Speech Recognition of Spoken Digits

Stefanie Peters

May 10, 2006

Prof. Dr.-Ing. Andreas König

Lecture Information

Sensor Signal Processing

Prof. Dr.-Ing. Andreas König

Institute of Integrated Sensor Systems

Dept. of Electrical Engineering and Information Technology

University of Kaiserslautern

Fall Semester 2005

What did we learn?

Signal Processing and Analysis

Feature Computation

Cluster Analysis

Dimensionality Reduction Techniques

What did we learn?

Data Visualisation & Analysis

Classification Techniques

Sensor Fusion

Systematic Design of Sensor Systems

Sensor Signal Processing Project

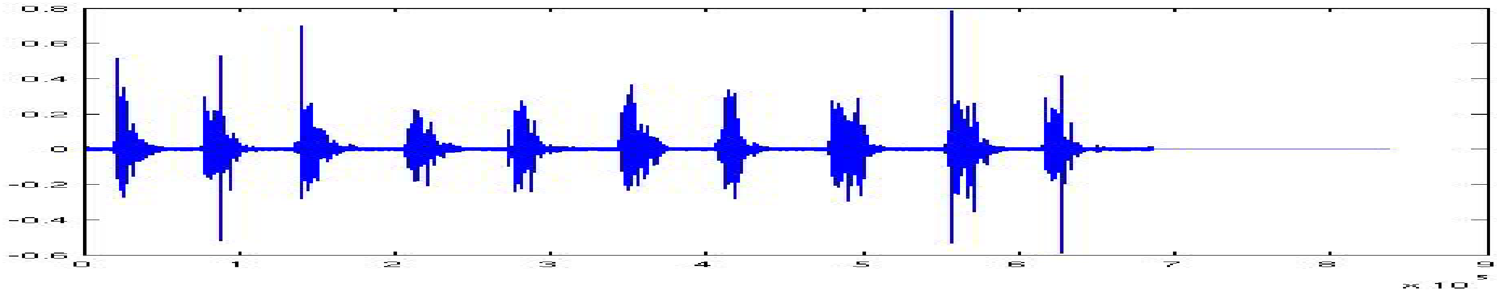
Case study: Speech Recognition of Spoken Digits

- General task for a project:
  - Design and implementation of a recognition system for either image or audio data with the programs Matlab and/or QuickCog
    - Recording / taking of training data
    - Preprocessing to enhance input signals for a ensuing feature computation
    - Selection and computation of suitable features
    - Classification of training and test data
  - Here:
    - Recording of spoken digits with a microphone
    - Implementation of a digit recognition system with Matlab

Training Data

- German Digits (0 to 9), only one speaker
- 10 audio recordings per digit in one wav-file
  (audio recording with 22050 Hz, mono, 16bit)

  -> approximately 8*10^5 sampling points per wav-file



Example for digit 3

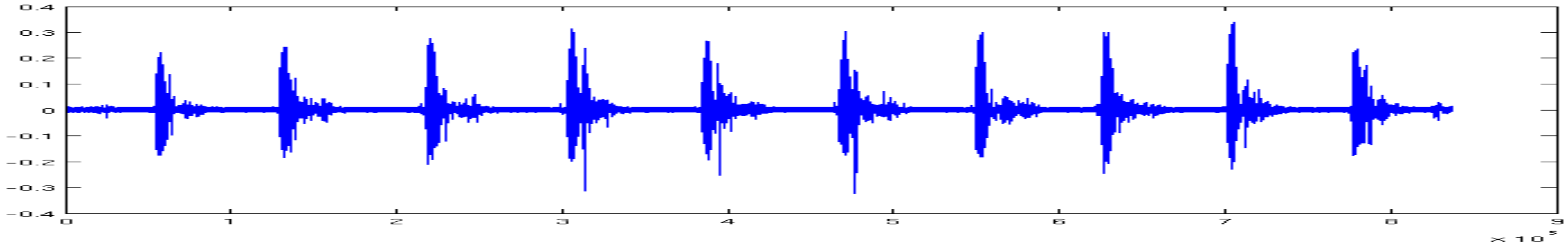
Preprocessing of the Audio Signal

Preprocessing of the audio signal (training and test data) completely in Matlab. Usage of Matlab and own functions:

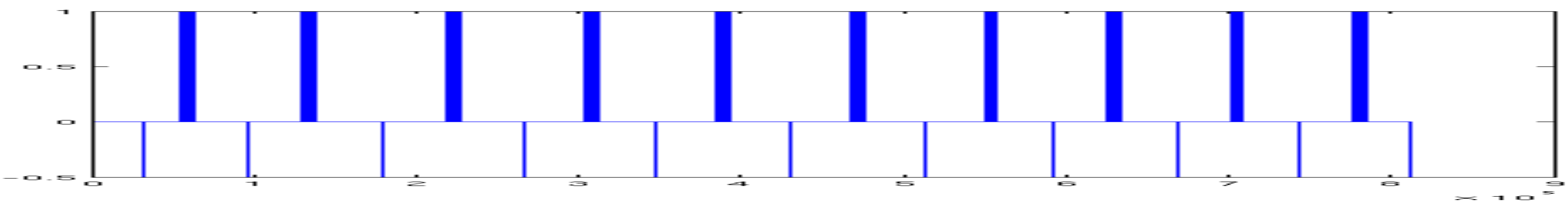
- Adjustment of the y-position of the signal depending on it‘s offset (Offset correction)
  - Signal = Signal – mean value (Signal).
- Noise reduction
  - Noise reduction via low pass filtering
    (the frequency response depends on the noise).
- Separation of the complete audio signal (a series of digits) to smaller audio signals. After the separation a contiguous signal contains only one digit.

Separation of the Audio Signal

- Input signal with noise reduction and offset correction (digit 1).

