Automatic Speaker Recognition

MATZ RADLOFF – SEMINAR „AKTUELLE THEMEN DER AUDIOSIGNALVERARBEITUNG“

Contents

Use Cases

Source: [3]

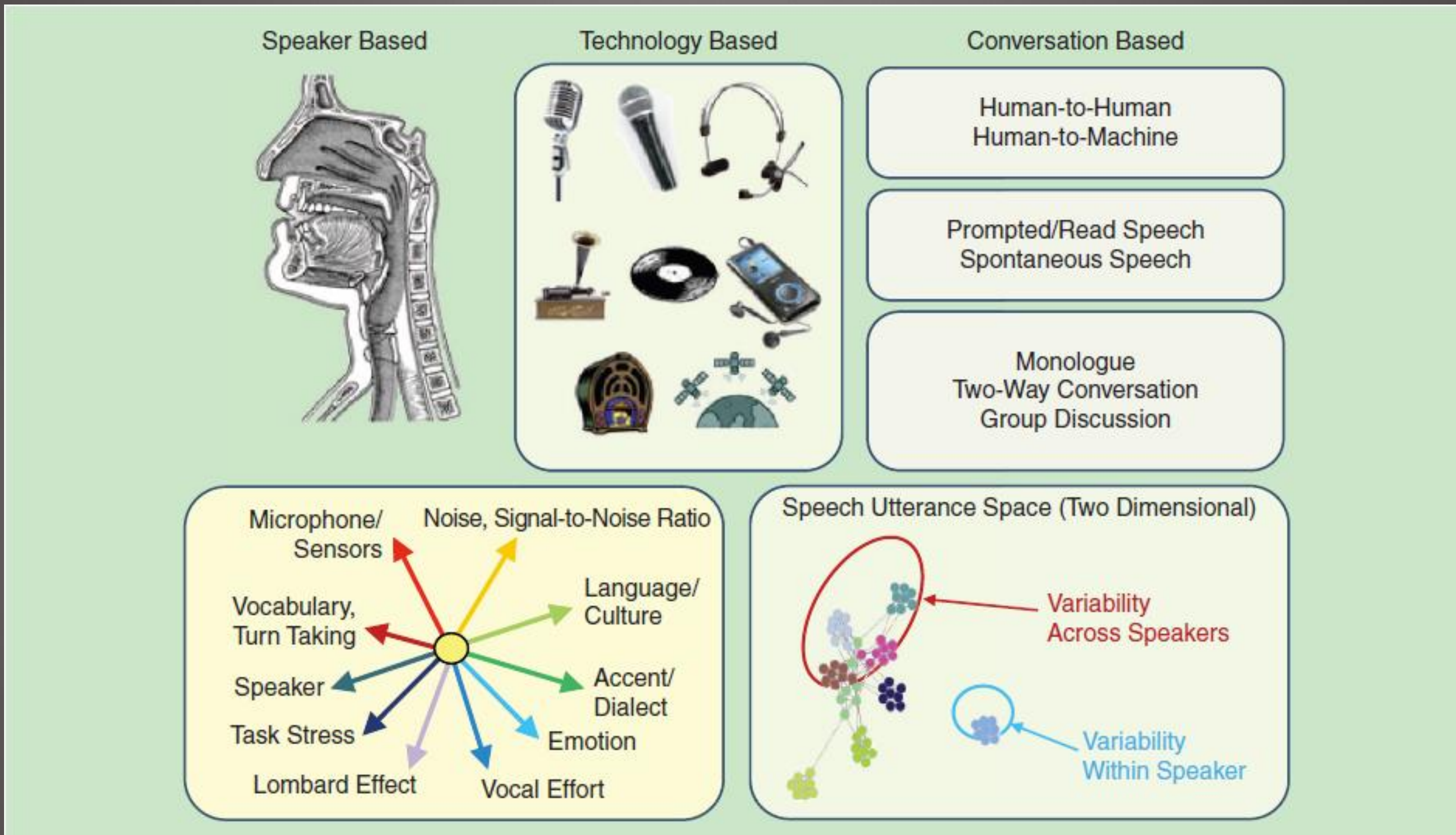
Use Cases

Source: [4]

