Assignment 5, speach recognition

Albert Öst, Per Emil Hammarlund

5/29/2019

Play example from database

Feature extraction

Features where extracted using the provided $GetSpeachFeatures$, and the MFCCS from $GetSpeachFeatures$ was then normalized. The normalized MFCCS was the feature extracted observation sequence used in the model.

HMM design

- One HMM for each number
- Numbers 0-9
- In each HMM, a hidden state for each phoneme and two silent states where added in the beginning and end.

number	phonemes
zero	Z IY R OW
one	W AH N
two	T UW
three	TH R IY
four	F AO R
five	F AY V
six	S IH K S
seven	S EH V AH N
eight	EY T
nine	N AY N

- To find a good approximation for the initialization of the transition prob. matrix. The average numner of time frames for each numbers MFCCS (in the train set) was calculated and then divided by the number of hidden states. This fraction for each class was then used as the prob. of transitioning to the next state.

```
avgs =

   0.200000000000000
   0.200000000000000
   0.166666666666667
   0.250000000000000
   0.200000000000000
   0.200000000000000
   0.250000000000000
   0.250000000000000
   0.166666666666667
   0.166666666666667
```

- For example, the first class (zero) was given the following initial trans. prob. matrix.

$$A_{zero} = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 & 0.2 \end{pmatrix}$$

- All trans. prob. matrices where left-right and finite.
- The emmisions from each hidden state were 13 dimensional vectors. The emitters from each hidden state was given a GMM with three components.
- Each of the threee components where given equal mixture coefficients π_i, a subjectivley non informative diagonal covariance matrix (the identity), and the same mean.
- Noise was added to the means and covariances in between hidden states.

Training and testing method

- According to the documentation of the dataset that we chose. The test set is is the first 10% of the recordings. So recordings 0-4 in each class and speaker where assigned to the test set, and recordings 5-49 where assigned to the training set. The HMMs where trained on their respecitve classes using the provided $hmm.train()$ method.
- The HMMs where tested by calculating the log prob for each HMM given each observation in the test set. And the accuracy was:

```
accuracy =

   1.000000000000000
   1.000000000000000
   0.850000000000000
   0.900000000000000
   0.900000000000000
   1.000000000000000
   0.800000000000000
   1.000000000000000
   0.700000000000000
   0.950000000000000
```

- So all classes had great accuracy except for the number eight. Looking at eight, we saw that perhaps three GMM components in the first non-silent hidden state seemed to only have two peaks, so we tried to use one less component in the GMM. This did not have any effect.
- Furthermore, a few recordings did not pronounce all phonemes. So a setup with one less hidden state was tried, but there was no significant difference.

Training data vs randomized from HMM

The training data and randomized sequences from the hmms took the following form.

