E9 261 - Speech Information Processing

Homework # 4
Due Date: 11:59PM, April 18, 2023

Please upload (in the course webpage) your codes (in multiple zip files with filenames having part1 part2 etc. each not exceeding 10Mb). In the zipped folder, the program names with README should be self explanatory. Mention the uploaded file names in your answer book. Filename of each program should contain the question number it is associated with.
The recording must be uploaded in the folder with your name in the following drive folder:
https://drive.google.com/drive/folders/1rJia9ksRuJVcDWsEGWuq0dCQBwc6W4sP?usp=share_link
This is required as all of you need to use each other's recordings in Part2 of this homework.

Part1

You will carry out consonant classification in this homework and find out the role of different features and distance measures for the classification task. Six different consonants will be used for this purpose, namely, /f/ (as in 'cuff'), /s/ (as in 'kiss'), /ch/ (as in 'church'), /v/ (as in 'vivid'), /z/ (as in 'zoo'), and /j/ (as in 'jam'). There will be four different classification tasks (CTs) that you need to perform as follows:

- Classification Task 1 (CT1): Two-class classification with class 1 (C1) comprising /f/, /s/, /ch/ and class 2 (C2) comprising /v/, /z/, /j/.
- Classification Task 2 (CT2): Three-class classification where C1: /f/, C2: /s/, and C3: /ch/
- Classification Task 3 (CT3): Three-class classification where C1: /v/, C2: /z/, and C3: /j/
- Classification Task 4a (CT4a): Direct six-class classification where C1: /f/, C2: /s/, C3: /ch/, C4: /v/, C5: /z/, and C6: /j/
- Classification Task 4b (CT4b): Hierarchical six-class classification where C1: /f/, C2: /s/, C3: /ch/, C4: /v/, C5: /z/, and C6: /j/. Here, you need to first run CT1. Depending on the predicted class from CT1, run either CT2 or CT3.

For this purpose, please record (at 16kHz sampling rate) vowel-consonant-vowel (VCV) spoken by yourself where C is one of the six consonants (/f/, /s/, /ch/, /v/, /z/, and /j/) and V is one of the following six vowels: V1 (as in Hid), V2 (as in Head), V3 (as in Had), V4 (as in Hudd), V5 (as in Hod), and V6 (as in Hood).

For a chosen C, you need to record the following six VCVs, each repeated FIVE times:

V1CV1, V2CV2, V3CV3, V4CV4, V5CV5, and V6CV6

Hence, for a chosen C you need to record and prepare the following 30 wav files:

V1CV1_1.wav, V1CV1_2.wav, V1CV1_3.wav, V1CV1_4.wav, V1CV1_5.wav, V2CV2_1.wav, V2CV2_2.wav, ..., V6CV6_5.wav

You need to repeat this for all six choices of C, i.e., /f/, /s/, /ch/, /v/, /z/, and /j/. Thus, a folder containing 30x6=180 recorded wavfiles needs to be submitted and used for the experiments. The naming convention should be following the list given above, i.e., V1pV1_1.wav, V1pV1_2.wav etc.

For each of the 180 wavfiles, you need to mark the following three timestamps:

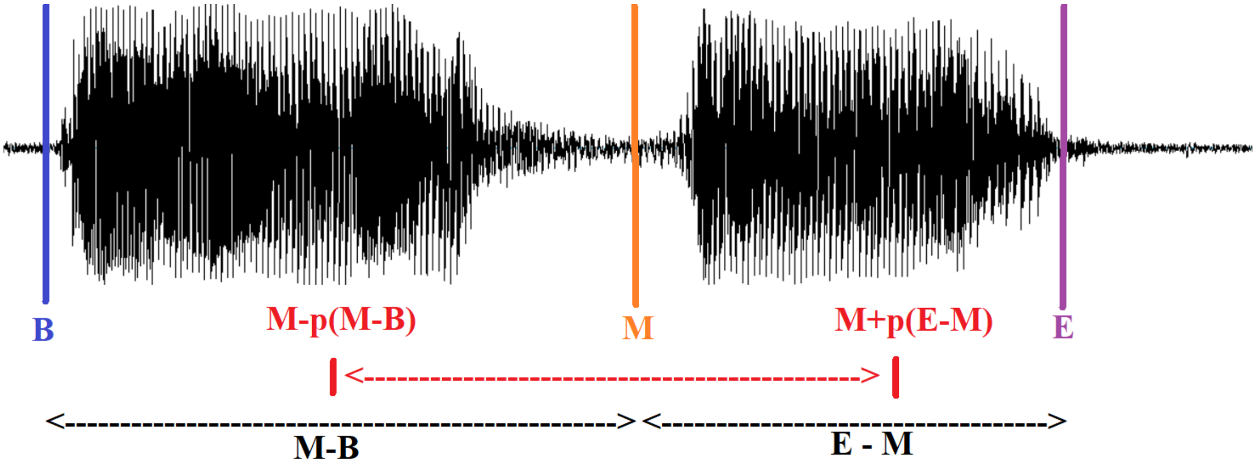
1. Begin (B) of pre-consonant vowel
2. Middle (M) of the consonant
3. End (E) of post-consonant vowel

