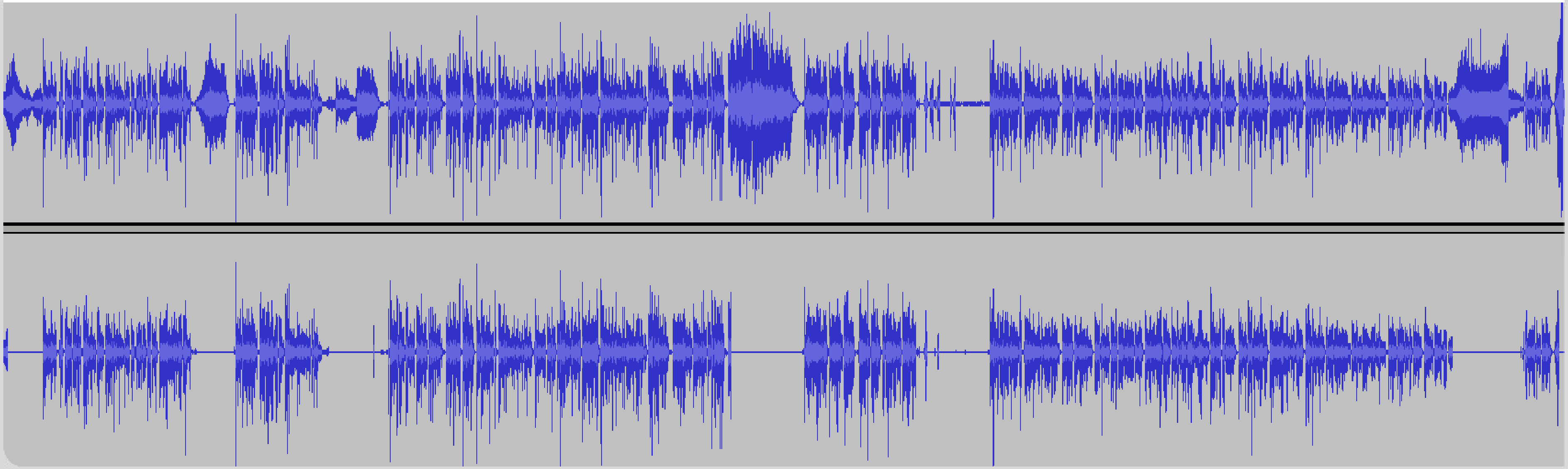
Audio Segmentation For Robust Real-Time Speech Recognition Based on Neural Networks

Micha Wetzel, Matthias Sperber, Alex Waibel

Institute for Anthropomatics and Robotics,
Karlsruhe Institute of Technology, Germany



Problems in Speech Recognition

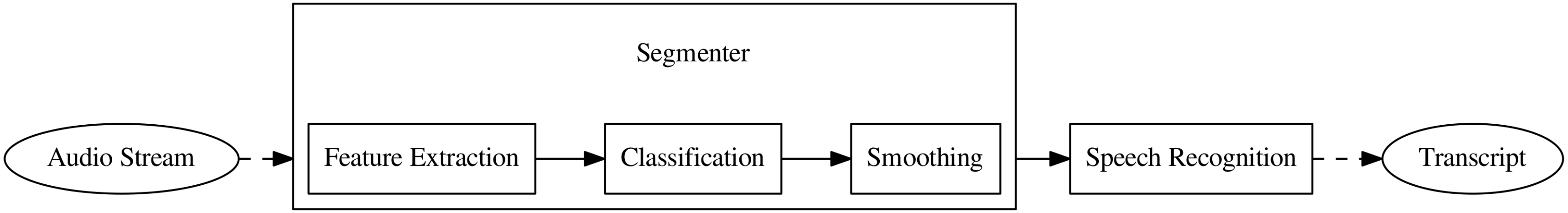
- Garbage in, garbage out
  - „ Thank you you You Can you you you ...“
  - „ you see you may and you see that them ...“
  - „ if if if if if if f F f..“
- Causes high latency

Challenges

- Real-time
  - Limited temporal information
  - Fast algorithm needed
- Classifying speech correctly
- No clear segment borders
- Humans need ~200ms to classify reliably[1]

