Exercise #3:
Generative models and phone classification

İlknur Baş
151226814
16.04.2023

Task 1: Implement a Gaussian model (GM) for phone classification

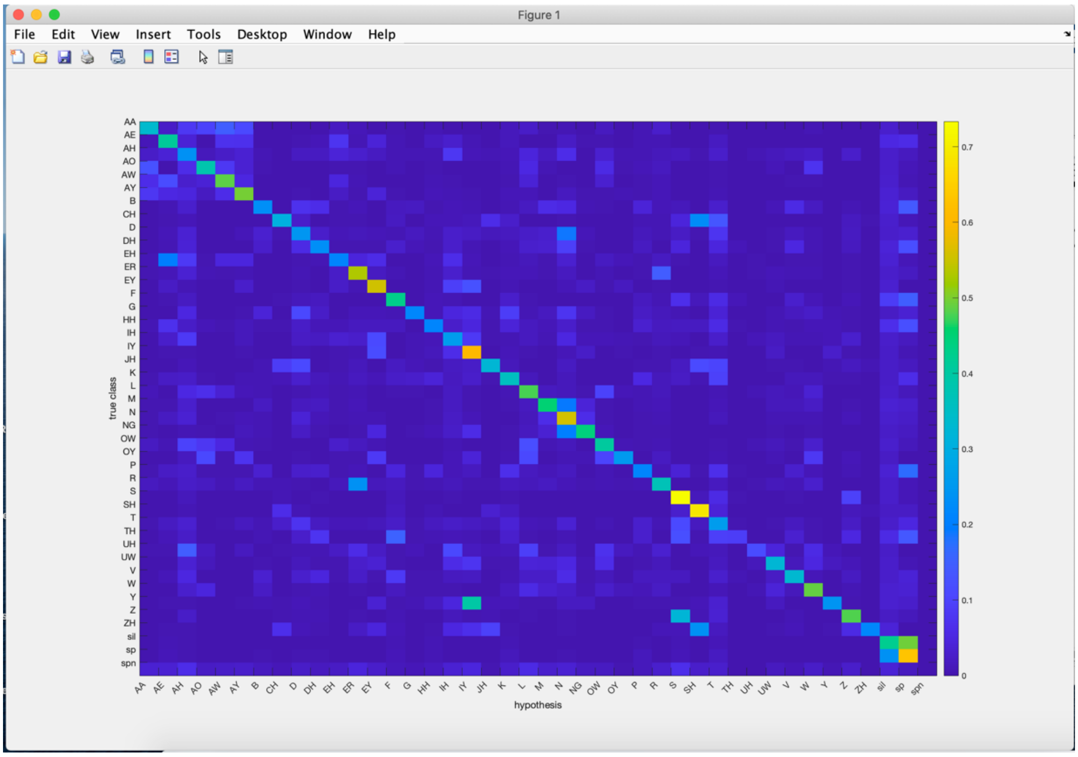


Figure 1: Confusion matrix for phone classification using Gaussian model

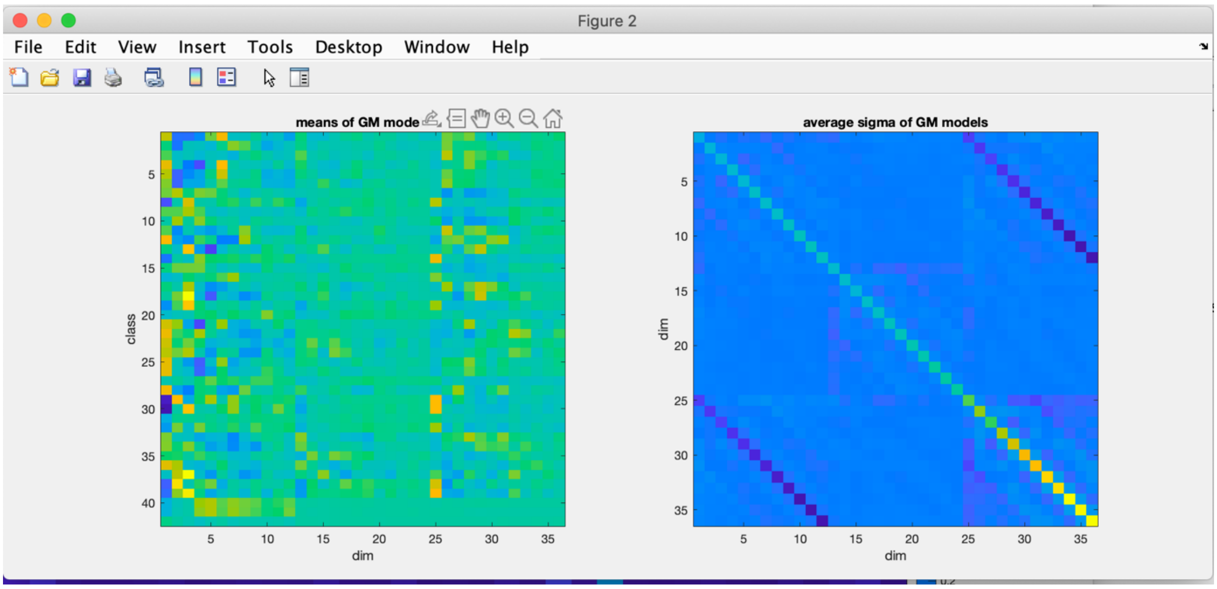
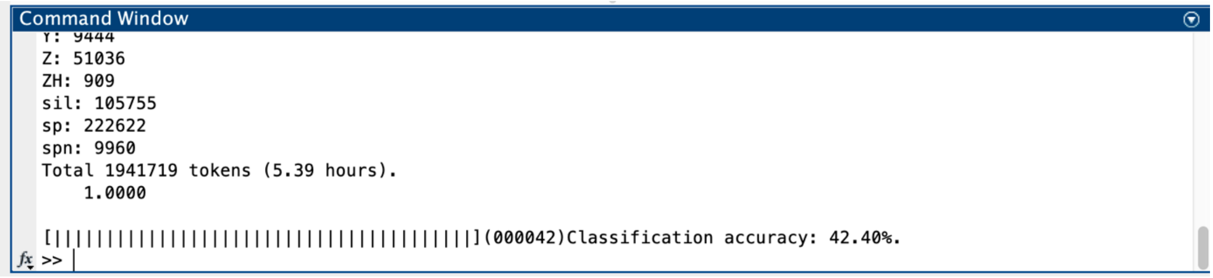


Figure 2: Gaussian model parameters



```
Y: 9444
Z: 51036
ZH: 909
sil: 105755
sp: 222622
spn: 9960
Total 1941719 tokens (5.39 hours).
    1.0000

[||||||||||||||||||||||||||||||||||||||||](000042)Classification accuracy: 42.40%.
>>
```

Figure 3: Accuracy of phone classification using Gaussian model

Q1.1:

Firstly, Gaussian distribution have 2 main parameters, mean and standard deviation. The mean shows the average value of the data, and the standard deviation shows the spread of the data. If we think this in the context of speech features, assume we have a feature vector (calculated using MFCC), the mean of each MFCC coefficient will help us to understand/compare the differences in different speech utterances and for example it could give us an idea about the value of those in certain phones. Likewise, the variance of each MFCC coefficient will tell us how a coefficient varies, we can calculate this for coefficients in different speech utterances and try to understand the which phones has certain variance pattern. This helps us to classify phones. Similarly, in previous exercises we have learned that vocal track gives information about the spectral envelope which also gives information about formants. So, when we model speech features with Gaussian distribution (which has mean and standard deviation as parameters as I mentioned above), standard deviation (similar term, variance) will denote the spectral envelope and the mean will denote the "mean" or "the most common" frequency. Also, when we observe the Gaussian distribution function, we see that there is a peak in the middle which in fact could denote the most common frequency. Because of these properties, it becomes easy to model speech features and classify them as well especially the phones that has a single most common frequency. Lastly, using Gaussian distribution is a suitable choice as in MATLAB, there are built-in functions for Gaussian distributions which makes it easy to use.

Q1.2:

In single multivariate Gaussian distribution, we assume that each phone is modeled using a single Gaussian distribution, meaning that for all samples of the particular phone class, we have the same mean and covariance matrix (calculated using feature vectors of those samples) for the distribution. In other words, each phone class has its own same mean and covariance matrix which are calculated from training data. This may lead not to express the variability in the data (in feature vectors) properly for a particular phone class. Because even though some samples are happened to be in the same phone class, they may differ a bit in terms of their acoustic features (MFCC etc.) due to the noise in environment etc. Another disadvantage is that assume there are outlier in our training data (even though the data we have used seems highly reliable one), this can affect the model parameters and as a result the classification can be affected in a negative way.

