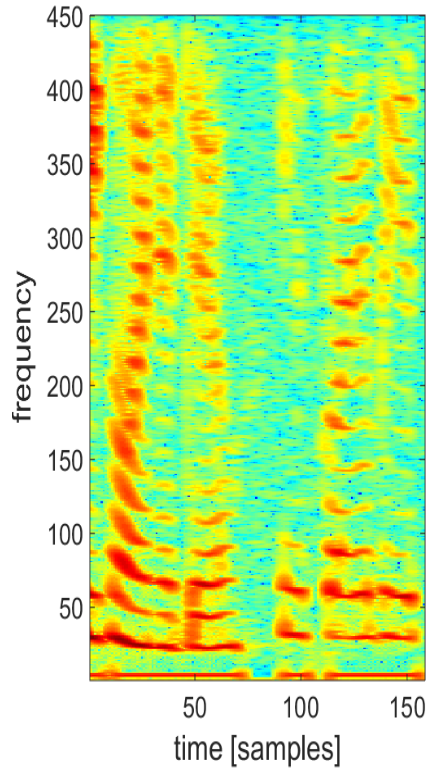
tract segments of one exclusive speaker



- This is a segment with one audio source
- The fundamental frequency is almost the same
- The shapes of the “wrinkles” are consistent

Where is the change?

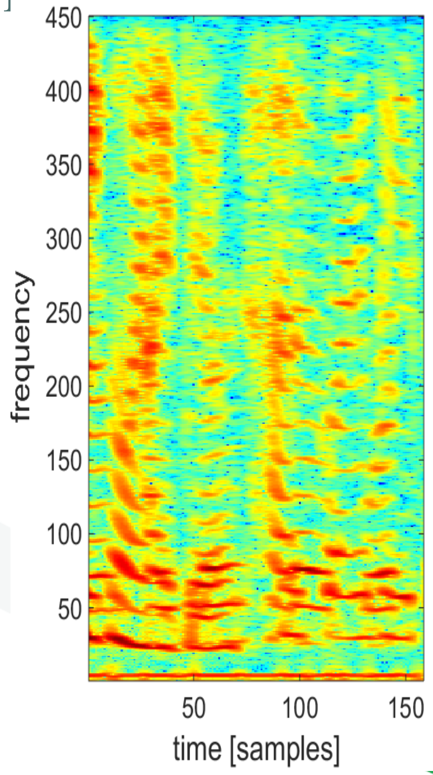
- The input of the CNN is a spectrogram covering 1.4 seconds of audio
- The goal is to extract se ısive speaker



  - In this segment a speaker change is present
  - The fundamental frequency changes
  - The shape characteristics of the “wrinkles” changes

Where is the change?