You need to mark these timestamps using Audacity software (use of any other softwares is also fine) by marking a segment from B to M (VC) and marking another segment from M to E (CV). For every wavfile, you need to prepare a text file (e.g., V1pV1_1.txt, V1pV1_2.txt etc.) containing the timestamps. These 180 timestamps files should be submitted together with the wavfiles. For the format of the text files, please have a look at an example recording and timestamps marking, when V1 is used and /z/ is chosen as C, at the following location:

https://drive.google.com/drive/folders/1zQUB1Y87rKmz7q1GY9sbFgZSBdt1PfQf?usp=sharing

For all the classification tasks, first THREE (*_1.wav, *_2.wav and *_3.wav) of the five repetitions should be used in training and remaining TWO (*_4.wav and *_5.wav) for every vowel and consonant combination should be used for testing.

You need to experiment with various segment lengths from each VCV recording centered around the middle of the consonant as illustrated in the figure below. A segment (as shown by the red colored part in the figure below) starts from M-p(M-B) and ends at M+p(E-M). Thus, the segment length is parameterized by p (0<p<1). When p=1, the entire VCV from B to E is used. You need to experiment with four values of p, namely, 0.25, 0.50, 0.75 and 1.0.



For the classification task, you need to consider a short-time feature (D-dimensional) every 20msec with a 10msec shift. Thus, each VCV recording will be represented as a sequence of D-dim features, i.e., a feature matrix, the length of which varies from one recording to another and also depending on the choice of p. You need to use the K Nearest Neighborhood (KNN) classifier for all the classification tasks. The distance metric to be used in the KNN classifier is dynamic time warped (DTW) distance between the test feature matrix and a training feature matrix. Choose K=10 in the KNN classifier.

A. Use the short-time spectrum computed using FFT order N as the feature vector. Vary N=128, 256, 1024. Experiment separately with the Euclidean and the Itakura Saito distance measure in computing DTW distance. For each of these two distance measures, report the classification accuracies for all combinations of p and N (in a table with rows and columns having different values of p and N, respectively). Do this separately for CT1, CT2, CT3, CT4a, CT4b. Interpret the classification accuracies you have obtained. In particular, which between the CT4a and CT4b results in the highest six class classification accuracy and why?

B. Use the short-time real cepstrum followed by liftering of order L as the feature vector. Vary L=5, 10, 20. Experiment separately with the Euclidean distance between two cepstral vectors and the Itakura Saito distance between spectral envelopes computed from liftered cepstrum as the two distance measures in computing DTW distance. For each of these two distance measures, report the classification accuracies for all combinations of p and L (in a table with rows and columns having different values of p and L, respectively). Do this separately for CT1, CT2, CT3, CT4a, CT4b. Interpret the classification accuracies you have obtained. In particular, which among the CT4a and CT4b results in the highest six class classification accuracy and why?

C. Use short-time linear prediction coefficients (LPCs) of order N as the feature vector. Vary N=10, 15, 20. Experiment separately with the Euclidean distance between two LPC vectors and the Itakura Saito distance between LPC modeled spectral envelopes as the two distance measures in computing DTW distance. For each of these two distance measures, report the classification accuracies for all combinations of p and N (in a table with rows and columns having different values of p and N, respectively). Do this separately for CT1, CT2, CT3, CT4a, CT4b. Interpret the classification accuracies you have obtained. In particular, which among the CT4a and CT4b results in the highest six class classification accuracy and why?

D. Use any short-time feature of your choice and report the classification accuracies for various values of p. Do this separately for CT1, CT2, CT3, CT4a, CT4b. Interpret the classification accuracies you have obtained. In particular, which among the CT4a and CT4b results in the highest six class classification accuracy and why? Does this feature yield a six-class classification accuracy better than the features in parts (A), (B), & (C)? Which among the features in parts (A), (B) and (C) performs the best and why?

If you are not implementing features or distance measure or classifier by yourself, please cite a reference and/or codebase and/or tool that you use.

