Machine learning for speech and computer vision

Case Study 1, Feb. 2022

A single zip file named as 'RollNumber_yourname_case1.zip' should be uploaded for getting the credits. The zip file should include code files (one for each question), audio files, and a single report file (.pdf file) showing all the figures and appropriate descriptions as asked in the questions. You can use MATLAB or Python for implementing the case study.

This case study corresponds to 5 % of the total marks.

1. Record the utterance "My name is "your-name" **in your own voice** with a sampling frequency of 16 kHz. Normalize the speech data to lie between -1 and +1.

   a) Save this utterance to a wave file named 'your-name_record.wav'

   b) Read back the file 'yourname_record.wav' to a variable (speech signal) and plot the speech signal with time as the x axis. (Note: sampling period = (1/sampling frequency)). What is the relation between the time duration of the utterance and the total number of samples in the utterance?

   c) Play and listen to the utterance using the sampling frequency.

   d) Now multiply the variable with 0.5 and 2 (do not normalize after multiplying), and listen to the sounds. Describe the difference between these sounds and the sound played previously in part c.

Note: Please use your first name or last name for 'your name'

2. Record two utterances **in your own voice**: the first utterance is three cardinal vowels a, i, u uttered in sequence, the second utterance is "She had your dark suit in greasy wash water all year" Obtain a spectrogram (use the FFT function to obtain after splitting the speech signal into frames) for both the utterances using Hamming window of length 30 ms (narrow band spectrogram) and 5 ms (wideband spectrogram) for window shift of 5 ms, and FFT length of 512.

   a) Plot the speech waveform, wideband, and narrowband spectrogram (as grayscale image) as subplots (use 'subplot' function in MATLAB) of a figure one below the other with time alignment for both the utterances.

   (**Help:** Use 'imagesc (time, freq, 20*log10(abs(fftspectra)))' function to convert fftspectra values of all frames into an image and it also displays it as image. Use sgmap=1.0- colormap(gray(128) followed by colormap(sgmap) to get the spectrogram in grayscale)

   b) Note down the first three formant frequencies for three cardinal vowels of the first utterance using the spectrogram. Also, obtain the magnitude spectrum at the center of three vowels. Obtain the first three formant resonant frequencies and pitch frequency from the spectrum for all three vowels.

   c) List the set of sounds heard in the IPA notation for the second utterance recorded. Briefly describe spectral characteristics for stop consonants (/k, t/), fricatives (/s, ʃ/**),** and semivowels (/w, r/). The description should match the spectrogram shown.

3. Obtain the time-domain features for the **second utterance** recorded in **question 2**.

   a) Obtain the short-time energy and short-time magnitude using a window length of 25 ms with a window shift of 10 ms. Plot both features overlapping on the speech waveform. Obtain the short-

time magnitude for window lengths of 35 ms and 45 ms with the window shift of 10 ms. Explain the effect of the change in window length on the short-time magnitude.

b) Obtain the short-time zero-crossing rate (ZCR) for a window length of 25 ms and window shift of 30 ms. Plot the ZCR overlapping on the speech waveform.

c) Obtain and plot the short-time autocorrelation for the center frame of vowel /a/ in word 'dark' and obtain the pitch period based on the autocorrelation function.

d) Briefly describe how the short-time energy/magnitude and short-time ZCR can be used to obtain the boundaries for voiced and unvoiced sounds using example values in the utterance.

4. Obtain the first 13 MFCC, delta MFCC, and delta-delta MFCC features for the **first utterance** recorded in question 2 using the method described in the class with 26 mel-filters, window length of 25 ms, and window shift of 10 ms. Use T=2 for calculating the delta and delta-delta coefficients, and α=0.9 for pre-emphasis. You should explain the code for MFCC computation if asked during evaluation.

a) Calculate the distance between the MFCC values at the center frame of the three cardinal vowel sounds. i) distance between /a, i/, ii) distance between /a, u/, and iii) distance between /i, u/. Explain your observation on the distance values.

For distance calculation, ignore the first coefficient and use the remaining 12 coefficients and calculate the distance as $\frac{10}{\ln(10)}\sqrt{2\sum_{l=1}^{12}\left(mfcc_1(l) - mfcc_2(l)\right)^2}$, where $mfcc_1$ and $mfcc_2$ are the MFCC coefficients for two vowels.

b) Obtain log-filterbank features at the center frame for three cardinal vowel sounds. Plot the log-filterbank features versus the frequency (linear frequency value) of the center frequencies of the filter-bank.

c) Compute log-filterbank features using 40 mel-filters for the center frame for three cardinal vowels. How do the log-filterbank features change when compared to 26 filters as in part b?

Bonus question (This is not mandatory. Marks for this question will be added if the marks scored in first four questions is less than maximum marks. This will correspond to 10% of maximum marks for the case study)

5) Obtain the linear prediction coefficients at the center frame for three cardinal vowel sounds using the first utterance of question 2. Use the 'lpc' function of MATLAB to compute the coefficients. Obtain the LPC spectrum for three vowels at the center frame using the vocal tract transfer function assuming G=1.