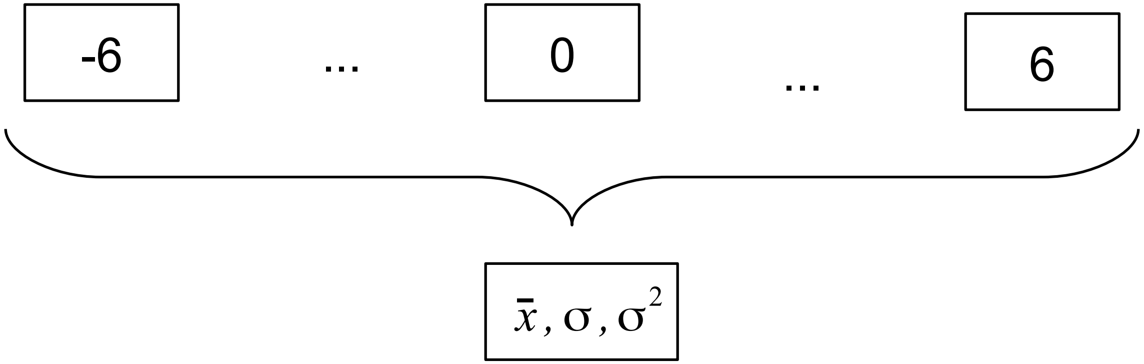
[1] H. Harb et al. Signal Processing and Its Applications Vol 2. 2003

Approach


Segmenter
Audio Stream
Feature Extraction
Classification
Smoothing
Speech Recognition
Transcript