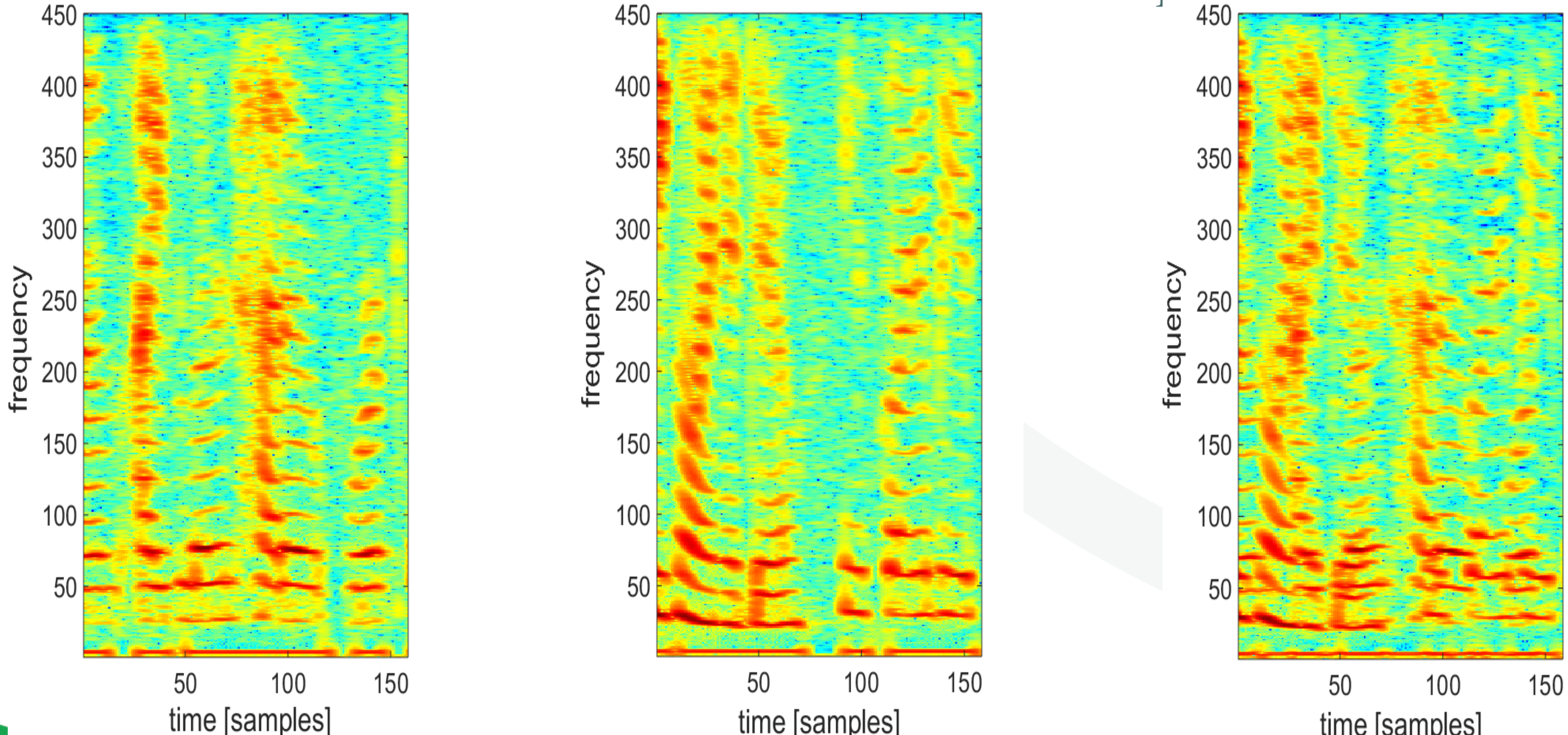
- The input of the CNN is a spectrogram covering 1.4 seconds of audio
- The goal is to extract segments of one exclusive s
  - This segment depicts an overlapped speech
  - There are a lot of non-harmonic frequencies
  - The shapes are chaotic



Where is the change?

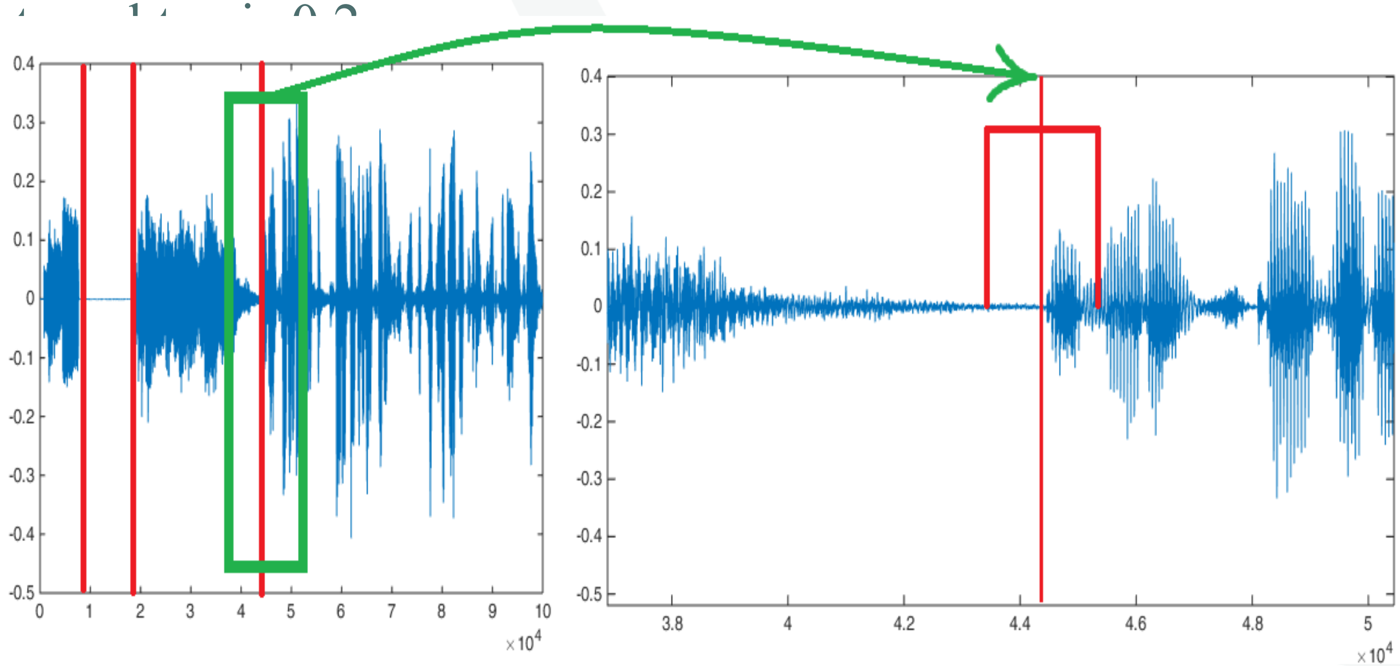
- The input of the CNN is a spectrogram covering 1.4 seconds of audio