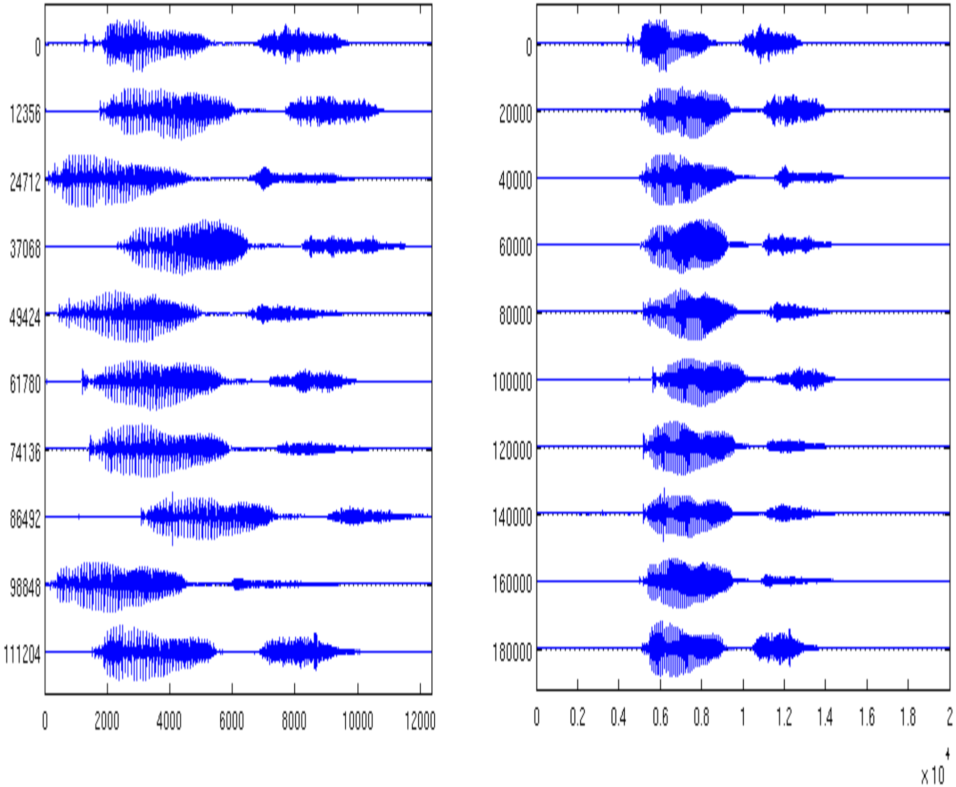


- Extreme values (fixed threshold, bars > 0) and resulting cutting positions (mean position between the extreme values of two different digits, bars < 0).



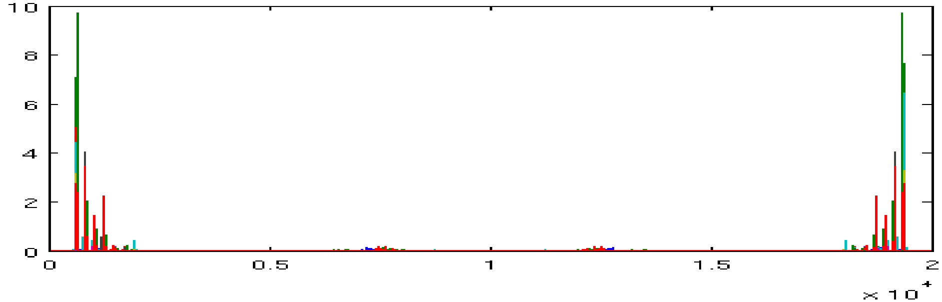
Separation of the Audio Signal

- Cutting of the audio signal in regions which contain only one digit.
- Amplitude scaling (for each digit separately) to the codomain [-1, 1].
- Positioning of each digit using correlation and / or center point adjustment.
- Each signal now consists of 20000 sampling points.



Feature Computation (1)

- Frequency analysis of the audio signal of one digit via Fourier transformation.
- Sub sampling of the signal to reduce the number of the sampling points (usage of low pass filters)
- Example for ten times the digit 1



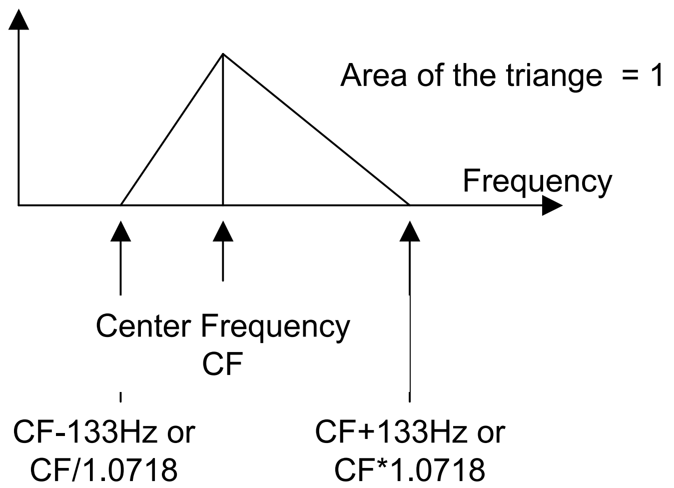
- In ensuing tests, this feature wasn't sufficient for a suitable discrimination between the ten different digits.