Challenges

- performance metric (how, not what)
- high variability
  - situational task stress (car, hands-free, distraction)
  - vocal style (whisper, shout)
  - emotion
  - physiological (illness, intoxication, aging)
  - disguise
  - technological (different microphones)
  - environmental (background noise, room acoustic)
  - data quality (duration, sample-rate, compression)

Challenges



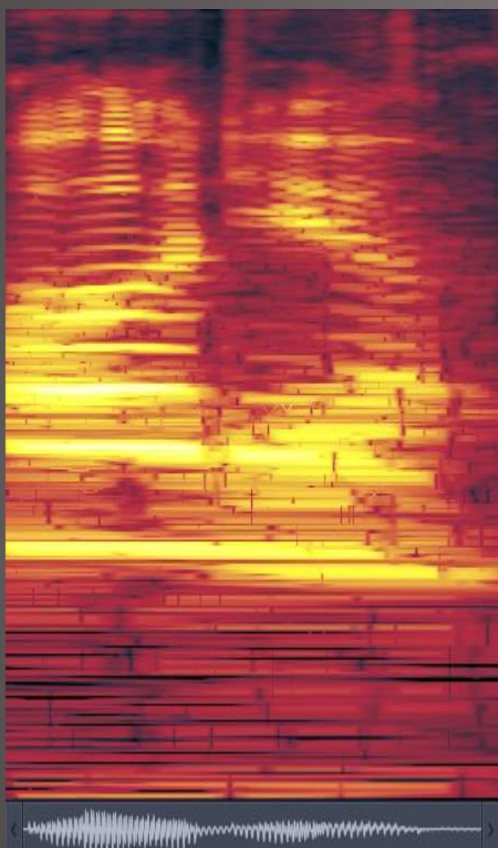
Source: [1]

Human vs. Machine

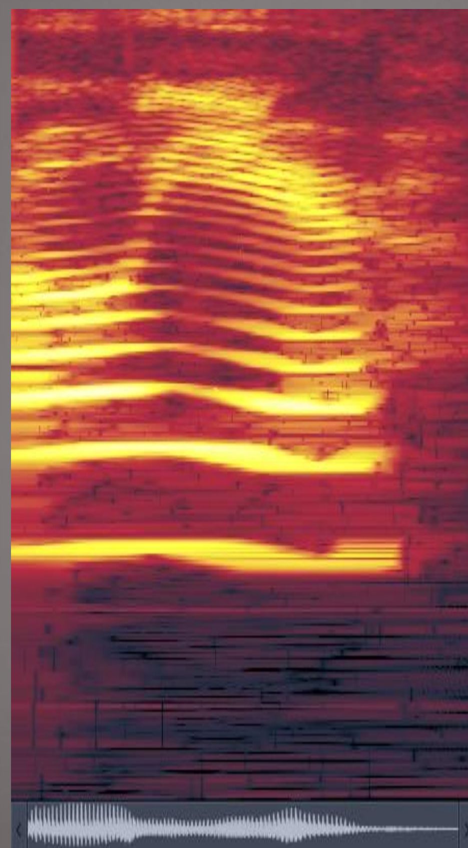
Human	Machine
aquired trait	requires sufficient training data
better if language if known	does not need to „know“ language
bad at identifying unfamiliar voices	consistent performance when adequate data is available
susceptible to bias	only biased towards training data