ract se isive s



Where is the change?

- The precision of the border of the segment is “noisy”
- The labels should reflect that – instead of one instance it is an interval



CNN architecture

- The shapes of the kernels in the first layer are chosen with the shapes of the high energy wrinkles in mind

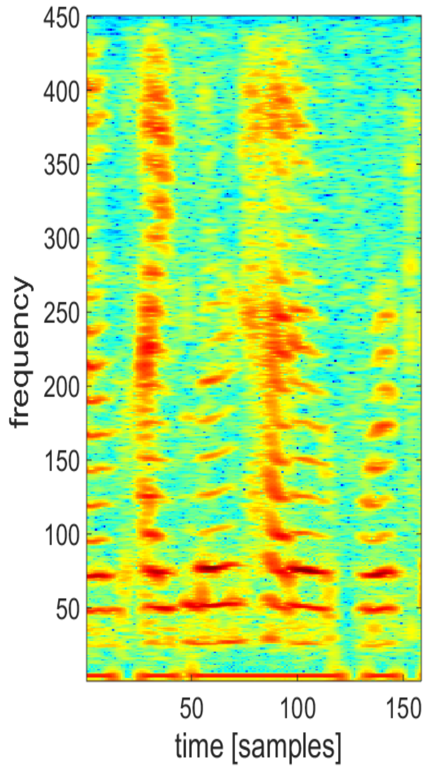


Table 1. Summary of the architecture of the CNN.

Layer	Kernels	Size	Shift
Convolution	50	16 x 8	2 x 2
Max pooling		2 x 2	2 x 2
Batch Norm			
Convolution	200	4 x 4	1 x 1
Max pooling		2 x 2	2 x 2
Batch Norm			
Convolution	300	3 x 3	1 x 1
Max pooling		2 x 2	2 x 2
Batch Norm			
Fully Connected	4000		
Fully Connected	1		

CNN Training

- Using Keras with Theano backend
- Stochastic Gradient Descent
- Batch size - 64
- Step-size learning rate
- Nesterov momentum
- In later stages RMSProp for fine-tuning
- Initialization: K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”, Feb 2015.

Experiment

- CallHome corpus – 8 kHz, telephone, wild speech, annotated
- We compare CNN to the baseline BIC method
- Each segment of 1.4 seconds is regressed to the interval <0; 1>
- Comparison according to DET curves (with linear axes)
- Training data – 5 hr 48 min – 35 conversations
- Testing data – 11 hr 20 min – unheard speakers – 77