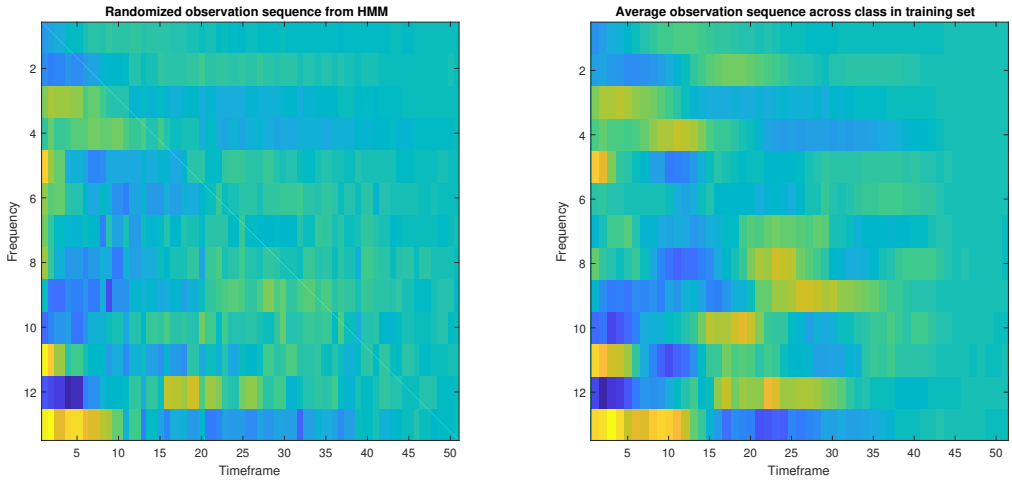

Figure 1: Randomized vs training data class zero

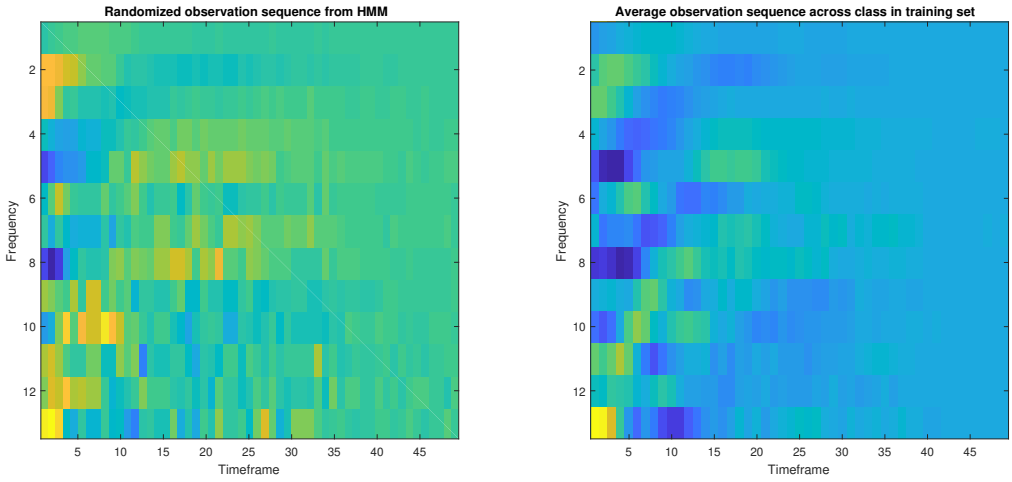

Figure 2: Randomized vs training data class one

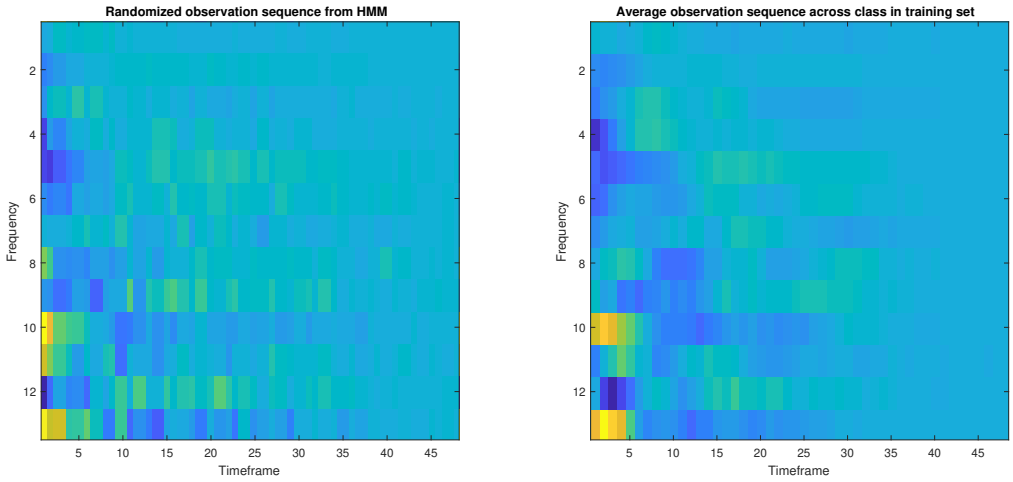

