Assignment 1

Gengcong Yan - 1009903
ELEC-E5510 - Speech Recognition

November 9, 2021

1 Question 1

1.1 A

Quantifies the gross-shape of the spectrum in identification of vowels. At the same time, it removes fine spectral structure, which is often less important. It thus focuses on that part of the signal which is typically most informative.

It's relatively easy to understand and computationally efficient for calculation in a few steps. It has been widely used for many years in research and industry, proving their good performance.

1.2 B

In spectrograms,the range of values is very non-uniform. In fact, by directly looking at power spectra, we do not really see anything informative. It's because most sounds humans hear are concentrated in very small frequency and amplitude ranges. we need to transform them into proper scale. The signal in spectrogram forms peaks and valleys, which correspond to the resonances of the vocal tract. These peaks are known as formants and they can be used to uniquely identify all vowels. Capturing or quantifying such macro-level structures is important because of the connection with the vowel-identity. Cepstrogram can better illustrate the structures better.

2 Question 2

```
#Calculate all ACC

train_acc=[]
test_acc=[]

for n in range(20,31,1):
    S = train_gmm(train_data_normalized, train_class, n)

    predictions = predict(train_data_normalized, S, n)
    train_acc.append(accuracy_score(predictions, train_class))

    pref_test= predict(test_data_normalized, S, n)
    test_acc.append(accuracy_score(pref_test, test_class))
```

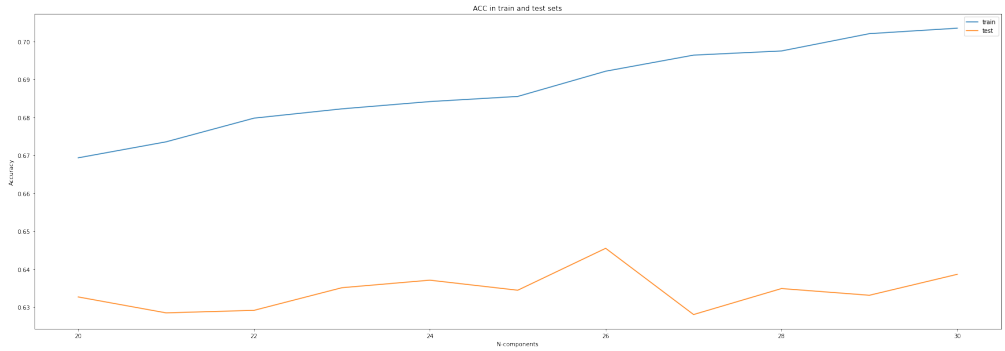

Figure 1: ACC in train and test sets

```
# Plot all ACC
x=list(range(20,31,1))
print(x)
plt.plot(x,train_acc, label = "train")

plt.plot(x,test_acc,label = "test")

plt.xlabel("N-components")
plt.ylabel("Accuracy")
plt.legend()
plt.title("ACC in train and test sets")
plt.show()
```

2.1 A

Because the train set and test set contains different samples. The training is based on train set, so naturally got better results of recognition. The test set is independent of training data, simulating the model's performance in a real environment. The error rate is normally higher.

2.2 B

From the plot in Fig1 we know, typically the more the number of components, the better the results in training. But when number reaching a maximum limit, there is no improvement anymore in test data. **I think 26 is a good number.**

3 Question 3

3.1 A

The lighter the color is, the more difficult the model can distinguish different phonemes. The models are not equally sensitive to recognizing different phonemes, and for some specific

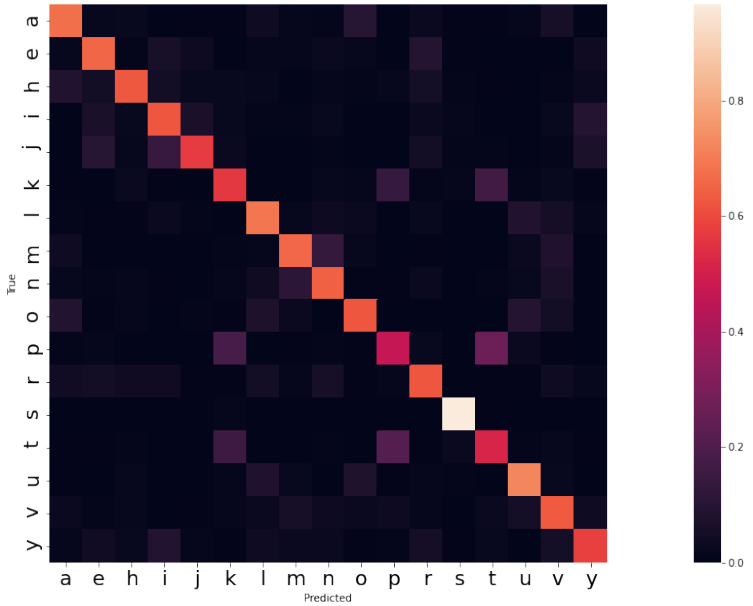

Figure 2: Confusion Matrix of GMM model

phonemes, it may be difficult for the model to distinguish them.