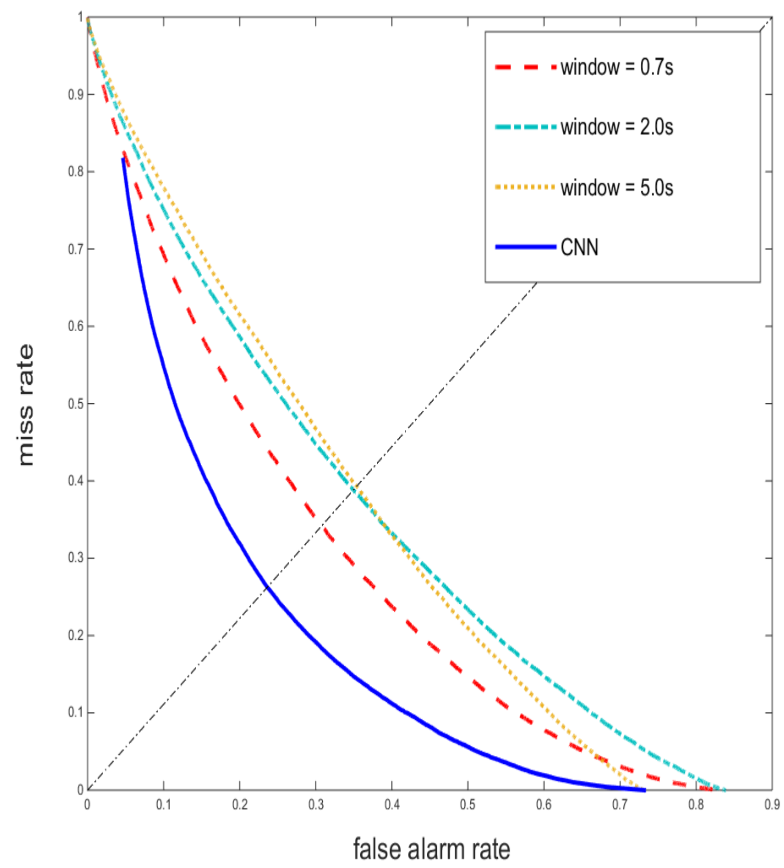
Speaker change detection - Results

- BIC baseline system
- 20 LFCC + delta

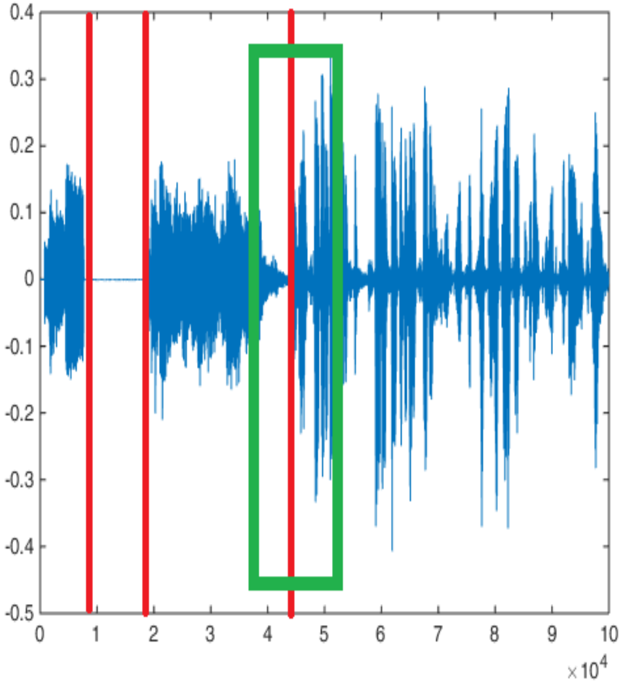
Table 2. EER values for different systems.