Human vs. Machine

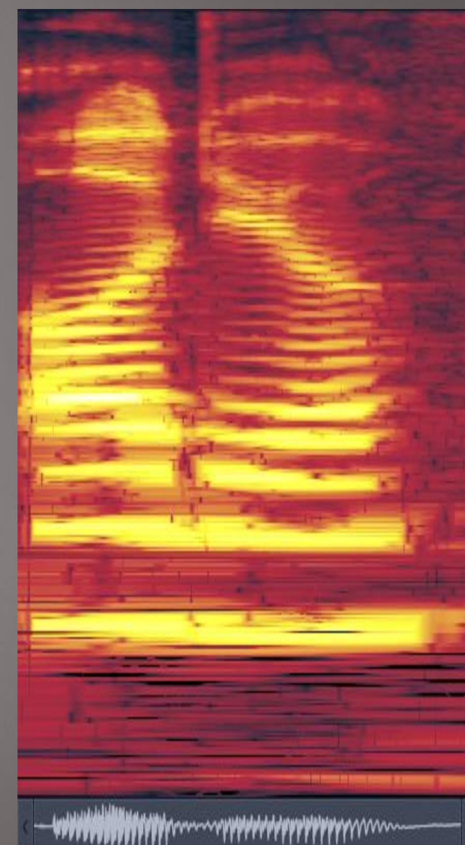
#1



#2



#3



NN-based Speaker Recognition

- "Neural Network Based Speaker Classification and Verification Systems with Enhanced Features"
- paper by Zhenhao Ge et al. [2]
- TIMIT 8K dataset (200 speakers)
- 100% classification rate

NN-based Speaker Recognition

- preprocessing:
  - normalization
  - VAD (Voice Active Detection)
  - MFCC (Mel-Frequency Cepstral Coefficients)
  - Concatenation
- neural network
  - shallow, 1 hidden layer (390:200:200)

NN-based Speaker Recognition

VAD

Short-Term Energy (remove environmental noise)