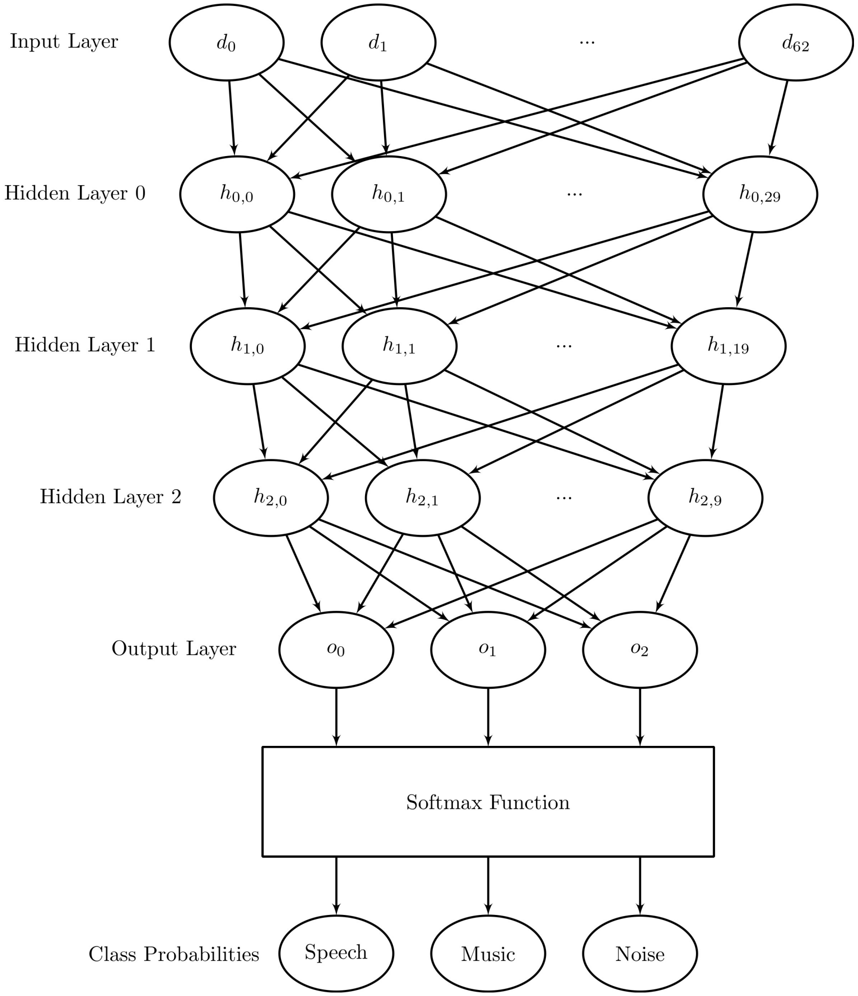
Feature Extraction

- 10 ms frames
- MFCC and ZCR
- 13 frames stacked
- Dimensionality reduction

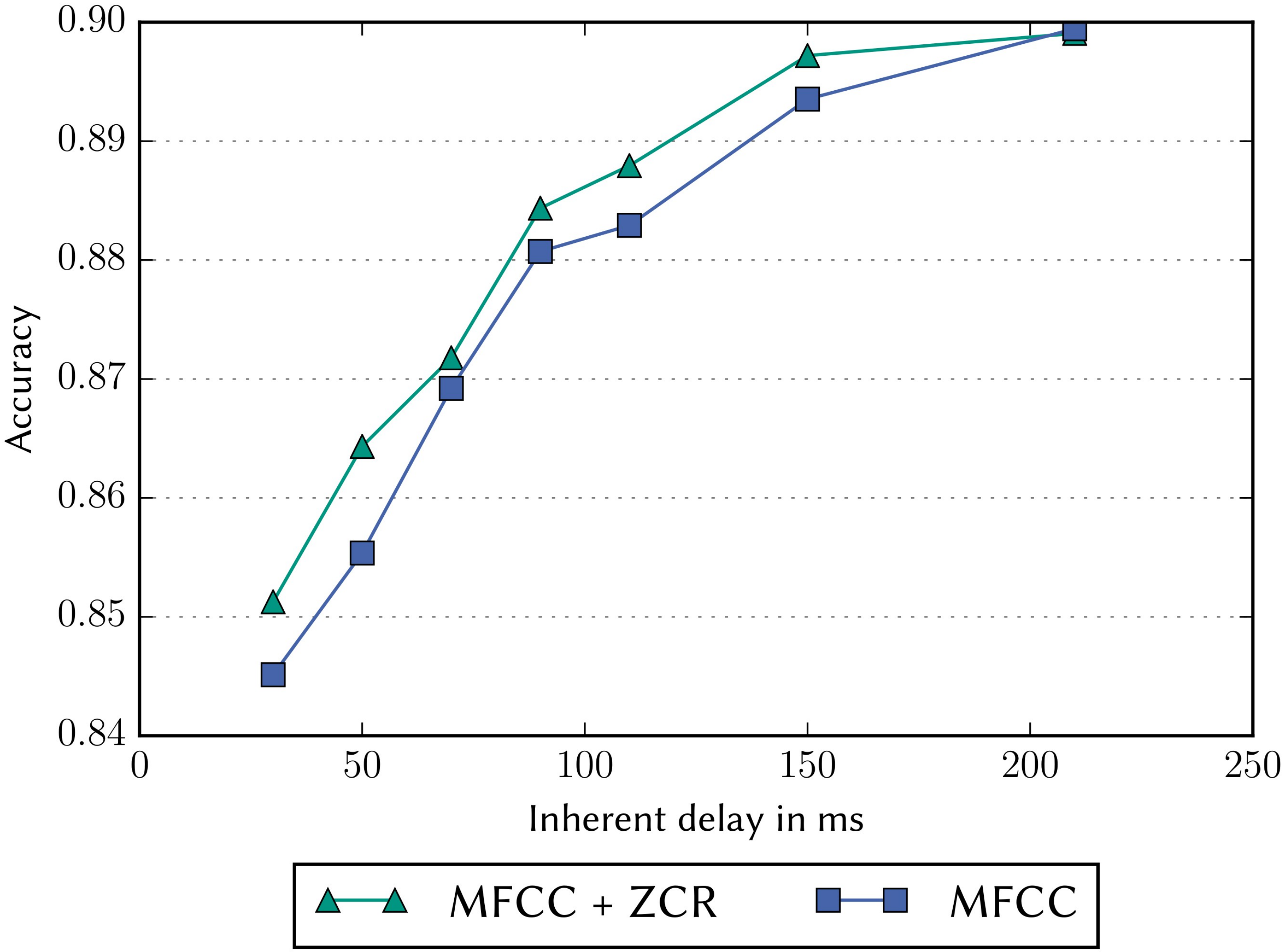


Classification

- Multilayer perceptron
- < 1ms computational latency
- ~70 ms inherent latency
- 87 % accuracy
- 2Class: 95 % accuracy