Q1.3:

First of all, the diagonal in the confusion matrix denotes the number of correctly classified samples for each phone. And the off-diagonal elements denote the misclassified samples. For that reason, I observed the off-diagonal elements that have "lighter" color in their cells as they will denote the classification errors (/the phones that are confused the most). In below, I will point out the most outstanding ones.

- Around 0.3-0.4 times, "Y" is got confused with "IY".
- Around 0.3 times "Z" is got confused with "S".
- Around 0.2 times "ZH" is got confused with "SH".
- Around 0.2 - 0.3 times "R" is got confused with "ER".

In fact, these confusions/classification errors make sense as in each case both phones have similar acoustic properties. Also, I observed that "sil" (silence) and "sp" got confused as well.

Task 2: Implement a Gaussian mixture model (GMM) for phone classification

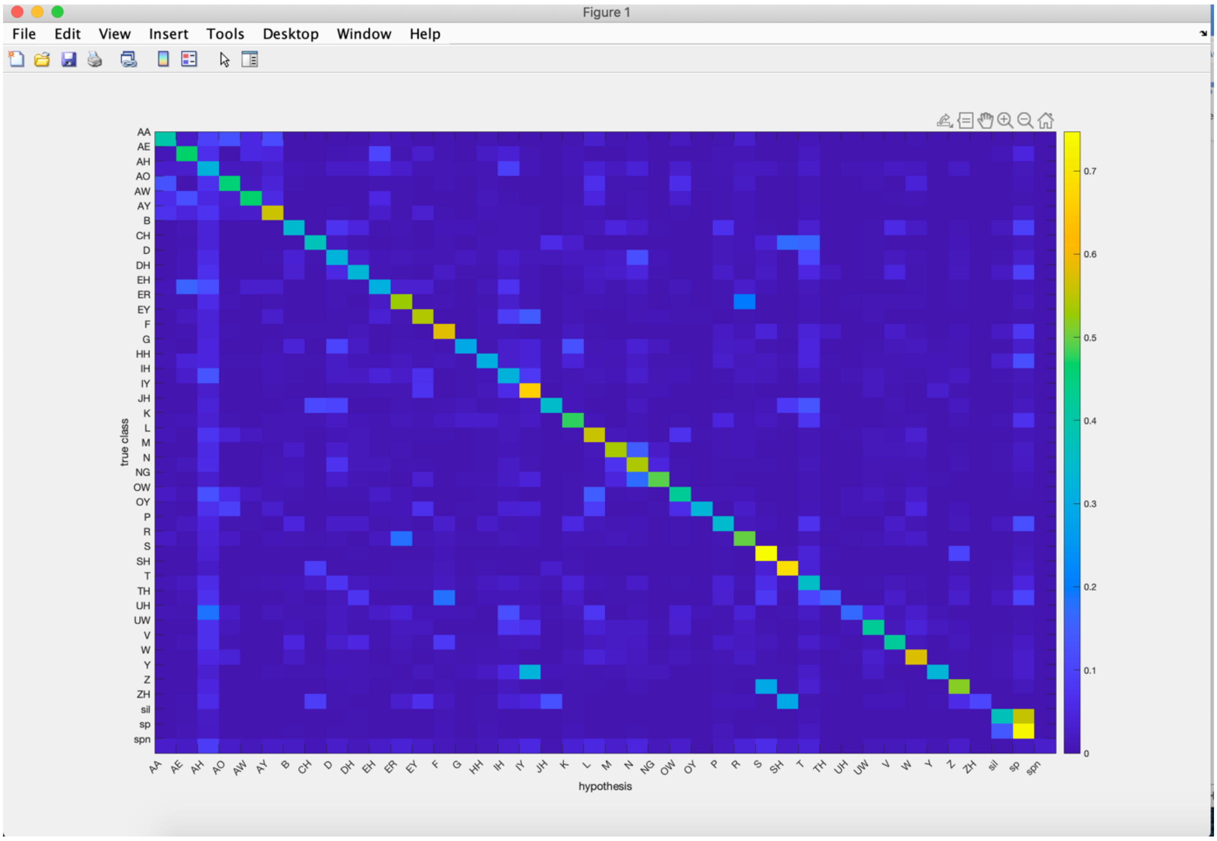


Figure 4: Confusion matrix for phone classification using Gaussian mixture model

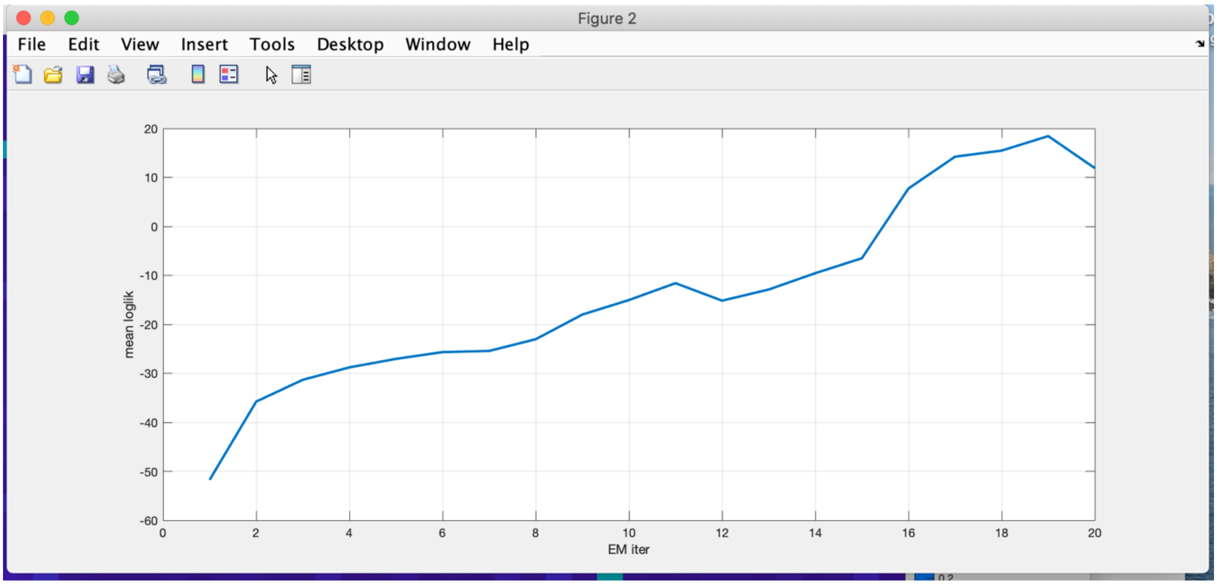
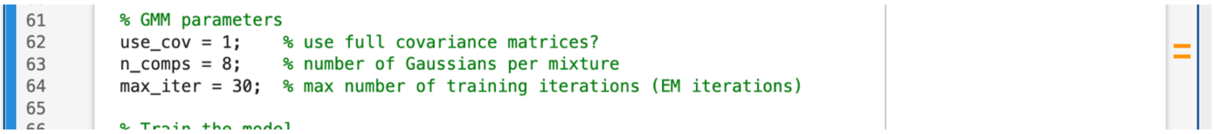


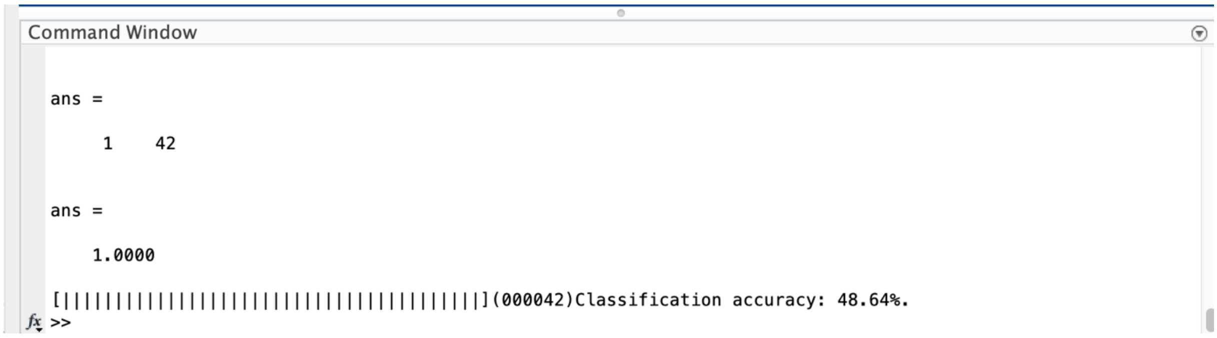
Figure 5: Mean of loglikelihood values in each iteration



```
% GMM parameters
use_cov = 1;     % use full covariance matrices?
n_comps = 8;     % number of Gaussians per mixture
max_iter = 30;   % max number of training iterations (EM iterations)

% Train the model
```