Feature Computation (2)

- Mel Frequency Cepstral Coefficients (MFCC)
  - Usage of the ‚Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work' for the MFCC Computation:
  - Windowing of the input signal (here: with a hamming window, usually sampling windows every 10 msec)
  - Discrete Fourier transformation of each window
  - Logarithm (base 10) of the Fourier coefficients
  - Mapping of the results to the „Mel-Scale" using triangle filters
  - Usage of the first 13 MFCC parameter curves

Feature Computation (2)

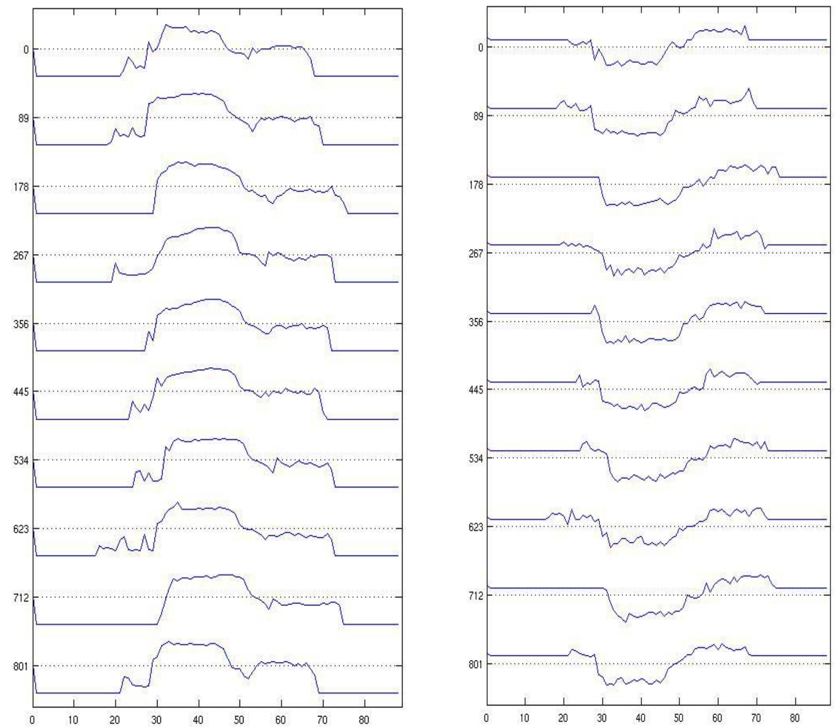
- Filter bank of triangle filters:



The filter bank is constructed using 13 linearly-spaced filters with a distance of 133.33 Hz between the center frequencies followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency).

Training and Classification with Matlab

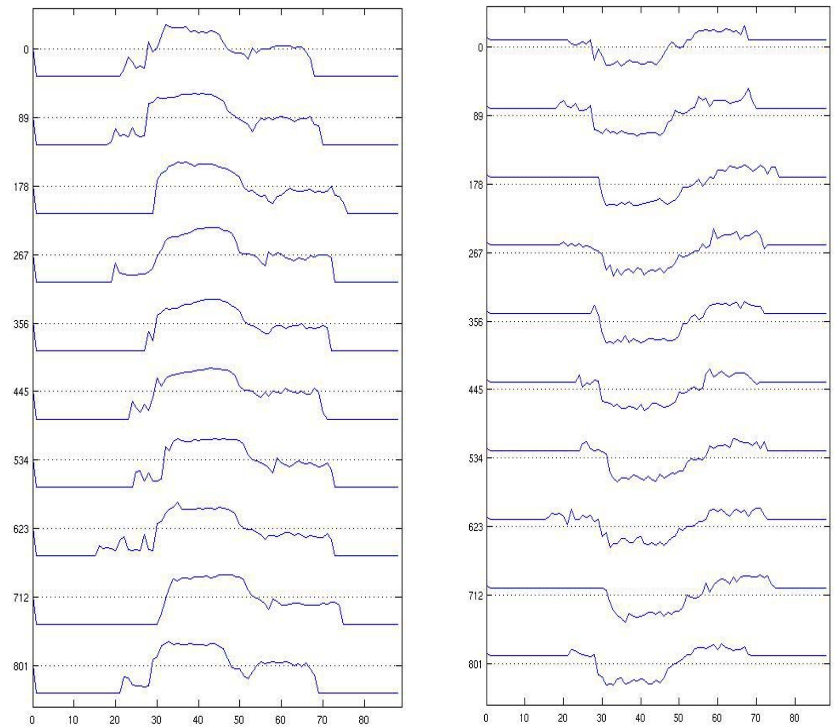
- Method:
  - Usage of the first 13 MFCC parameter curves of the training data in comparison to the test data parameter curves
  - Each MFCC parameter curve contains 89 values (size audio signal (one digit): 20000 sampling points, sampling rate audio signal: 22050Hz, frame rate hamming window: 100Hz)
  - Scaling of the parameter curves to the codomain [0,1]