Classification Results



Classification - Problems

- Class fluctuation
- Misclassifications

  → ASR performs worse

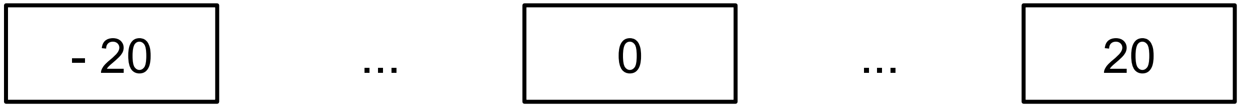
Smoothing

- Need more temporal information
- Class off adjacent frames is correlated
- Make use of knowledge about the past

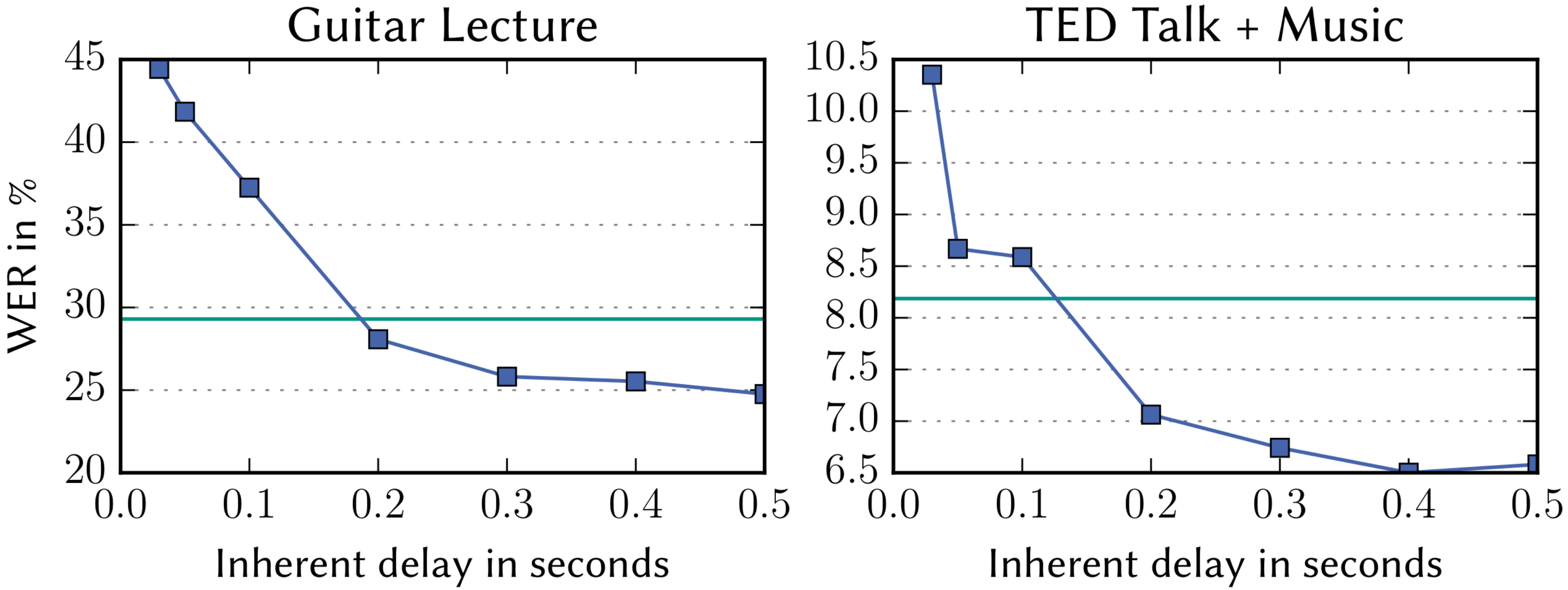
Mode Smoothing

- Mode of adjacent frames
- Remove misclassifications
- Additional temporal information
  → additional 200 ms latency