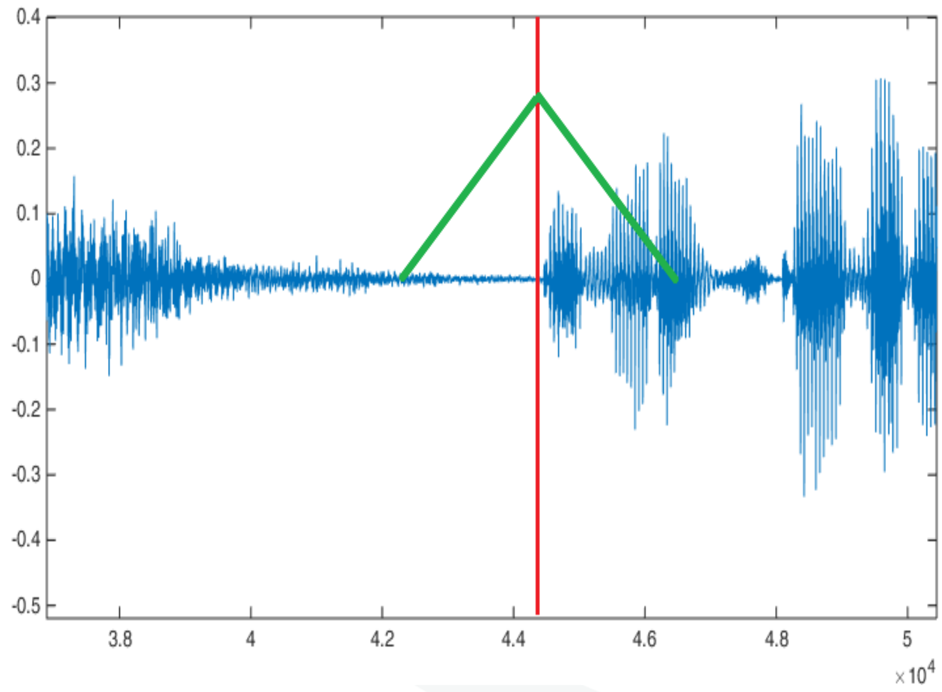
System	BIC 0.7	BIC 2.0	BIC 5.0	CNN
EER	0.3229	0.3679	0.3704	**0.2482**

- CNN – binary labelling
- Another type of labelling?