MFCC 1 (left), MFCC 5 (right) for the ten audio signals of the digit 1

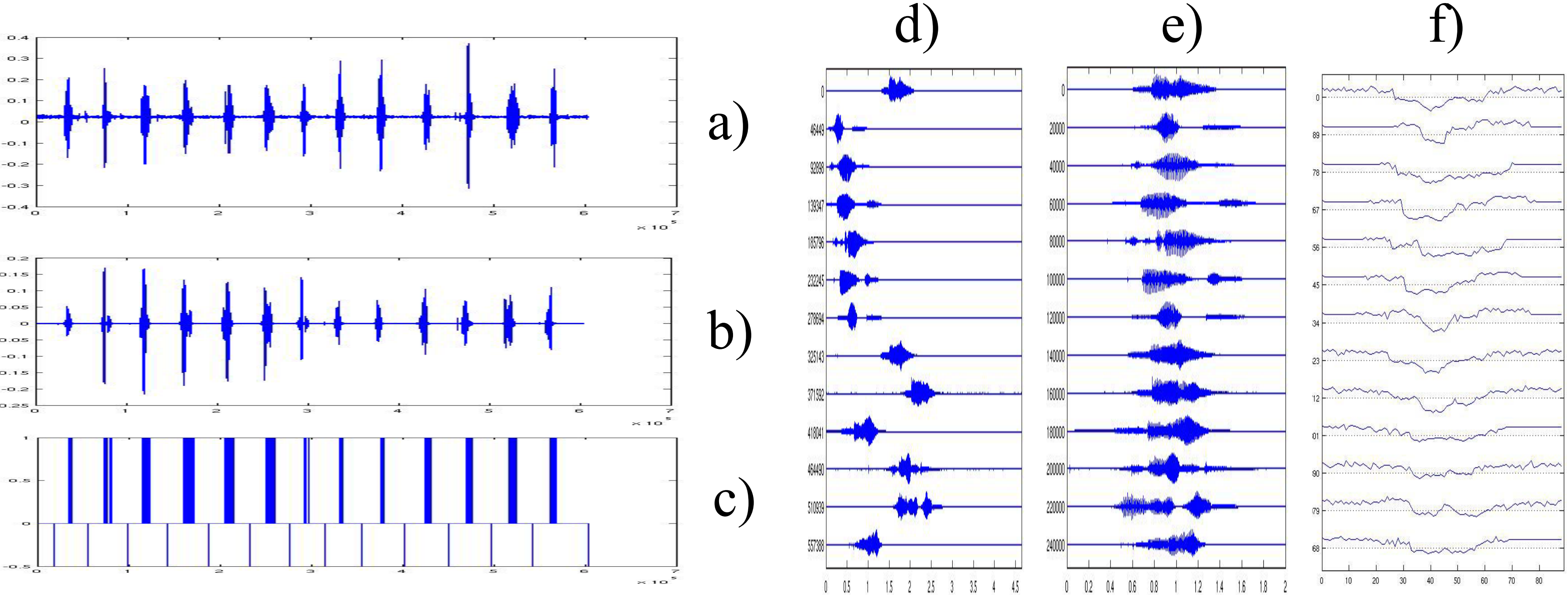
Training and Classification with Matlab

- Method:
  - Computation of the correlation (of the 13 MFCC parameter curves) between training data and test data.
  - The maximal correlation of all MFCC parameter curves is added up for each training sample.
  - The training sample with the highest sum (maximal correlation) sets the affiliation of a test digit.