- Erosion & dilation tested but not needed



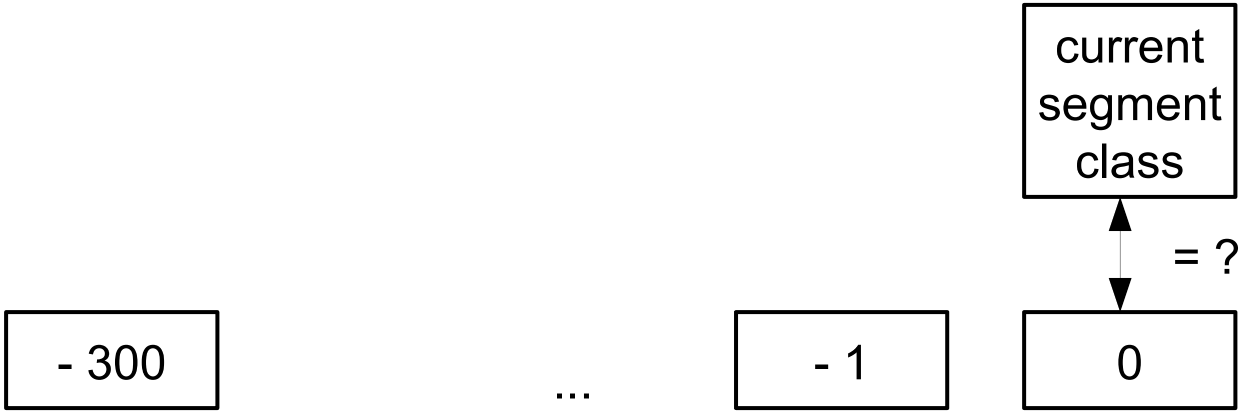
Mode - Results



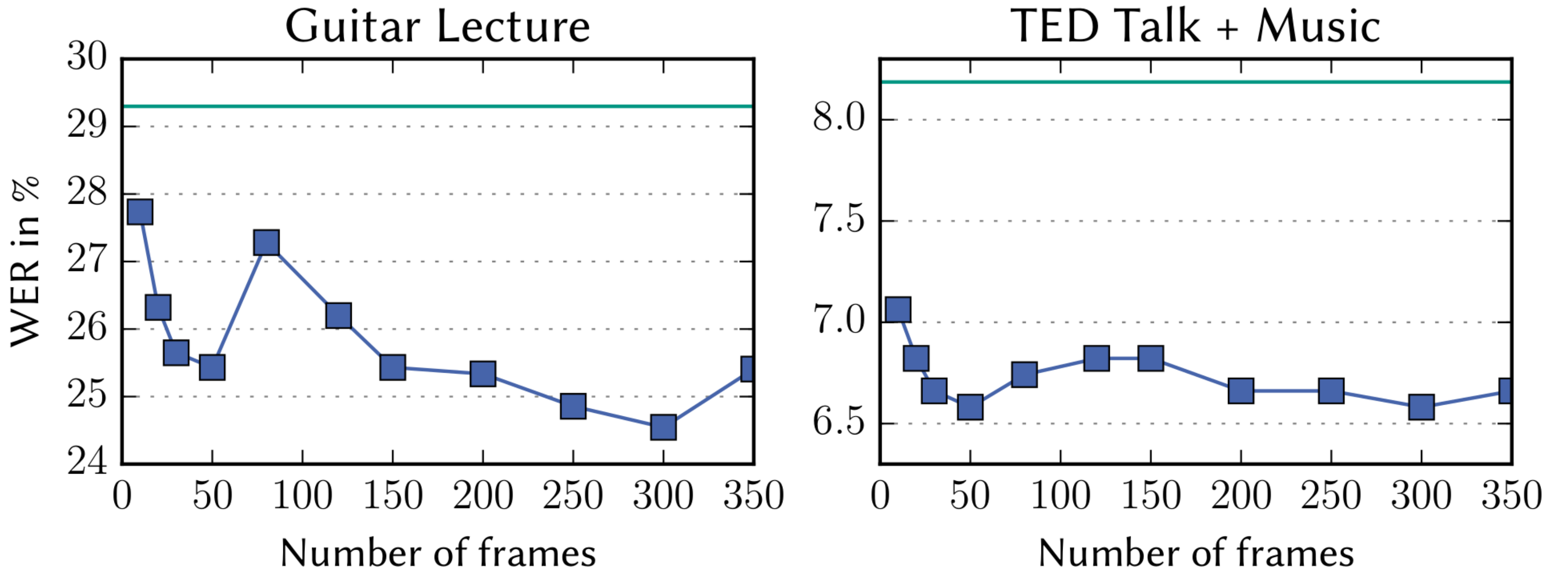
- Tradeoff latency ↔ accuracy