$$E = \frac{1}{N} \sum_{n=1}^{N} |s(n)|^2$$

NN-based Speaker Recognition

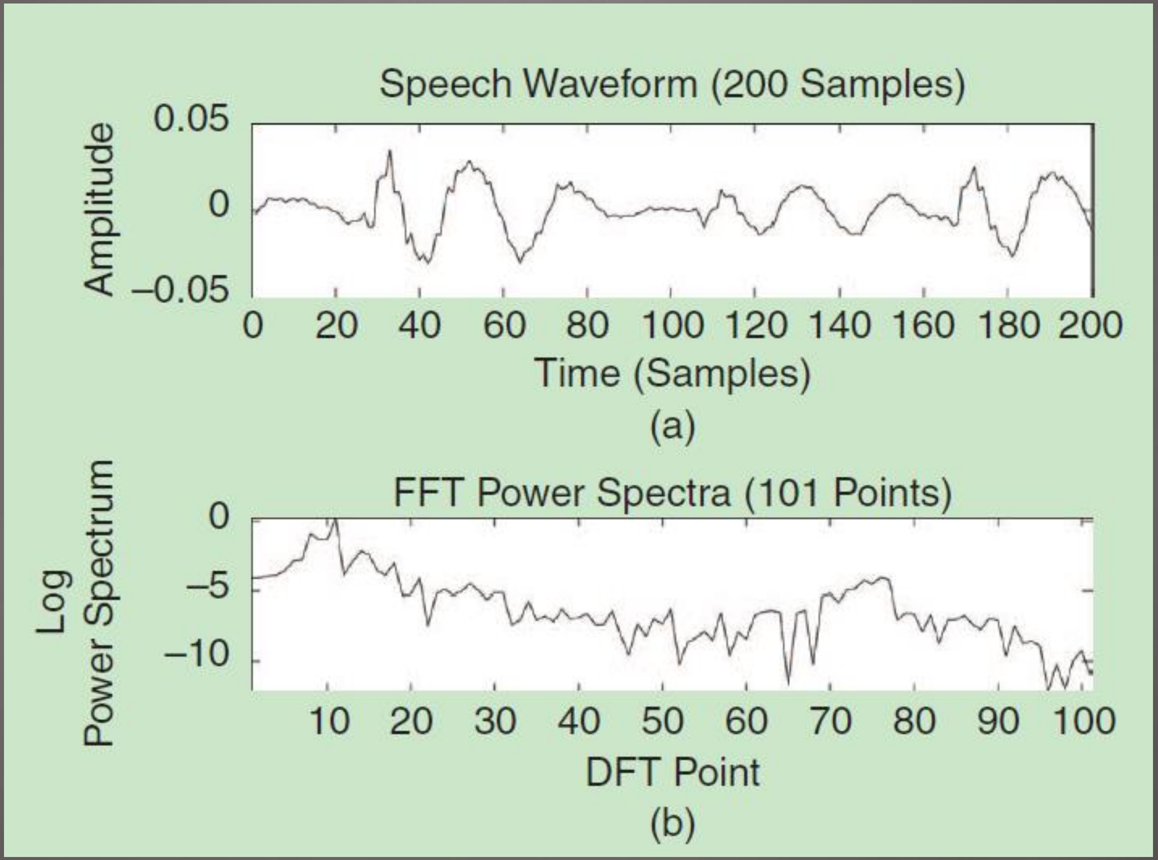
VAD

- Spectral Centroid (remove non-environmental noise)
- „center of mass“ of spectrum