MFCC 1 (left), MFCC 5 (right) for the ten audio signals of the digit 1

Example: Phone Number



a) No.: 0631/316004574, b) Offset correction and lowpass filtering, c) Extremal values and resulting cutting positions, d) Separated Digits, e) Positioning, f) MFCC 5

-> Recognition rate: 100% (using 13 MFCC parameter curves)

Test results

- As a result of the classification method all training samples are classified correctly.
- In the following table, the results of the classification for different test audio signals are shown, using only one MFCC parameter curve at a time as well as the combined results for all 13 parameter curves.
- Displayed are the recognition rates using the normalizes MFCC parameter curves. Wrong classification is highlighted in red color.

Test results (1)

MFCC (scaled: [0,1])	A: 5 times each digit _1 , 2, 3, 4, 5, 6, 7, 8, 9, 0		B (like A, different micro, more noise):	
MFCC 1:	5/5, 5/5, 1/5, 3/5, 5/5, 5/5, 4/5 ,5/5, 5/5, 3/5	41/50	5/5, 5/5, 1/5, 3/5, 5/5, 5/5, 4/5 ,5/5, 5/5, 3/5	41/50
MFCC 2:	5/5, 5/5, 1/5, 4/5, 5/5, 2/5, 2/5 ,5/5, 5/5, 5/5	39/50	5/5, 4/5, 4/5, 1/5, 2/5, 2/5, 0/5 ,5/5, 2/5, 4/5	20/50
MFCC 3:	5/5, 5/5, 4/5, 1/5, 3/5, 5/5, 5/5 ,5/5, 5/5, 5/5	43/50	5/5, 5/5, 5/5, 0/5, 2/5, 4/5, 2/5 ,5/5, 5/5, 1/5	34/50
MFCC 4:	5/5, 5/5, 5/5, 4/5, 4/5, 4/5, 3/5 ,5/5, 5/5, 5/5	45/50	5/5, 5/5, 5/5, 1/5, 2/5, 4/5, 1/5 ,5/5, 5/5, 1/5	34/50
MFCC 5:	5/5, 5/5, 5/5, 3/5, 5/5, 5/5, 5/5 ,5/5, 5/5, 5/5	48/50	5/5, 5/5, 5/5, 0/5, 3/5, 4/5, 1/5 ,4/5, 4/5, 2/5	33/50
MFCC 6:	5/5, 3/5, 2/5, 4/5, 5/5, 3/5, 5/5 ,5/5, 5/5, 3/5	40/50	5/5, 2/5, 4/5, 3/5, 3/5, 5/5, 3/5 ,5/5, 4/5, 0/5	34/50
MFCC 7:	5/5, 2/5, 0/5, 4/5, 3/5, 5/5, 3/5 ,5/5, 5/5, 5/5	37/50	5/5, 5/5, 2/5, 1/5, 0/5, 5/5, 1/5 ,1/5, 1/5, 0/5	22/50
MFCC 8:	5/5, 0/5, 5/5, 3/5, 4/5, 4/5, 3/5 ,5/5, 5/5, 2/5	36/50	5/5, 0/5, 4/5, 0/5, 1/5, 4/5, 1/5 ,1/5, 5/5, 2/5	23/50
MFCC 9:	5/5, 4/5, 2/5, 1/5, 5/5, 5/5, 2/5 ,5/5, 5/5, 4/5	38/50	5/5, 3/5, 0/5, 0/5, 2/5, 5/5, 0/5 ,0/5, 5/5, 2/5	22/50
MFCC 10:	3/5, 0/5, 4/5, 4/5, 0/5, 3/5, 1/5 ,5/5, 5/5, 5/5	30/50	3/5, 5/5, 3/5, 0/5, 0/5, 0/5, 0/5 ,5/5, 2/5, 3/5	21/50
MFCC 11:	5/5, 4/5, 5/5, 3/5, 0/5, 4/5, 1/5 ,2/5, 4/5, 4/5	32/50	5/5, 5/5, 1/5, 2/5, 0/5, 0/5, 0/5 ,1/5, 5/5, 0/5	19/50
MFCC 12:	1/5, 1/5, 5/5, 4/5, 2/5, 2/5, 4/5 ,5/5, 5/5, 5/5	34/50	1/5, 2/5, 4/5, 1/5, 1/5, 4/5, 3/5 ,5/5, 4/5, 2/5	27/50
MFCC 13:	2/5, 1/5, 2/5, 1/5, 1/5, 3/5, 4/5 ,5/5, 1/5, 2/5	22/50	2/5, 1/5, 2/5, 0/5, 0/5, 1/5, 0/5 ,5/5, 5/5, 0/5	16/50
All:	5/5, 5/5, 5/5, 5/5, 5/5, 5/5, 5/5 ,5/5, 5/5, 5/5	50/50	5/5, 5/5, 5/5, 4/5, 3/5, 5/5, 3/5 ,5/5, 5/5, 4/5	44/50

