Biometric Person Authentication – Seminar
Summer Semester 2023
Quality and Usability Lab, TU Berlin

Assignment 2 – Due on 19 June 2023 at 12:00 CEST

7 Questions a) to g), Total marks = 8

Perform a text-dependent speaker identification task. You are given a speech signal for the word "six", spoken by an unknown speaker, 2 speech signals for the word "six", spoken by known speaker M1, and 2 speech signals for the word "six", spoken by known speaker M3. Based on the central vowel /ɪ/[1] in "six", you need to decide whether the utterance is more likely to have come from M1 or from M3. To this purpose, you have to identify the vowel frames of all the signals, train speaker models for M1 and M3, using the speaker's vowel frames, and determine the likelihood ratio of the unknown speaker's vowel frames against the M1 and M3 models.

Training files for Speaker M1: 06M1SET0.wav and 06M1SET1.wav
Training files for Speaker M3: 06M3SET0.wav and 06M3SET1.wav
Test file from unknown speaker: 06X1S1T0.wav

PART I – FEATURE EXTRACTION

In sub questions a) to c), you extract 12-dimensional cepstral vectors for the vowel /ɪ/ from all 5 speech files.

a) Perform speech analysis on the 4 training files AND the test file to compute the short-time energy E_i in dB, and the voicing parameter VC_i for each frame i of each of the 5 signals.
Print the arrays E and VC for each of the 5 files. [2 marks]

Use the following analysis parameters, given sampling frequency fSampl=12500 [Hz]:
Frame duration: `frameDuration = 0.02048` [s] or `frameSamples=256`;
Frame shift: `frameShift = 0.01024` [s] or `shiftSamples=128`;
Energy threshold: `eThr = -45` [dB];
Voicing threshold: `vcThr = 0.4`;
Minimum fundamental frequency: `f0Min = 80` [Hz];
Maximum fundamental frequency: `f0Max = 200` [Hz];

Note: The first frame comprises Samples 1 to 256, the second frame Samples 129 to 384 etc. Voicing is the maximum signal autocorrelation for the frame over the range of lags from `kmin=floor(fSampl/f0Max)` to `kmax=ceil(fSampl/f0Min`.

b) Assume that all signal frames with E_i > -45 *and* VC_i > 0.4 belong to the vowel (because they are of high energy and are voiced!).
For each of the 5 files, print the frame numbers of all vowel frames. [1 mark]

[1] International Phonetic Alphabet (IPA) notation for the vowel in the word "six"

c) Perform speech analysis on all 5 signals by multiplying with a Hamming window each of the vowel frames that you determined in b) and then computing a 12-dimensional cepstral vector c_1 to c_{12} for the Hamming-windowed frame.
For each file, print the first and last cepstra with the file name and frame number of each.
[2 marks]

Hint: You can use the Matlab/Octave function `rceps()` to compute the full ("real") cepstrum c_0 to c_{255} of a frame. However, you only need the values c_1 to c_{12} (ignore c_0). Remember that Matlab/Octave indices run from 1 to 256, ***not*** from 0 to 255 and c_0 is the first element!

At this point you have determined a 12-dimensional feature (cepstrum) vector for each vowel frame in the 4 training files AND in the test file.

PART II – MODEL BUILDING FROM THE TRAINING DATA

Next, you build a Gaussian model for each known speaker from that speaker's vowel cepstra.

d) Determine each speaker's 12-dimensional Gaussian PDF for the vowel by determining the mean vectors muM1 and muM3 and the diagonal covariance matrices sigmaM1 and sigmaM3.
For each PDF, print the mean vector and covariance matrix. [1 mark]

At this point you have completed the PDFs for the vowel /ɪ/ for each of speakers M1 and M3.

PART III – COMPARISON OF THE TEST DATA AGAINST THE 2 MODELS

For each vowel frame of the test file, you must now calculate the likelihood of the frame belonging to speaker M1 and the likelihood of the frame belonging to speaker M3.

e) For each vowel cepstrum $\boldsymbol{c}_j$ of the test file, calculate the 2 log likelihoods $\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M1}, \boldsymbol{\Sigma}_{M1})$ and $\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M3}, \boldsymbol{\Sigma}_{M3})$.
Print all log likelihoods you have obtained with the file name and frame number of each.
[1 mark]

Hint: You can use the Matlab/Octave function `mvnpdf()` to compute the likelihoods $p(\boldsymbol{c}_j|.)$.

f) Determine the mean log likelihoods $\overline{\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M1}, \boldsymbol{\Sigma}_{M1})}$ and $\overline{\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M3}, \boldsymbol{\Sigma}_{M3})}$ over the test file vowel cepstra $\boldsymbol{c}_j$, the log likelihood ratio $LLR = \overline{\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M1}, \boldsymbol{\Sigma}_{M1})} - \overline{\ln p(\boldsymbol{c}_j|\boldsymbol{\mu}_{M3}, \boldsymbol{\Sigma}_{M3})}$ and finally the likelihood ratio $LR = e^{LLR}$.
Print the 2 mean log likelihoods, the log likelihood ratio and the likelihood ratio.
[0.5 marks]

PART IV – CONCLUSION

g) At this point you have completed the speaker identification problem. State the solution both in terms of which of the 2 speakers M1 and M3 is the likely speaker of the test utterance and in terms of the confidence you have in that decision.
Write a paragraph in plain English, interpreting your result. [0.5 marks]

Submission

A single zip archive Ass2.zip which comprises

- a Matlab file Ass2.m with the code of your main script;
- a Matlab file extractCepstra.m with the code of your feature extraction function;
- a text file Ass2.txt with the unredacted output of your Matlab program; and
- a document Ass2.docx or Ass2.pdf with your answer to Q.g).

Matlab Notes

1. In Q.a), autocorrelations $\boldsymbol{r}(k)$ need to be calculated. Matlab function `r=xcorr(x, kMax, coeff)` returns a vector of size $2k_{max} + 1$ containing $\boldsymbol{r}(-k_{max}, \cdots, -1,0,1 \cdots, k_{max})$. For our purpose, we only require $\boldsymbol{r}(k_{min}, \cdots, k_{max})$ where $k_{min} = f_{sampl}/f0_{max}$ corresponds to the largest expected fundamental frequency, and $k_{max} = f_{sampl}/f0_{min}$ to the smallest expected fundamental frequency. Use Matlab functions `floor()` and `ceil()` to round k_{min} down and k_{max} up, respectively, and take care to determine the required maximum autocorrelation of the frame for the correct range of Matlab vector `r`.
2. In Q.c, you may use Matlab functions `hamming()` to multiply a frame with a Hamming window and `rceps()` to compute the real cepstrum c_0 to c_{255} of a frame of 256 samples. Of those you only need c_1 to c_{12}, ignoring c_0. Therefore, the values c_1 to c_{12} are the second to thirteenth elements of the Matlab array returned by `rceps()`.
3. For Q.d, you should use the Matlab functions `mean()` and `cov()` as in Assignment 1. As you did in Assignment 1, use unbiased sample statistics for the calculation of the covariances. To make the 2 matrices' diagonal, use the function `diag()`.
4. For Q.e, you should use the Matlab function `mvnpdf(.)` as in Assignment 1.
5. Remember to avoid **unnecessary** loops for vector or matrix operations in Matlab!

--- End of Assignment 2 ---