$$C = \frac{\sum_{k=1}^{K} kS(k)}{\sum_{k=1}^{K} S(k)}$$

NN-based Speaker Recognition

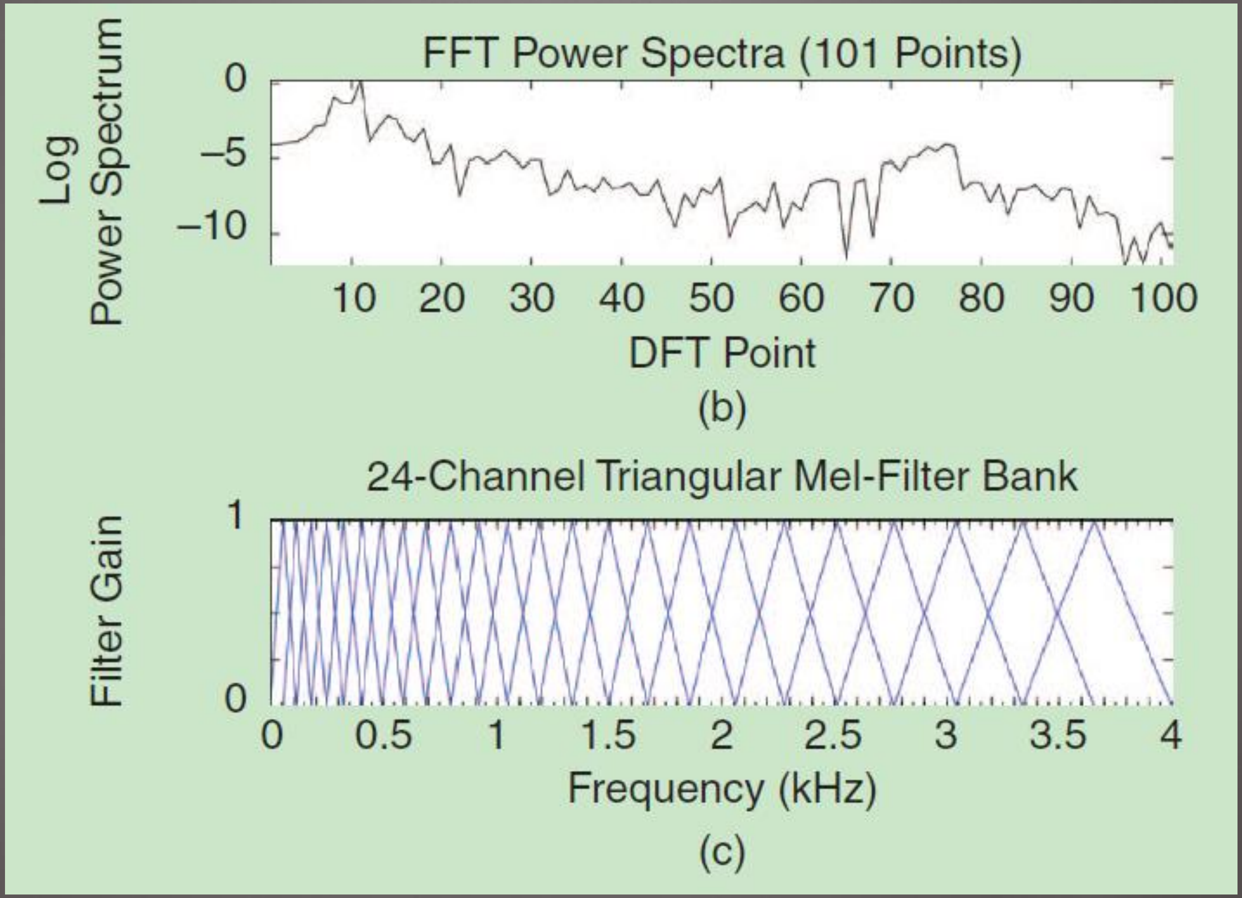
Mel-Frequency Cepstral Coefficients



Source: [1]

