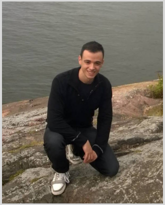
Speech Recognition - EQ2340

HMM (Hidden Markov Models)

Alessio Russo



Control Engineering (Politecnico di Milano)

Lars Lindemann

Systems, Control and Robotics (KTH)

Royal Institute of Technology (KTH, Stockholm)

3rd November, 2015

Content

Speech Recognition of a limited speech corpus

Speech Recognition of a limited speech corpus

- High demand in industry
- Usage in current systems (e.g. Siri)
- Easy to understand general principle

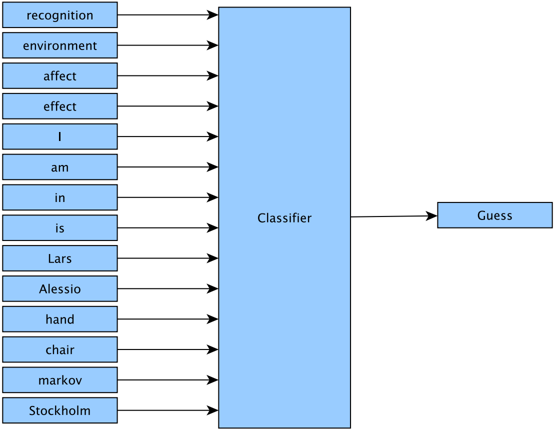
Speech Recognition of a limited speech corpus

- High demand in industry
- Usage in current systems (e.g. Siri)
- Easy to understand general principle

Speech Corpus

- General speech corpus to form sentences
- Multisyllabic, similar and short words to challenge the system

Speech Recognition of a limited speech corpus

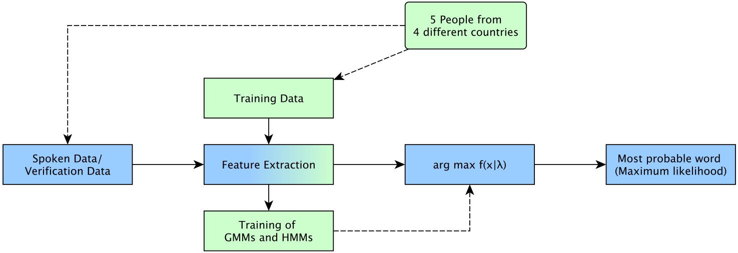


Speech Recognition of a limited speech corpus

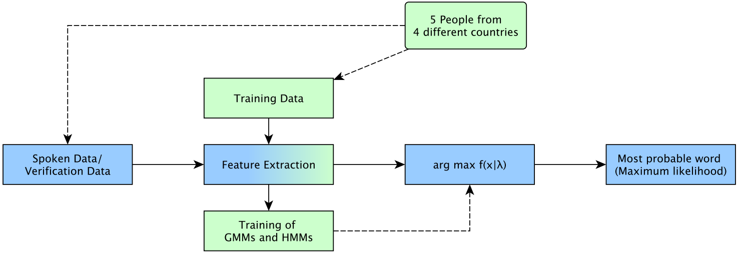
- Two examples...

Lars: affect
Natalie: recognition

Overview of the Implementation



Overview of the Implementation



- Distinguish between training and validation/live demonstration

Content

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Continuous feature vectors

- 13 MFCC (Mel-frequency cepstrum coefficients)
- 26 dynamical features (independent of absolute value)
- 30 ms time frame

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Continuous feature vectors

- 13 MFCC (Mel-frequency cepstrum coefficients)
- 26 dynamical features (independent of absolute value)
- 30 ms time frame

Possible problems

- Pitch
- Different speakers (absolut output value)
- Noise

Continuous feature vectors

- 13 MFCC (Mel-frequency cepstrum coefficients)
- 26 dynamical features (independent of absolute value)
- 30 ms time frame

Number of States for left-right HMM

- Trade off: too few parameters vs. amount of training data
- Also: Limited training data
- State assignment due to syllables + start/end state

Number of States for left-right HMM

- Trade off: too few parameters vs. amount of training data
- Also: Limited training data
- State assignment due to syllables + start/end state

Number of States for left-right HMM

- Trade off: too few parameters vs. amount of training data
- Also: Limited training data
- State assignment due to syllables + start/end state