Fuzzy labelling



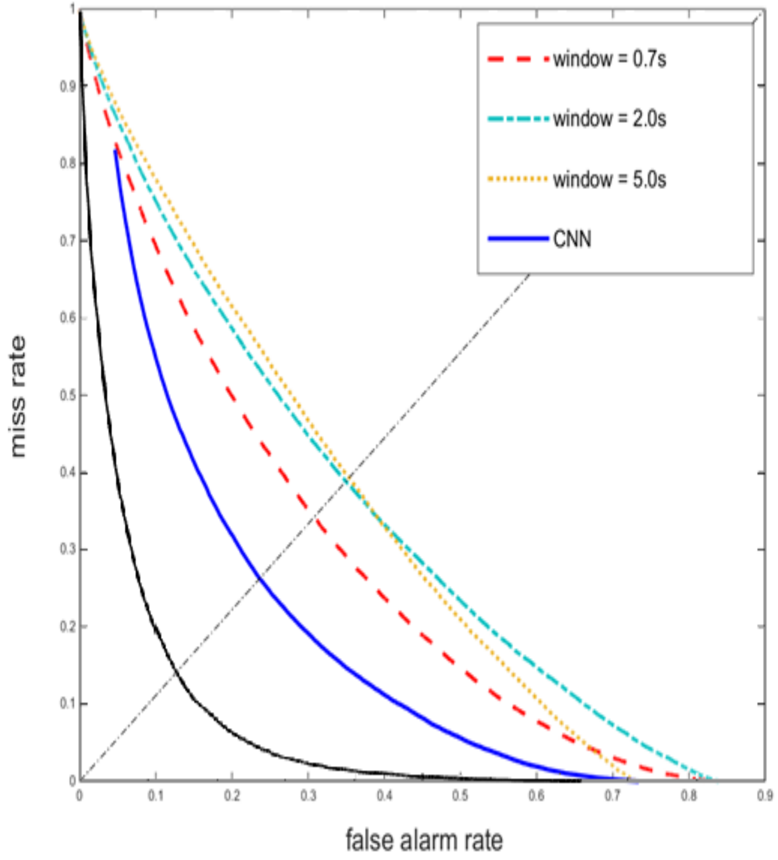


Fuzzy labelling results

- Even better results
- EER = 0.1405

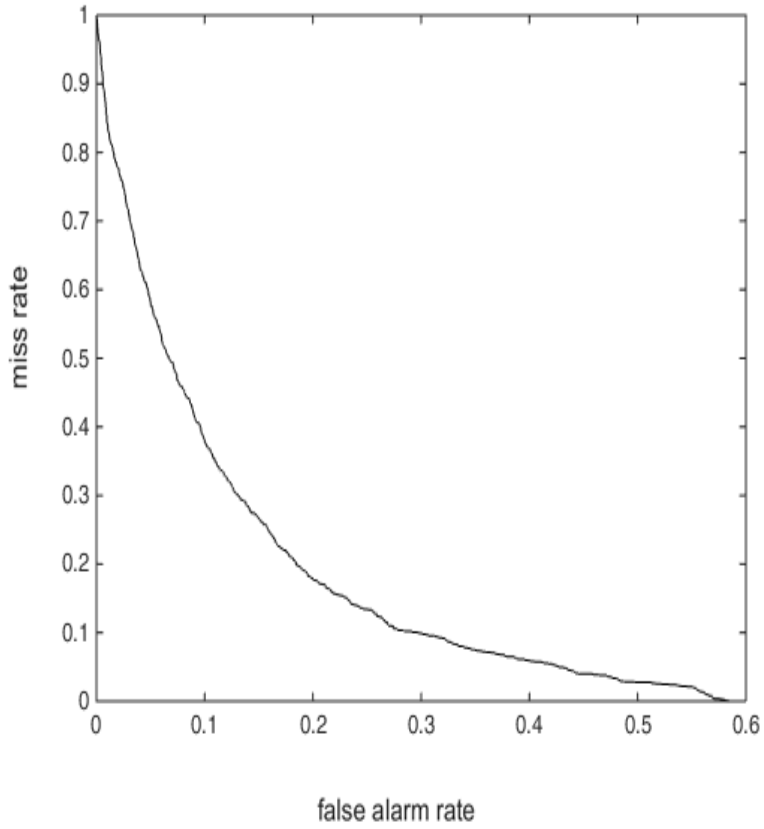
Table 2. EER values for different systems.

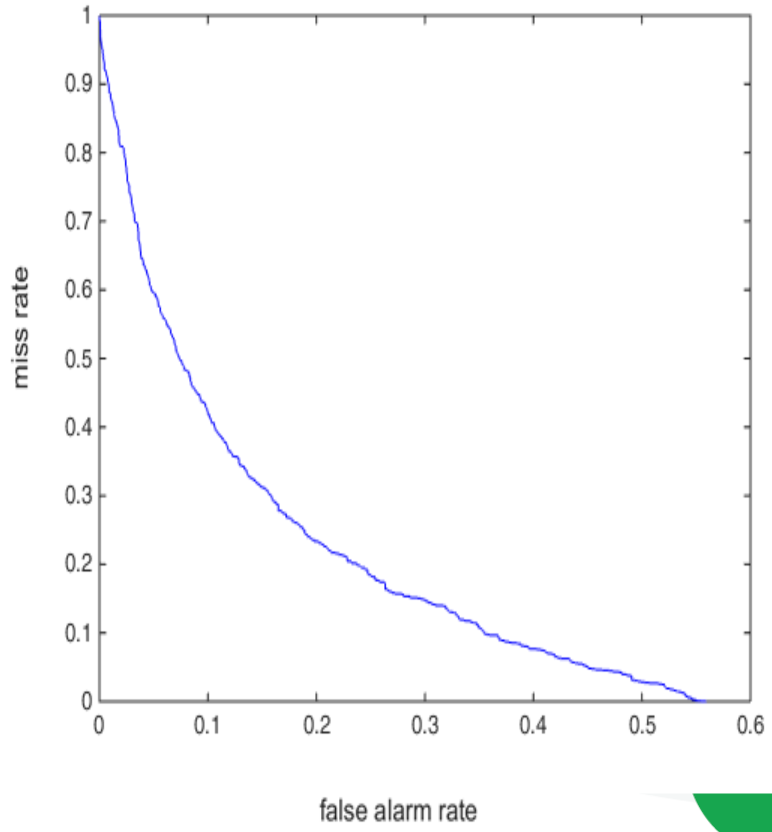
System	BIC 0.7	BIC 2.0	BIC 5.0	CNN
EER	0.3229	0.3679	0.3704	**0.2482**



Czech language data

- EER = 0.1908 (male – female)   EER = 0.2166 (male – male)





THANK YOU FOR YOUR ATTENTION