NN-based Speaker Recognition

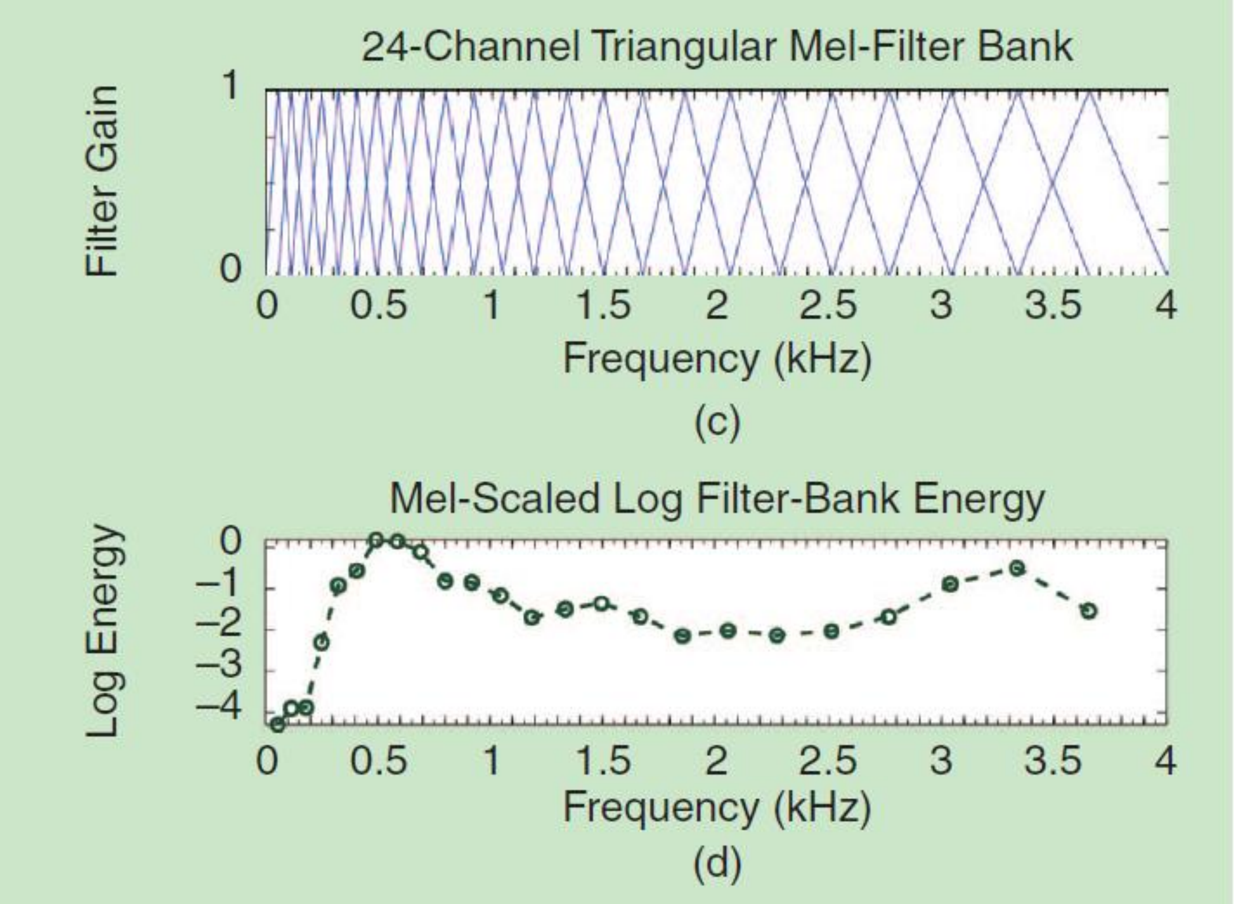
Mel-Frequency Cepstral Coefficients



Source: [1]