Minimum Change Support Smoothing

- Use knowledge about past
- Create big segments
- Favour speech

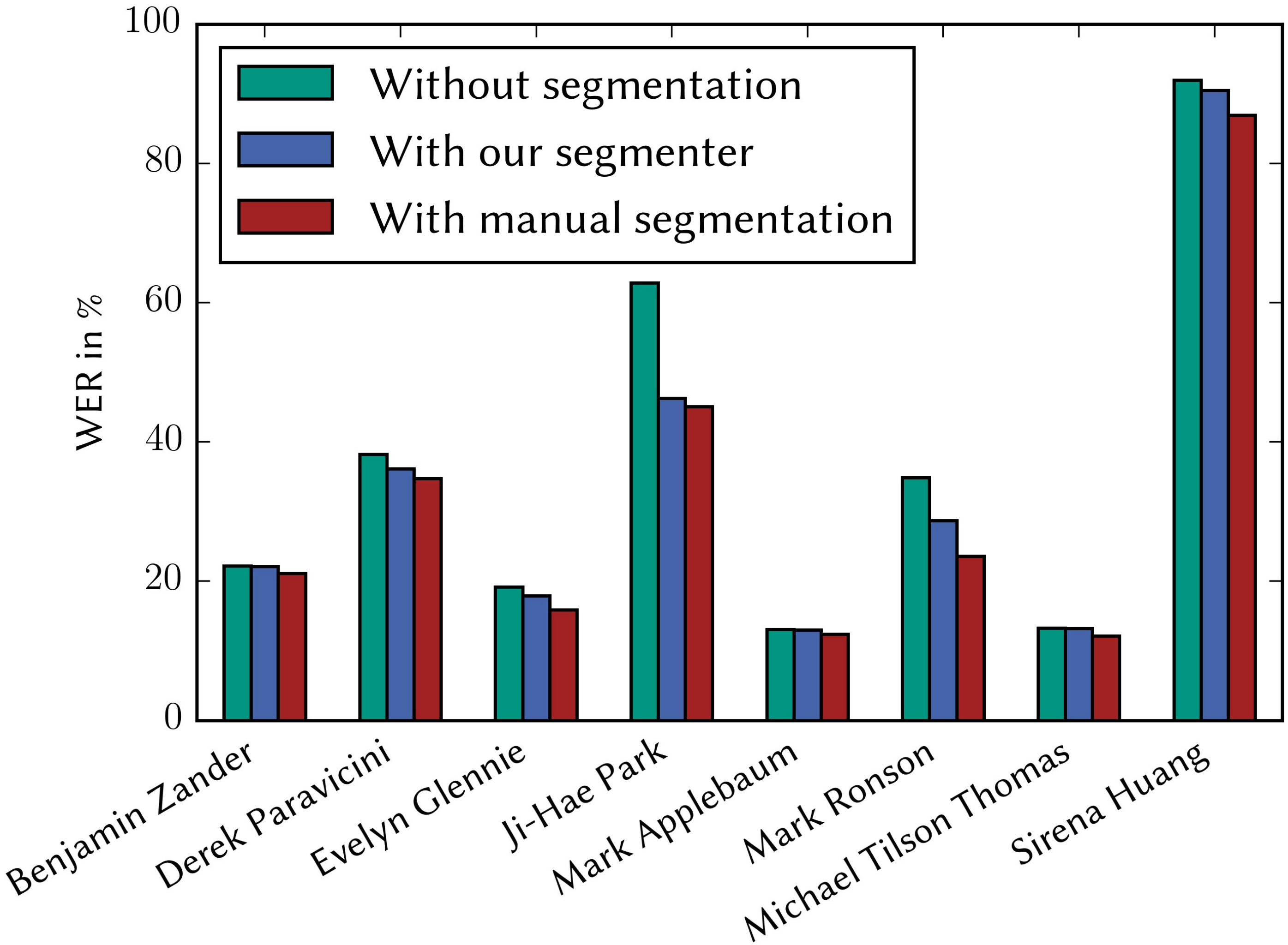


Minimum Change Support - Results



- With mode smoothing