Figure 3: Randomized vs training data class two

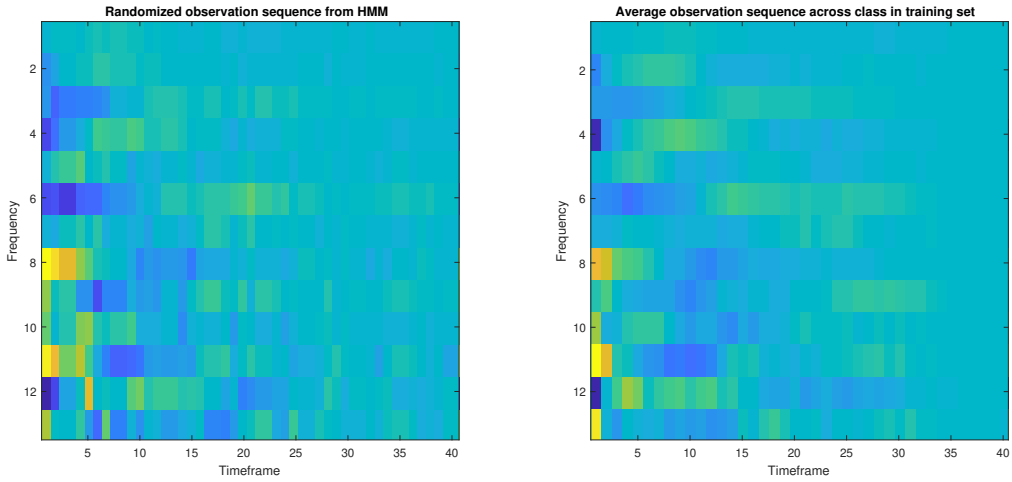

Figure 4: Randomized vs training data class three

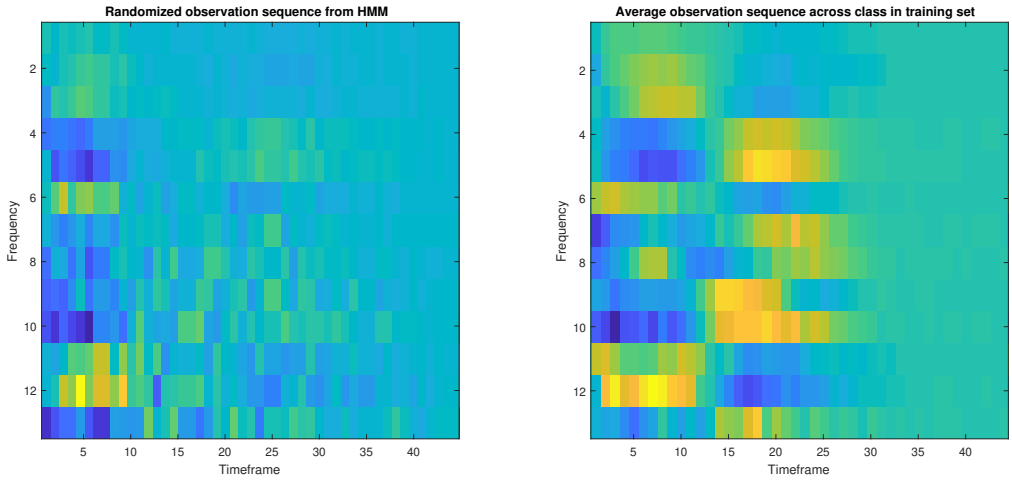

Figure 5: Randomized vs training data class four

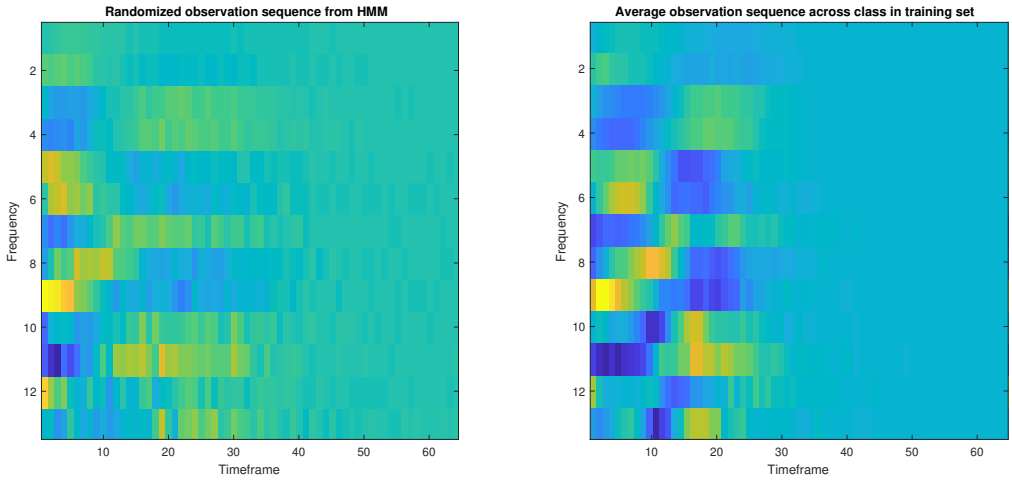