NN-based Speaker Recognition

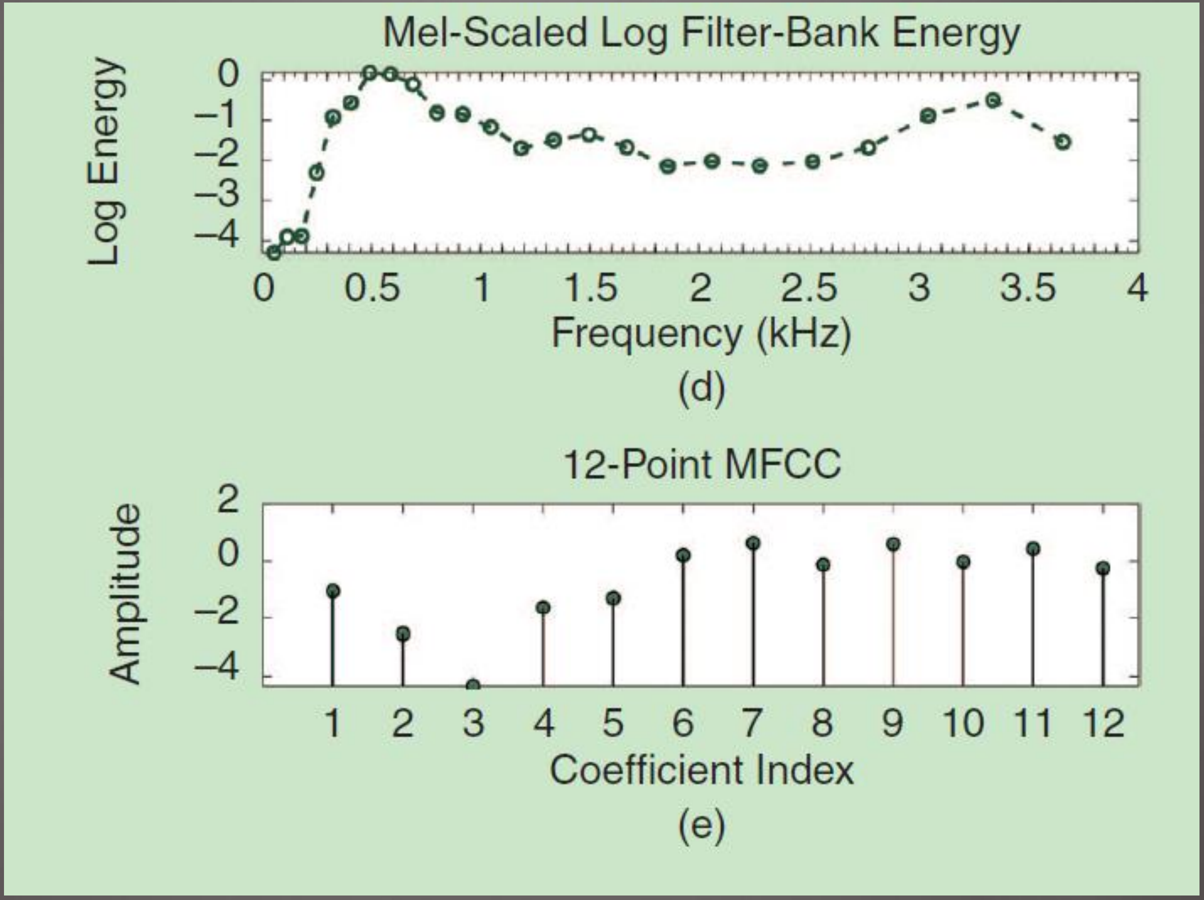
Mel-Frequency Cepstral Coefficients



Source: [1]

NN-based Speaker Recognition

Mel-Frequency Cepstral Coefficients



Source: [1]

NN-based Speaker Recognition

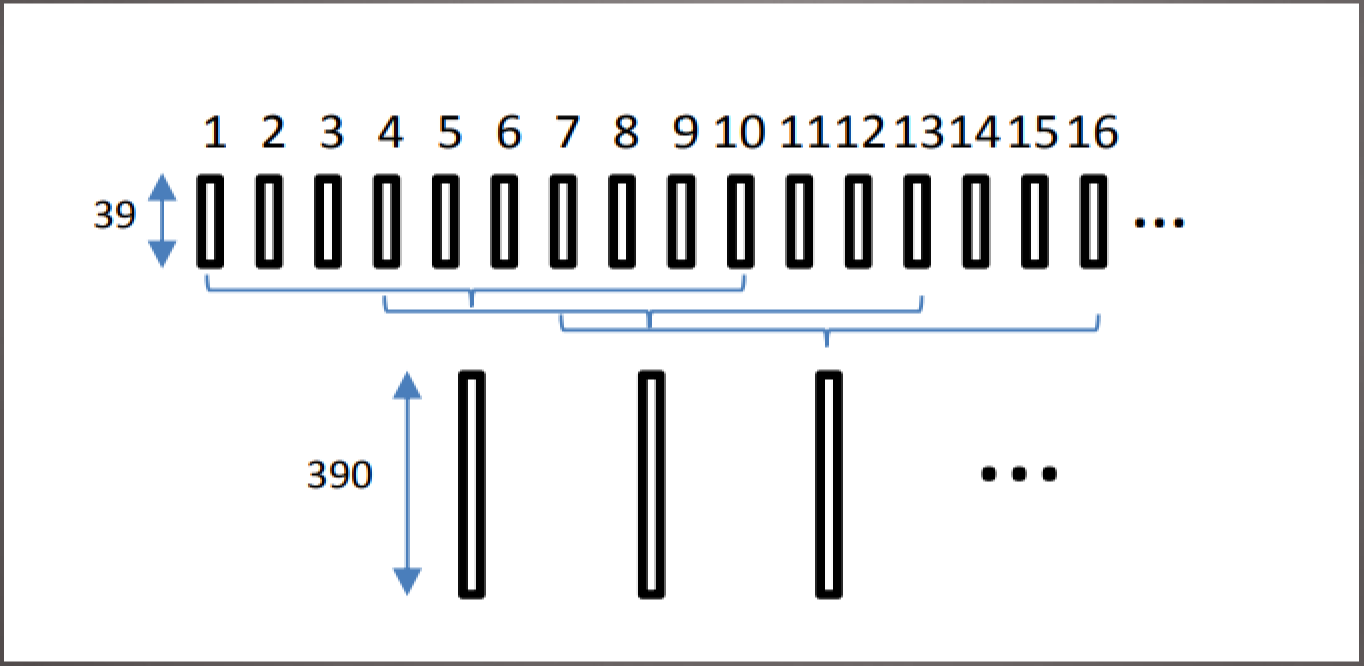
Concatenation

- 39-point MFCC
- 25ms overlapping windows (10ms hop)
- normalization with SMVN (speaker-level multivariate normal distribution)