Test results (2)

	C: Phone Number: 0049 / 613316004574	D: (like C, different micro, more noise): 0049 / 613316004574	E: a series of digits: 1234567890
MFCC 5: (best results when using only one MFCC)	5539 / 631316774574	0039 / 631316004574	1233567899
All :	0049 / 631316004574	0049 / 631316004574	1234567899
	F: Pizza Service 1 (Joey's): 0631 / 10865	G: Pizza Service 2: 004963110865	H: Pizza Service 3: 10865
MFCC 5:	0641 / 10865	6023 / 62116065	10165
All :	0631 / 10865	0049 / 63110865	10865
	I: Digits 1-5 (soft-spoken)	J: Digits 1-5 (loud)	K: Phone Number 2: 0180 / 333 999
MFCC 5:	62345	12025	9186 233 999
All :	12345	12345	9180 333 999
	Other Female: 1739	Other Female: 3456	Other Male: 3578
All:	1730	3436	3662

Conclusions

- Results:
  - Suitable training of a recognition system that recognizes spoken digits from one speaker.
  - More than one MFCC parameter curve is needed for the classification.
  - Pattern Matching via correlation needs a lot of time.
    (Training: 100 digits -> 1 min, Testing: 10 digits -> 2 min)
  - No suitable digit recognition for different speakers.

- Further problems:
  - Scaling of the length of a spoken digit is difficult to implement.

Questions

???

Matlab Functions

One example function for training and testing with Matlab:

function SpeechRecognitionExp ()

- S1 = loadTrainingData1(0); // load training data (here: 10 wav files (->100 digits))
- T1 = loadTestData1(0); // load test data (here: 10 wav files (->50 digits))

- PlotParam = 0; // 1: plot results (filtered signal cut signal, mfcc parameter, …)
- FilterParam = [800, 0.005, 0.99999]; // Filter size and frequencies for a band pass filter
- CutsParam = [0.1, 0.05, 10000]; // Thresholds for the separation of the digits, minimum distance between two digits (in sampling points)
- MFCCParam = 0; // 0: all MFCC parameters, >0 only one MFCC parameter
- NormalizeParam = 1; // 1: scale MFCC parameters to [0,1]

- [MFCCSignal, FilterParamTrain, CutsParamTrain] = SpeechRecogTrain(S1, MFCCParam, PlotParam, FilterParam, CutsParam);
- SpeechRecogTest(MFCCSignal, T1, MFCCParam, NormalizeParam, PlotParam, FilterParamTrain, CutsParamTrain);