- No additional latency

Segmentation Results



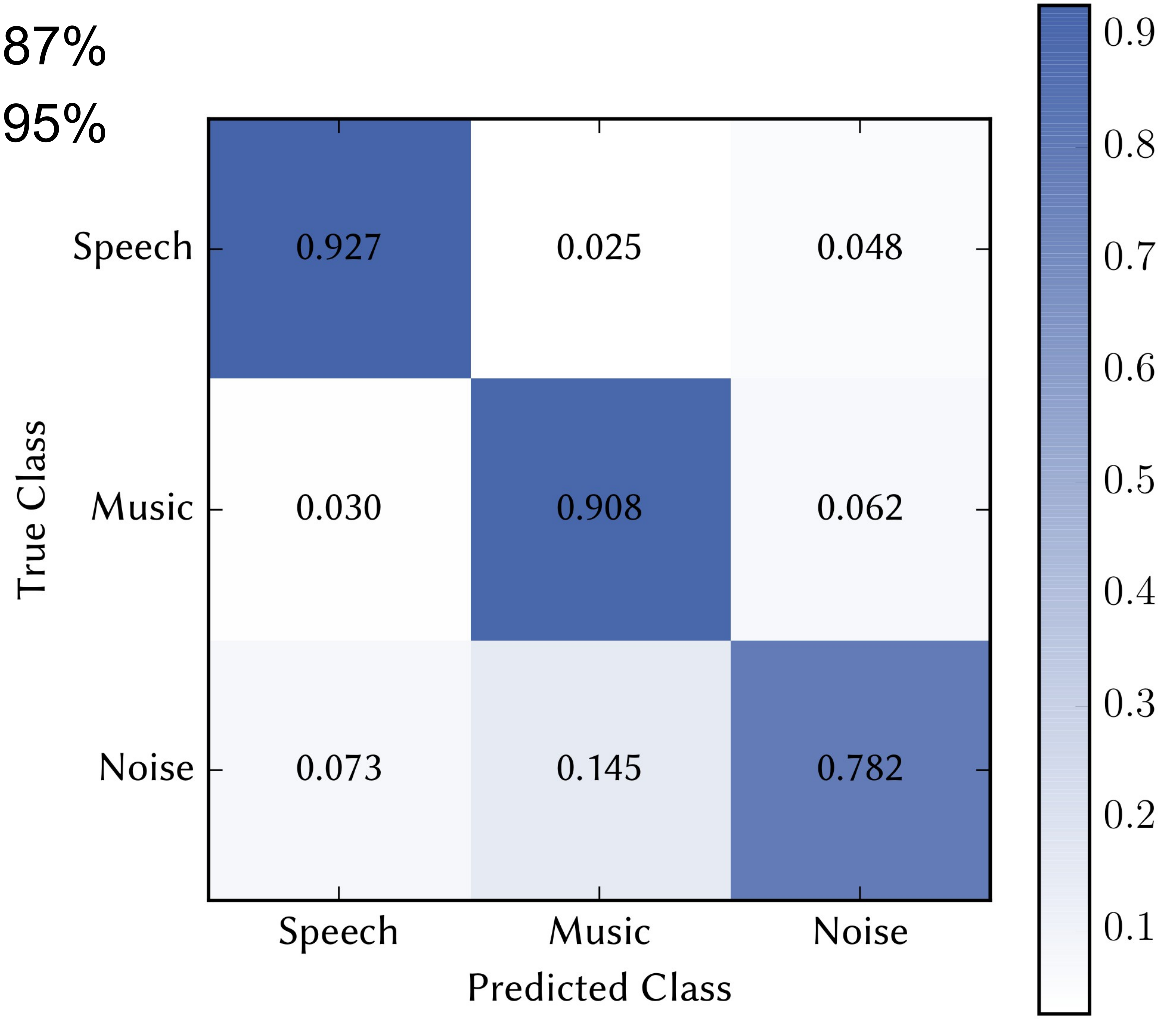
Summary

- Fast computation
- ~270 ms latency
- Removes 39 % of segmentation based errors
- Language independent