Figure 6: Randomized vs training data class five

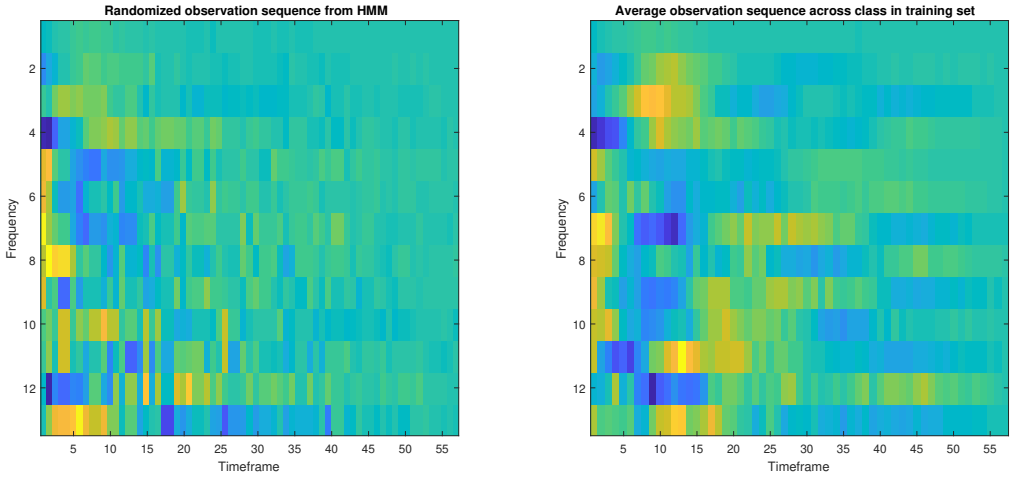

Figure 7: Randomized vs training data class six

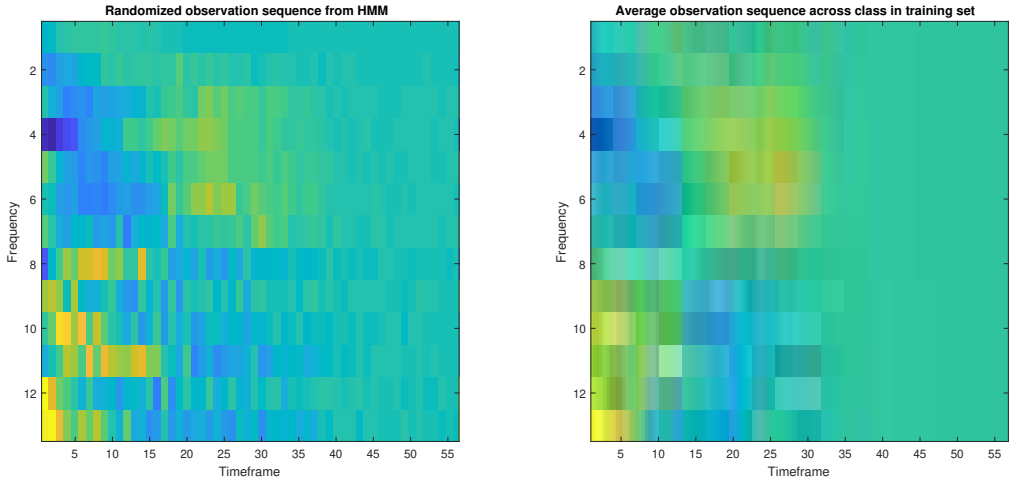

Figure 8: Randomized vs training data class seven