NN-based Speaker Recognition

Concatenation

- 10 frames concatenated (3 frames hop)
- 39 * 10 = 390 (NN input-vector size)



Source: [2]

NN-based Speaker Recognition

Neural-Network

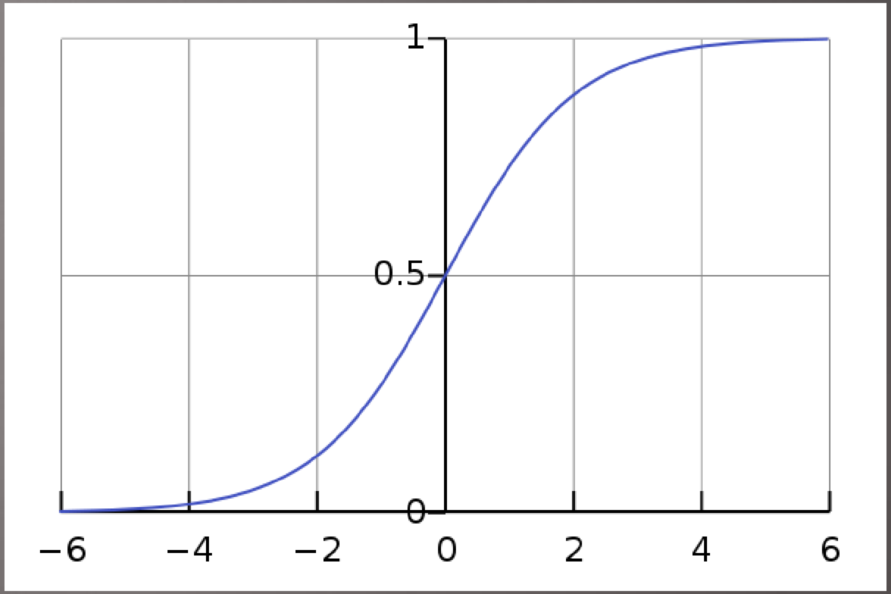
- 390:200:200
- forward-backward propagation
- sigmoid activation-function

NN-based Speaker Recognition

Neural-Network

Sigmoid function:

$$\frac{1}{1+e^{-z^{(l)}}}$$



Source: [5]

NN-based Speaker Recognition

Neural-Network

- Output: 200-dimensional vector
- „likelihood“ (0-1) of speaker

Performance / Conclusion

13.55 frames (0.48s) needed on average to achieve 100% accuracy

TABLE I. NN-BASED SPEAKER CLASSIFICATION PERFORMANCE WITH FIRST 200 MALE IN 8K TIMIT (0.1 SEC./FRAME, ~2.5 SEC./FILE)

Dataset	**Accuracy (%)**		**Frame (sec.) needed for 100% accuracy**		
	frame	***file***	***min***	***mean***	***max***
train	93.29	100	2 (0.13)	3.23 (0.17)	5 (0.22)
test	71.42	100	6 (0.25)	13.55 (0.48)	37 (1.18)

Source: [2]

Sources

[1] Speaker Recognition by Machines and Humans, John H.L. Hansen and Taufiq Hasan, IEEE signal processing magazine, Nov 2015

[2] Neural Network Based Speaker Classification and Verification Systems with Enhanced Features, Zhenhao Ge et al., Intelligent Systems Conference 2017, last access: 11.12.2017 14:43

[3] https://cdn0.vox-cdn.com/thumbor/FLjQuk0OsV2LEAUWcL7X_Fpex7k=/0x37:1848x1005/fit-in/1200x630/cdn1.vox-cdn.com/uploads/chorus_asset/file/9259995/OHYvMm8.jpg

[4] http://www.jobmail.co.za/blog/wp-content/uploads/2015/08/forensic-investigation.jpg

[5] https://en.wikipedia.org/wiki/Sigmoid_function#/media/File:Logistic-curve.svg