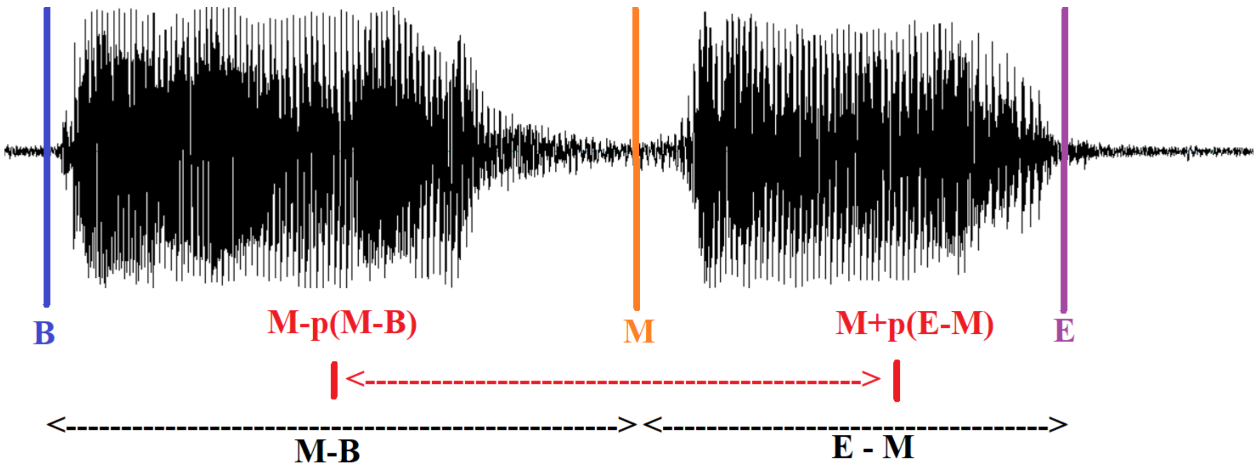
Part2

Now you will use the recordings from all students in the class to carry out CT2 and CT3.

Also, unlike KNN with DTW distance measure in Part1, you will use discrete and continuous observation density hidden Markov model (HMM) [1] as well as time delay neural network (TDNN) [2] as discussed in the class.

For both the classification tasks, first THREE (*_1.wav, *_2.wav and *_3.wav) of the five repetitions from all students should be used in training and remaining TWO (*_4.wav and *_5.wav) for every vowel and consonant combination should be used for testing.

You need to experiment with various segment lengths from each VCV recording centered around the middle of the consonant as illustrated in the figure below. A segment (as shown by the red colored part in the figure below) starts from $M-p(M-B)$ and ends at $M+p(E-M)$. Thus, the segment length is parameterized by p ($0<p<1$). When $p=1$, the entire VCV from B to E is used. You need to experiment with four values of p, namely, 0.25, 0.50, 0.75 and 1.0.



For the classification task, you need to consider a 39-dim MFCC with delta and delta delta coefficients computed with a window size of 20msec with a 10msec shift. Thus, each VCV recording will be represented as a sequence of 39-dim features, i.e., a feature matrix, the length of which varies from one recording to another and also depending on the choice of p.

A) Perform K-means clustering to quantize the MFCC features from all training VCV recordings using the centers of K clusters. Vary K=16, 32, 64, 128. Use the quantized features as the observations and build a discrete observation density HMM separately for each class. The classification during the test phase is done using a maximum likelihood criterion. Vary the number of states N in HMM as 3, 5, 7. Report the unweighted average recall for both CT2 and CT3 for different values of K and N. For the best case, report the confusion matrices as well. Summarize your observations.

B) Repeat part A) using original MFCC features and continuous density HMM where the observation density is modeled using Gaussian mixture model (GMM). Use a diagonal covariance matrix in GMM and vary the number of mixtures (M) as 4, 8, 16, 32. Vary the number of states N in HMM as 3, 5, 7. Report the unweighted average recall for both CT2 and CT3 for different values of M and N. For the best case, report the confusion matrices as well. Summarize your observations.

C) Use TDNN [2] to carry out the classification tasks, CT2 and CT3. You can linearly interpolate the feature matrices to be of the same length, if required. Report the unweighted average recall for both CT2 and CT3 together with their confusion matrices. Vary the hyperparameters and report the results of your experiments. Alternatively, you can experiment with any other neural network architecture to improve the classification accuracy. Describe neural network you may use in detail.

D) Compare performance in part A), B), C) together with the KNN classifier with DTW distance measure from Part1. Summarize your observations.

E) Is there any other way you propose to further improve the classification accuracy for both CT2 and CT3?

For implementation of HMM and Neural Networks, you can choose suitable toolboxes (e.g. [Kevin Murphy's toolbox](https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html) [3] for HMM).

<u>References:</u>

[1] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

[2] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, *37*(3), 328-339.

[3] https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html