ECE 537 Fundamentals of Speech Processing
Problem Set 4

UNIVERSITY OF ILLINOIS
Department of Electrical and Computer Engineering

Assigned: Friday, 9/30/2022; Due: Monday, 10/3/2022
Reading: Velichko & Zagoruyko, "Automatic Recognition of 200 Words," 1970

1. The five bandpass filters used by Velichko & Zagoruyko have relatively wide bandwidth, compared to the use of a spectrum based on the fast Fourier transform. It turns out that Euclidean distance between spectra is only reasonable if the filters have relatively wide bandwidth. This problem will explore the reasons.

   Suppose that you have two utterances of the same vowel, $s_1[n]$ and $s_2[n]$. Suppose that both $s_1[n]$ and $s_2[n]$ are examples of the same vowel, with Dudley vocoder amplitudes of $A_1 = 1$, $A_2 = 10$, $A_3 = 2$, $A_4 = 2$, $A_5 = 0$, $A_6 = 3$, $A_7 = 3$, and no energy above 2100Hz (remember that, in a Dudley vocoder, spectral amplitude A_ℓ scales frequency components in the range $300(\ell - 1) < f < 300\ell$). The difference between them is that $s_1[n]$ has a pitch period of N_1 samples, while $s_2[n]$ has a pitch period of N_2 samples.

   (a) (1 point) Assume an analysis window of length N samples, where N is the least common multiple of N_1 and N_2, and the discrete Fourier transform (DFT) is

   $$S_i[k] = \sum_{n=0}^{N-1} s_i[n] e^{-j\frac{2\pi kn}{N}}$$

   Assume that the integers $\frac{N}{N_1}$ and $\frac{N}{N_2}$ have no common multiple, i.e., none of the harmonics of $X_1[k]$ are in the same DFT bin as any harmonic of $X_2[k]$. Define the following normalized spectra:

   $$F_1[k] = \frac{|S_1[k]|}{\sqrt{\sum_{k'=0}^{N-1} |S_1[k']|^2}}, \quad F_2[k] = \frac{|S_2[k]|}{\sqrt{\sum_{k'=0}^{N-1} |S_2[k']|^2}}$$

   What is the Euclidean distance between these two normalized spectra? Is this a large number or a small number?

   Solution: The Euclidean distance is

   $$\rho = \sqrt{\sum_{k=0}^{N-1} |F_1[k] - F_2[k]|^2}$$

   Since $F_1[k] > 0$ only when $F_2[k] = 0$ and vice versa, this is just

   $$\rho = \sqrt{\sum_{k=0}^{N-1} F_1^2[k] + \sum_{k=1}^{N-1} F_2^2[k]} = \sqrt{2},$$

   which is the largest possible value of this measure. There is no other combination of spectra that would have a larger value.

(b) (1 point) Now, instead of using a DFT, use the spectral energy features defined by Velichko & Zagoruyko. In order to make the calculation easier, assume that $s_1[n]$ and $s_2[n]$ each has the same fraction of its harmonics within each of the five V& Z sub-bands. Find ρ, the Euclidean distance between their log sub-band energies.

Solution: If each has the same fraction of its harmonics in sub-band 1, and if those harmonics are scaled by gthe same amplitude $A_1 = 1$, then $\ln\left(\frac{E_0}{E_1}\right)$ will have the same value for each of these two vowels. The same reasoning applies to the other four sub-bands. These two vowels therefore have exactly the same feature values, so that $\rho = 0$.

2. Suppose that you have a word that is two frames long, and a word that is three frames long. The spectral similarities $a_{i,k}$ are as shown in the following table, where i is the row index and k is the column index:

$a_{i,k}$	1	2	3
1	0.484	0.153	0.464
2	0.405	0.624	0.146

What is the best alignment, and what is the average spectral similarity computed along the best alignment?

Solution: Setting up the cumulative similarity matrix, initializing it with an all-zero first row and column, and taking the max path to each row, we get:

$a_{i,k}$	0	1	2	3
0	0	0	0	0
1	0	0.484	0.484	0.484
2	0	0.484	1.108	1.108

The best alignment is therefore $((0,0),(1,1),(2,2),(2,3))$, and its average spectral similarity is

$$\frac{1}{2}(1.108) = 0.554$$