Figure 6: The hyperparameter values of GMM



```
ans =

     1    42


ans =

    1.0000

[||||||||||||||||||||||||||||||||||||||||](000042)Classification accuracy: 48.64%.
>>
```

Figure 7: Accuracy of phone classification using Gaussian mixture model

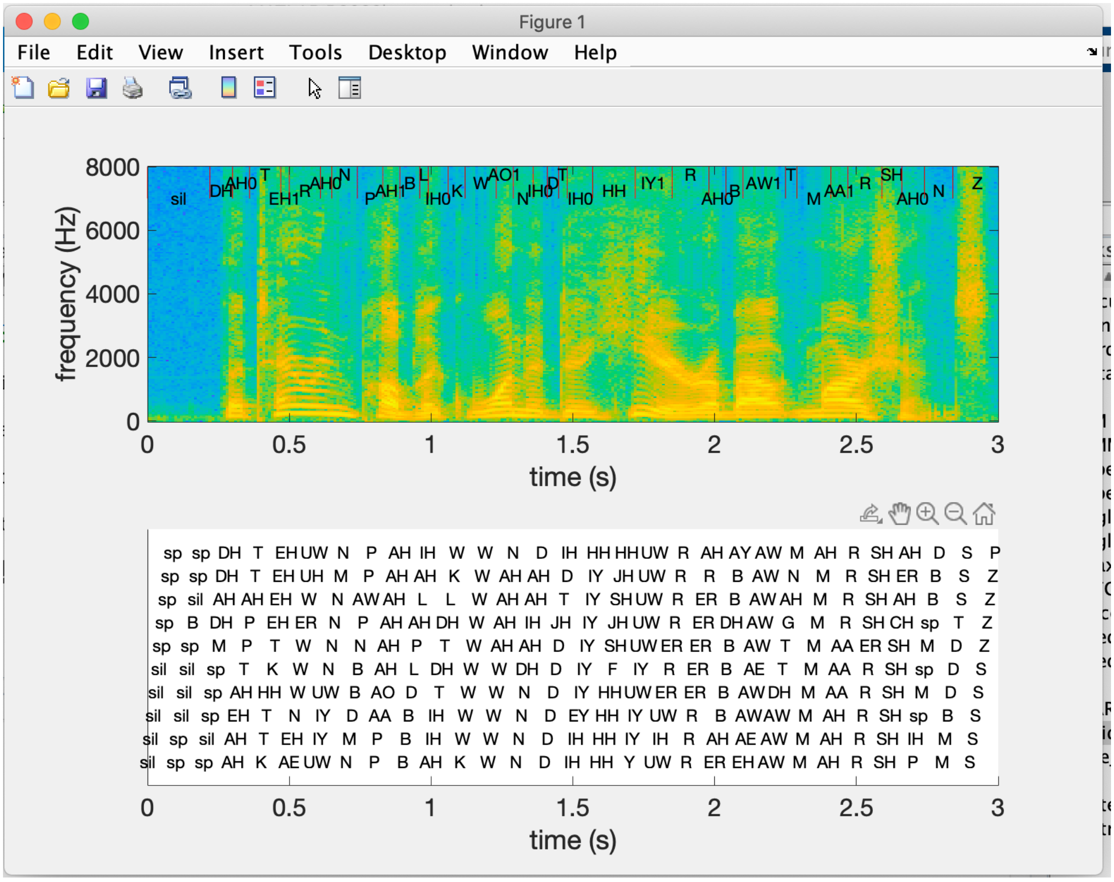


Figure 1
File
Edit
View
Insert
Tools
Desktop
Window
Help
8000
6000
4000
2000
0
frequency (Hz)
0
0.5
1
1.5
2
2.5
3
time (s)
sil
DH
AH0
T
EH1
R
AH0
N
P
AH1
B
L
IH0
K
W
AO1
N
IH0
D
T
IH0
HH
IY1
R
AH0
B
AW1
T
M
AA1
R
SH
AH0
N
Z
sp sp DH T EH UW N P AH IH W W N D IH HH HH UW R AH AY AW M AH R SH AH D S P
sp sp DH T EH UH M P AH AH K W AH AH D IY JH UW R R B AW N M R SH ER B S Z
sp sil AH AH EH W N AW AH L L W AH AH T IY SH UW R ER B AW AH M R SH AH B S Z
sp B DH P EH ER N P AH AH DH W AH IH JH IY JH UW R ER DH AW G M R SH CH sp T Z
sp sp M P T W N N AH P T W AH AH D IY SH UW ER ER B AW T M AA ER SH M D Z
sil sil sp T K W N B AH L DH W W DH D IY F IY R ER B AE T M AA R SH sp D S
sil sil sp AH HH W UW B AO D T W W N D IY HH UW ER ER B AW DH M AA R SH M D S
sil sil sp EH T N IY D AA B IH W W N D EY HH IY UW R B AW AW M AH R SH sp B S
sil sp sil AH T EH IY M P B IH W W N D IH HH IY IH R AH AE AW M AH R SH IH M S
sil sp sp AH K AE UW N P B AH K W N D IH HH Y UW R ER EH AW M AH R SH P M S
0
0.5
1
1.5
2
2.5
3
time (s)

Figure 8: Decoding output for a speech utterance using GMM and its spectrogram

Q2.1:

By using more than one Gaussian component to model the data, we can achieve to show the variability in the data within the phone classes more accurately. As I explain in Q1.2, in a single Gaussian distribution, we have same model parameters for a particular phone class, which limits us to show the variability in the data. However, in this case, by modeling each phone classes using several Gaussian components, we can show the small differences in the samples (in the feature vectors) of a particular class since each component have their own model parameters (mean, covariance, weight). And each captures the different aspect of the training data distribution within the particular phone class. Also, during the testing when we are finding the log-likelihoods of each class for every sample, we make use of these parameters. As a result, we would see better classification accuracy compared to the single multivariate Gaussian distribution.

Q2.2:

If we increase the number of Gaussian components, we can encounter with the overfitting problems. The aim in introducing "components" idea is that to show the variability in the data by each component having their own model parameters (mean, covariance, weight) within the particular phone class. By increasing the component number too much, our model might fit the training data too closely, and as a result, it does not perform well on the unseen test data in the testing phase. That means, the model has memorized the training data. Also, in the assignment it is written that as the number of Gaussian components increase, covariance of the component approaches zero and the determinant becomes 0. Which makes sense, because as the component number increases, there will be much smaller number of samples represented in each component, meaning that components start to specialize to individual data points. Since in our case, covariance represents the variability of the data in specific component, there will be less variability in that components and the covariance of it will approach zero.

Q2.3:

I think the classification confusions are highly similar in the GMM compared to single multivariate Gaussian distribution. When I put two confusion matrix side to side, the only difference was the coloring of the cells. In the off-diagonal cells, especially in places where we observe confusions as explained in Task 1, the colors of those cells become a bit darker. It indicates that the number of samples that are misclassified in each particular cell has decreased, which is promising as it shows GMM performed better than the single multivariate Gaussian distribution. The accuracy results also prove that.

Q2.4:

In speech, phones are not independent since a phone can be influenced by the previous and next phones of that phone. This is called coarticulation as we learnt during the lectures. But in GMMs, we assume that there are no temporal dependencies between frames, where a single phone denoted in each frame. So, in that case, the MFCC feature vectors for each frame would not fully capture temporal dependencies between frames (even though we use sliding window approach where some temporal dependencies can be captured). But in reality, acoustic properties of a phone also depend on where that phone is located (near which phone etc.) Hence, we can say that GMMs may not be preferable as they are not good at modeling temporal dependencies in speech. That is why Hidden Markov Models or Deep Neural Networks such as RNNs can be preferred instead. Additionally, in the assignment, it is written that GMM training is sensitive to the initial parameters and number of components as they affect the performance of the model. As it might be difficult to find optimal parameters that yields to good classification performance, this can be seen as one of the shortcomings of GMMs.