System Design, Training and Testing

HMM

Number of States for left-right HMM

- Trade off: too few parameters vs. amount of training data
- Also: Limited training data
- State assignment due to syllables + start/end state

environmnent: 6 states *chair*: 4 states

Number of States for left-right HMM

- Trade off: too few parameters vs. amount of training data
- Also: Limited training data
- State assignment due to syllables + start/end state

environmnent: 6 states *chair*: 4 states

Output distributions

- GMM (Gaussian Mixture Models)

Recorded data

- 5 people: 15 recordings per word
- One person has been disregarded
- 840 recordings in total, 60 per word

System Design, Training and Testing

Training data and validation set

Recorded data

- 5 people: 15 recordings per word
- One person has been disregarded
- 840 recordings in total, 60 per word

k-fold approach

- k=5 sets
- 48 training and 12 validation samples

Final HMM

- 5 sets have been tweaked on their validation set and compared on the whole set.
- Recognition rate in table below

Final HMM

- 5 sets have been tweaked on their validation set and compared on the whole set.
- Recognition rate in table below

Final HMM

- 5 sets have been tweaked on their validation set and compared on the whole set.
- Recognition rate in table below

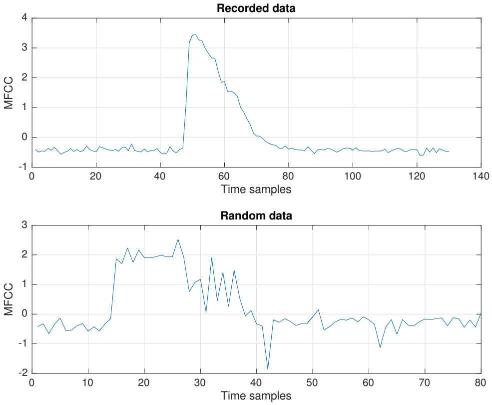
word	1	2	3	4	5	6	7
validation set	1.00	1.00	1.00	1.00	1.00	1.00	1.00
overall	0.76	0.95	1.00	1.00	1.00	1.00	1.00
word	8	9	10	11	12	13	14
validation set	1.00	1.00	1.00	0.833	1.00	1.00	1.00
overall	1.00	0.95	1.00	0.7	1.00	0.95	1.00

System Design, Training and Testing

Testing and tweaking

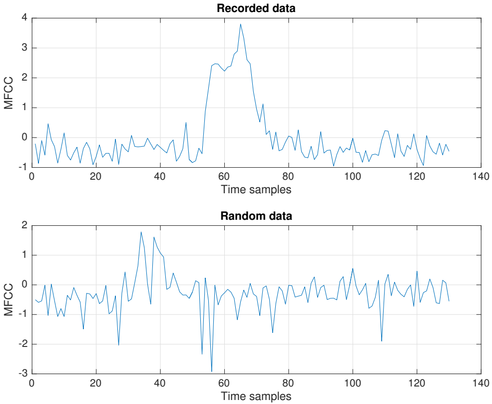
Some realizations

- word *I* and 1st MFCC coefficient



Some realizations

- word *I* and 2nd MFCC coefficient



Content

Classification Errors

- Average Classification Error: 1.2 %(validation) and 4.9 % (overall)
- Most commonly missclassified: *hand* with 16.6 % and 30 %

Classification Errors

- Average Classification Error: 1.2 %(validation) and 4.9 % (overall)
- Most commonly missclassified: *hand* with 16.6 % and 30 %

Results

System Performance

C =

$$\begin{pmatrix}
46 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 14 & 0 & 0 & 0 & 0 \\
0 & 57 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 57 & 0 & 1 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 & 0 & 0 \\
0 & 0 & 7 & 0 & 0 & 11 & 0 & 0 & 0 & 0 & 42 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 57 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 60
\end{pmatrix}$$

Results

System Performance

Lars: recognition
Martin: recognition
Natalie: recognition

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Take aways

- If data is rare, use smaller k in k-fold approach
- Good training data is important
- Collect more training data (remember trade-off)
  but: then be aware of overfitting!

The system

- Satisfying overall recognition rate
- Problems with the words *hand* and *recognition*
- less problems with *affect* and *effect* or short words

Results

Live Demonstration

...