3.2 B

There are some pairs of phoneme hard to classify in the model: **t-p, t-k, u-o, v-m, v-n, p-k**

3.3 C

The visualized confusion matrix are shown in Fig.2.

4 Question 4

```
# tw1_normalized = z.transform(tw1)
# tw2_normalized = z.transform(tw2)
# tw3_normalized = z.transform(tw3)

best_n=30
best_S = train_gmm(train_data_normalized, train_class, best_n)
pred1 = predict(tw1_normalized, best_S,best_n)
pred2 = predict(tw2_normalized, best_S,best_n)
pred3 = predict(tw3_normalized, best_S,best_n)

print([phonemes[i]for i in pred1])
print("--------------------")
print([phonemes[i]for i in pred2])
print("--------------------")
print([phonemes[i]for i in pred3])
```

4.1 A

After converted the numbers in predictions to letters in phonemes, we can obtain 3 arrays correspond to 3 finish words, but there are a lot repetition in arrays. The 3 arrays are shown below:

1. 'p', 't', 'p', 't', 'k', 'k', 't', 'k', 't', 't', 'p', 't', 'p', 'k', 'k', 'k', 'k', 'k', 'k', 'y', 'k', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'u', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'o', 'l', 'l', 'l', 'l', 'l', 'l', 'l', 'l', 'l', 'a', 'o', 'o', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'i', 'i', 'i', 'e', 'e', 'e', 'e', 'i', 'y', 't'

2. 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 'y', 'e', 'e', 'e', 'y', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'i', 'i', 'i', 'i', 'y', 'i', 'i', 'i', 'i', 'i', 'i', 'i', 'i', 'i', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 't', 't'

3. 'h', 'e', 'h', 'h', 'k', 'h', 'h', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'e', 'l', 'l', 'l', 'l', 'i', 'r', 'n', 'm', 'm', 'm', 'm', 'm', 'm', 'm', 'v', 'v', 'v', 'r', 'i', 'i', 'i', 'y', 'k', 'k', 't', 'p', 'k', 'k', 'k', 'k', 'u', 'k', 'k', 'u', 'v', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'u', 'o', 'u', 'u', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 's', 'a', 'v', 'o', 'a', 'a', 'a', 'a', 'a', 'a', 'v', 'v', 'n', 'n'

Since I don't know Finnish, I tried to search for the most likely words on google. **I think they are kolme (3), sisu, Helsinki.**

4.2 B

There are a lot of repeated letters in the results based on frame classification. It's hard to know the exact length of the word.

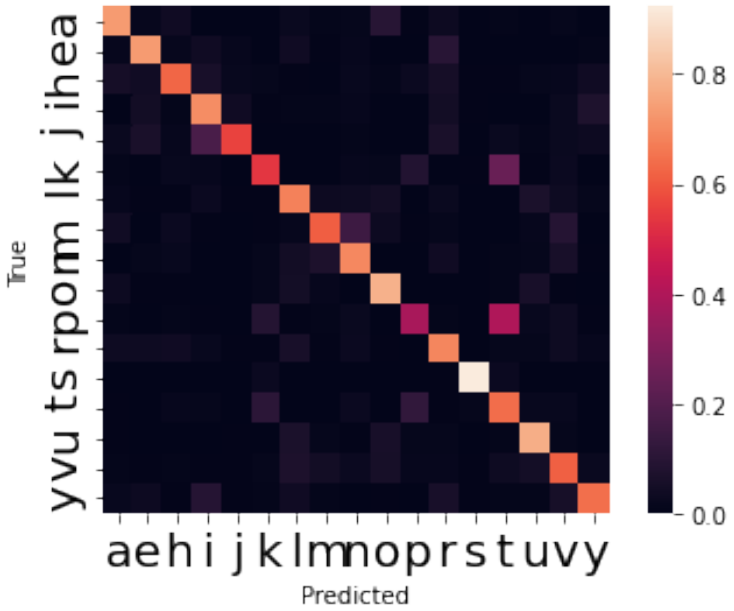

Figure 3: Confusion Matrix of DNN model

4.3 C

We can improve the way we extract features, try to avoid extracting duplicate data, for example, avoid overlapping when windowing the data. But sometimes this does not preserve enough information in the data.

5 Question 5

5.1 Question 5.1

DNN is better with accuracy 67.82%, whereas the accuracy of best GMM is around 63%. The visualized confusion matrix of DNN model are shown in Fig.3.

5.2 Question 5.2

I sum all the parameters in the dictionary returned by 'train_gmm': $17 \times (26 \times 26 + 26 \times 26 + 1 + 26) = 23443 < 142865$. But the total number of parameters in the MLP model is 142865. So DNN requires more parameters.

6 Question 6

6.1 A

GMM has lower classification error. Because these sample data are obtained from the GMM model,now we test them in the same model where they are produced, They can naturally

receive better results. We are using the knowledge of the model we gonna test to generate data that is certainly more relevant to it. The results does not represent the true performance of the model.

GMM is a generative approach which learn prominent features based on the joint probability distribution on given observable data. MLP is a discriminative approach is to just look at a bunch of data and figure out what the differences are. So it's hard to ask the discriminative model generating sampled data.