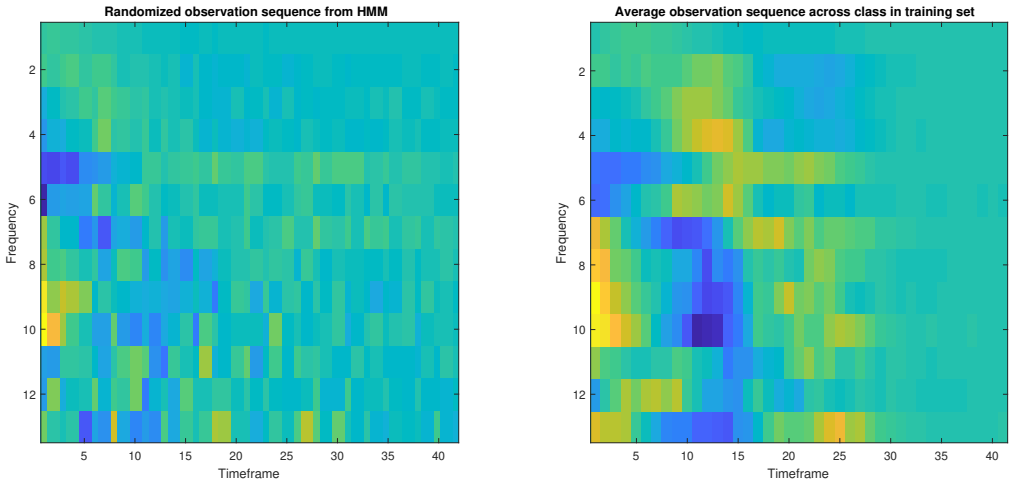

Figure 9: Randomized vs training data class eight

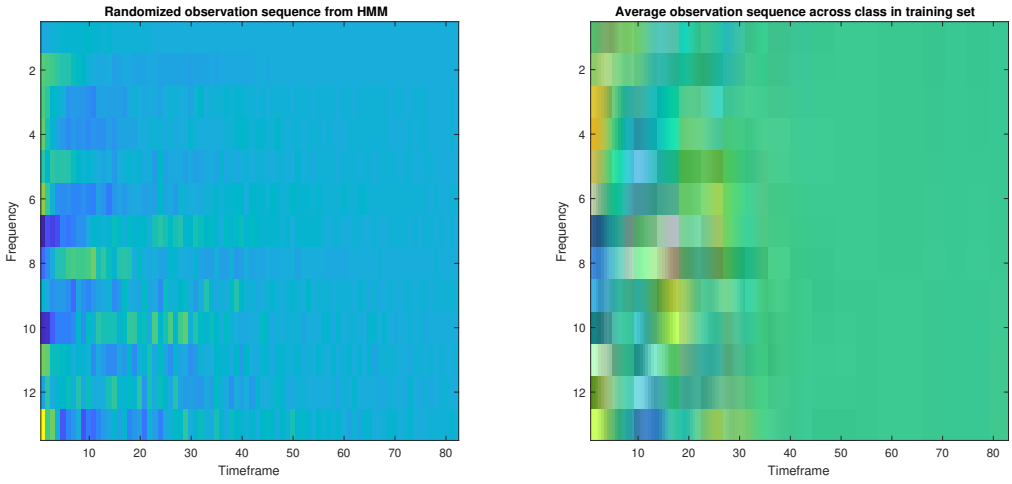

Figure 10: Randomized vs training data class nine

Missclassified instances

Some missclassified examples can be seen below:

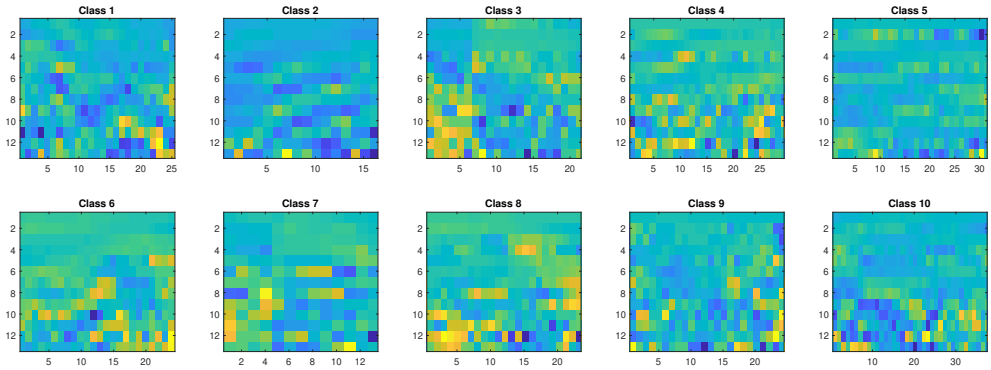

Figure 11: Example missclassified instances

Confusion matrix

The confusion matrix took the following form:

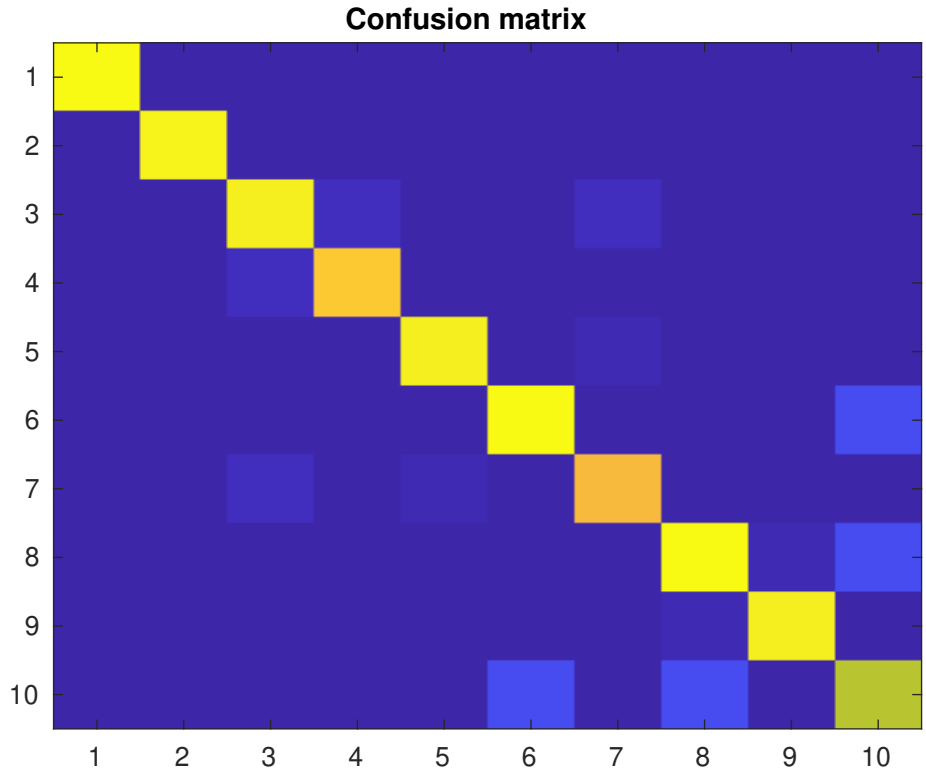

Figure 12: Confusion matrix

Live demo

Did not work at all when trained on the given dataset. But we created our own dataset, and trained a new model which worked significantly better...

Conclusions

- Though we had a rather large dataset with 4 speakers and 50 recordings per class (a total of 2 000). It still was not enough to generalize to our voices.
- Variablilty in pronouciation can lead to different amounts or outright different phonmemes, leading to issues when trying to generalize.