The End

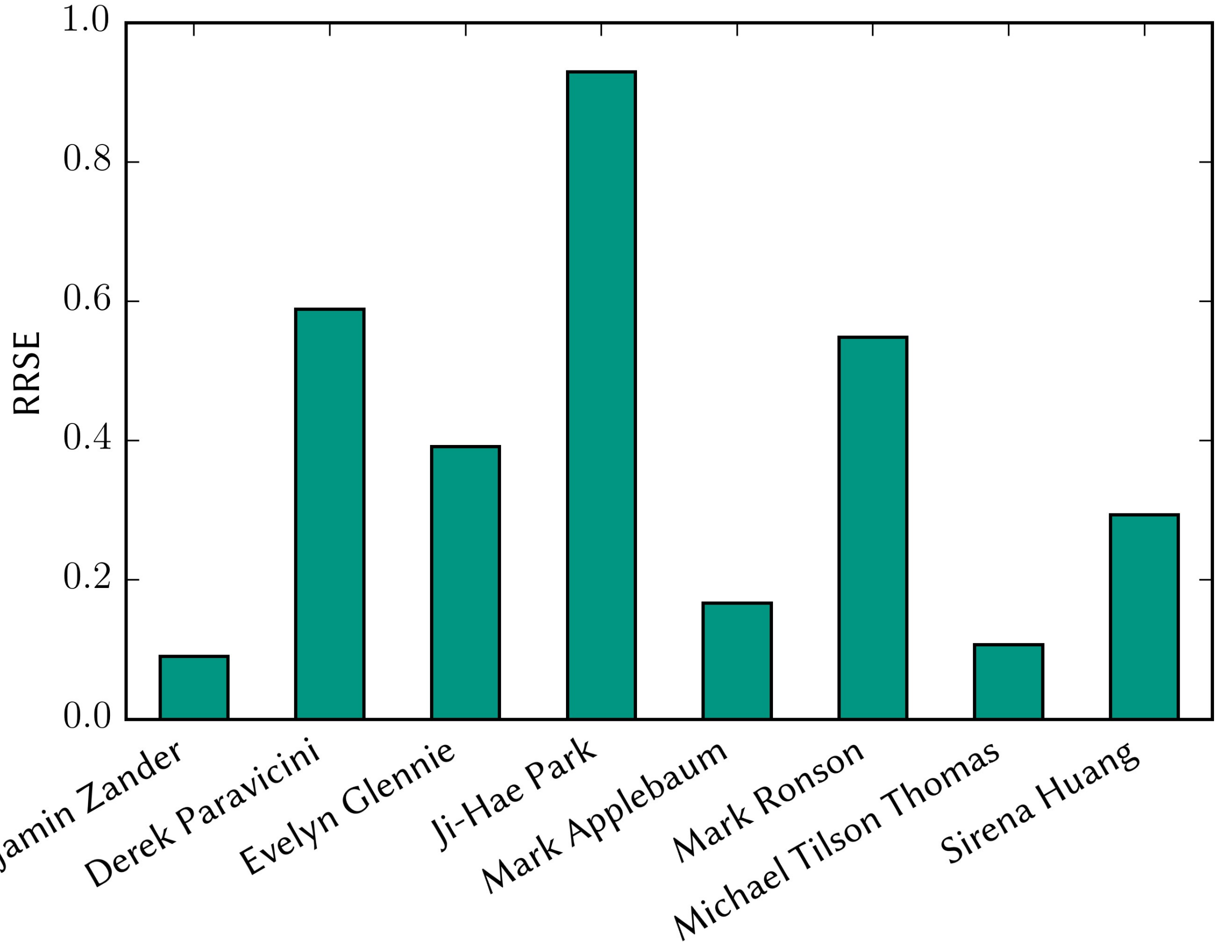
Classification Results

- 3Class accuracy: 87%
- 2Class accuracy: 95%



True Class \ Predicted Class	Speech	Music	Noise
Speech	0.927	0.025	0.048
Music	0.030	0.908	0.062
Noise	0.073	0.145	0.782

Segmentation Results



- Rate of Resolved Segmentation Errors = 1 – (S – M) / (N - M)

Music Example