Voice recognition with a linear time invariant system

Alejandro Ribeiro

February 23, 2021

To recognize spoken words we can compare the spectra of prerecorded words with the spectrum of signals observed at classification time. More specifically, say we want to discern between the spoken word "one" and the spoken word "two." We do so by recording K waveforms y_i for the spoken word "one" and K waveforms z_i for the spoken word "two." We normalize the respective signals to have unit energy and compute their DFTs to get

$$Y_i(k) = \mathcal{F}\left(\frac{y_i}{\|y_i\|}\right) \quad \text{and} \quad Z_i(k) = \mathcal{F}\left(\frac{z_i}{\|z_i\|}\right). \tag{1}$$

These DFTs are stored to construct the training sets $\mathcal{Y} := \{Y_i\}_{i=1,...,K}$ and $\mathcal{Z} := \{Z_i\}_{i=1,...,K}$.

With the training sets acquired, we proceed to observe a new signal x and compare the DFT $X = \mathcal{F}(x)$ with the DFTs contained in the training sets $\mathcal{Y}$ and $\mathcal{Z}$. To that end, we define the score function $q(X_1, X_2)$ that compares the DFTs $X_1(k)$ and $X_2(k)$ of two signals x_1 and x_2 of length N as

$$q(X_1, X_2) = \sum_k |X_1(k)|^2 \cdot |X_2(k)|^2, \tag{2}$$

for k in any contiguous set of N integers (e.g., $k = -N/2 + 1, ..., N/2$ or $k = 0, \ldots, N-1$). Note that the score function in (2) is slightly different than the one from Lab 6. This change allows us to interpret the score function $q(X_1, X_2)$ as the energy of the linear filtering of the signal x_1 by a filter with frequency response $H_2 = |X_2|$. Indeed, if we filter the x_1 using $h_2 = \mathcal{F}^{-1}(H_2)$, the resulting signal $x_3 = x_1 * h_2$ has spectrum $X_3 = \mathcal{F}(x_3)$ given by

$$X_3 = X_1 H_2 = X_1 |X_2|, \tag{3}$$

whose energy $\|X_3\|^2$ is given by (2). This filtering interpretation can be used to propose a time-domain implementation of voice recognition that does not involve computation of the DFT of the signal x_1. Due to Parseval, it amounts to filtering x_1 by h_2 and computing the energy of the output, i.e., $\|x_3\|^2$.

1 Comparison with average spectrum. For each of the training sets define the average spectra

$$\bar{Y} = \frac{1}{K}\sum_{i=1}^{K}|Y_i| \quad \text{and} \quad \bar{Z} = \frac{1}{K}\sum_{i=1}^{K}|Z_i|. \tag{4}$$

Interpret $H_y = \bar{Y}$ and $H_z = \bar{Z}$ as frequency responses of filters $h_y = \mathcal{F}^{-1}(H_y)$ and $h_z = \mathcal{F}^{-1}(H_z)$. Determine these impulse responses and use them to compute the score functions $q(X, \bar{Y})$ and $q(X, \bar{Z})$ *without using the DFT* of the signal x. Assign the signal to the digit with the largest score function.

2 Online operation. An advantage of the implementation in Part 1, as opposed to the method using the DFTs, is that they can be run online, i.e., as a system that runs continuously and detects digits as they are spoken. Explain how this can be done. Implement it.

1 Time management

This lab is intended to be very short since you are supposed to be studying for your midterm. One or two hours should be sufficient.