# Voice recognition

Alejandro Ribeiro

February 24, 2020

The goal of this lab is to use your accumulated knowledge of signal and information processing to design a system for the recognition of a spoken digit.

Figure 1 shows four different realizations of the DFT of the signal recorded when I spoke the word "one." These four DFTs are different to each other because there are variations in the sounds that I produce, but they also have discernible patterns. E.g. in all four DFTs you can see two well defined frequency spikes close to frequencies 0.5kHz and 0.7kHz. That these patterns are specific to the word "one" can be verified by the four different realizations of the DFT of the signal recorded when I spoke the word "two" that are shown in Figure 2. The two characteristic spikes of the DFTs in Figure 1 are absent from this second set of DFTs, which, instead, seem to all have a large frequency component a little below 0.4kHz. Another feature that arises upon comparison is that the spectra associated with the word "two" have their energy more evenly spread out than the energy of the word "one." Regardless of the specific features, the general level conclusion is that the four DFTs in Figure 1 are more like each other than they are like the DFTs in Figure 2, which are also more like each other than they are to the DFTs in Figure 1.

We could spend days talking about the physical meaning of these differences. Large frequency components are generally associated with vowels that produce high energy sounds at definite frequencies. Consonants generate sounds with less power because the vocal cords are not involved in their generation. Consonant sounds also tend to be more spread out in frequency because they are not associated with oscillating tones. The differences we see between the DFTs of the spoken words "one" and "two" are because their vowel sounds are different—so the spikes are in different locations—and the consonant sound in "two" is longer than the consonant sound in "one"—so the energy in "two" is more spread out.

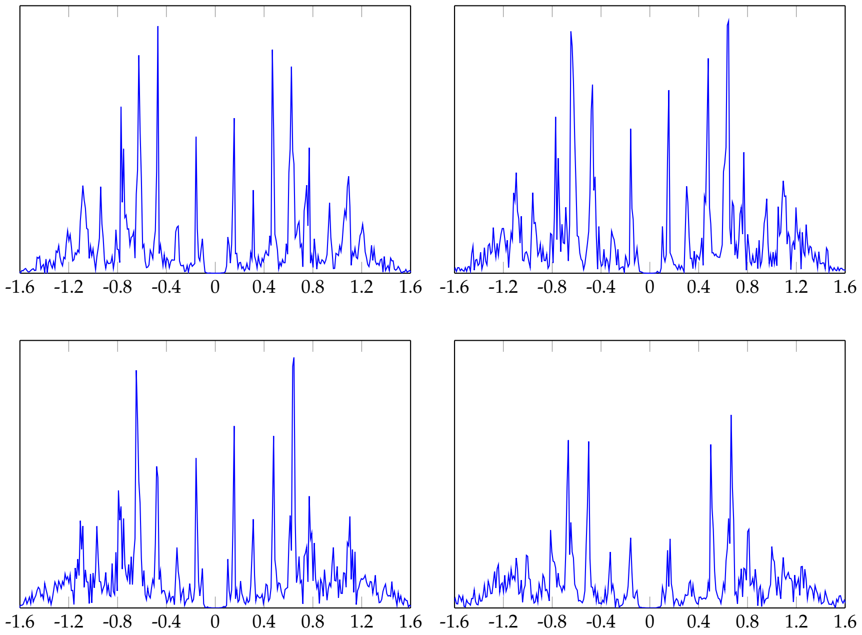

**Figure 1.** Four different observations of the Fourier transform of the signal recorded when speaking the word "one." These transforms are different from each other but they are more like each other than they are to the ones in Figure 2. (Frequency axis in kHz, sampling frequency set to $f_s = 8$ kHz, signal duration $T = 2$s)

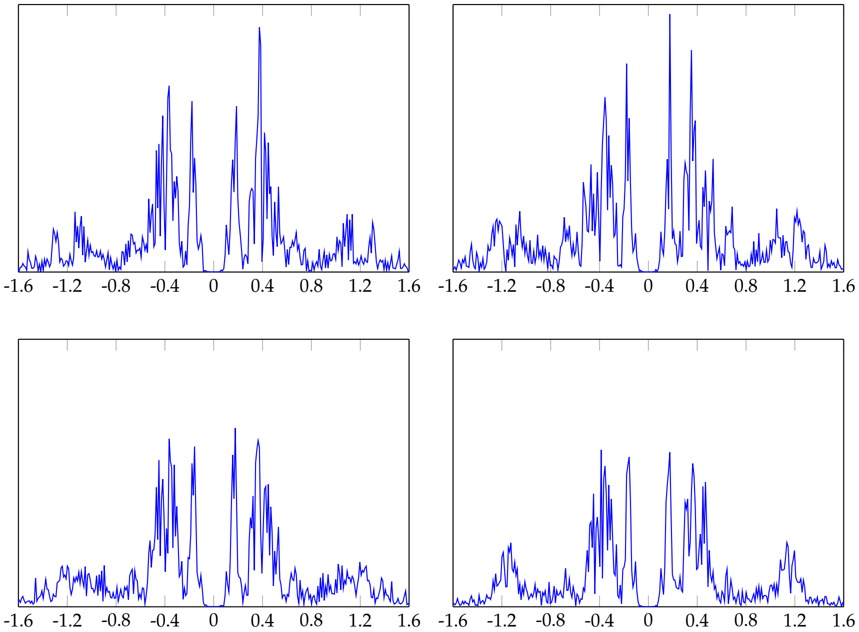

**Figure 2.** Four different observations of the Fourier transform of the signal recorded when speaking the word "two." These transforms are different from each other but they are more like each other than they are to the ones in Figure 1.

However, our interest today is not on analyzing these differences but on using them to detect a spoken digit. To that end, start by recording $N$ waveforms $y_i$ for the spoken word "one" and $K$ waveforms $z_i$ for the spoken word "two." The respective DFTs are denoted as $Y_i = \mathcal{F}(y_i)$ and $Z_i = \mathcal{F}(z_i)$. The sets of all DFTs $\mathcal{Y} := \{Y_i\}_{i=1,\ldots,N}$ and $\mathcal{Z} := \{Z_i\}_{i=1,\ldots,N}$ are called training sets. We assume that the signals in the training sets have been normalized to have unit energy.

**1 Acquire and process training sets.** Acquire and store $N = 10$ recordings for each of the two digits "one" and "two." Compute the respective DFTs and normalize them so that they have unit energy ($\|Y_i\|^2 = \|Z_i\|^2 = 1$). You can use the provided `record_sound()` function. If you do, please remember to check `help record_sound` before using it.

Having acquired and processed the training sets $\mathcal{Y}$ and $\mathcal{Z}$, we acquire a new signal $x$ that contains the utterance of either the word "one" or the word "two." We want to (correctly) identify which of these words was spoken. To do so, we compare the magnitude of the DFT $X = \mathcal{F}(x)$—which we assume has been normalized to have unit energy—with the magnitudes of the DFTs $Y_i$ and $Z_i$ that were stored in the training sets. There are different choices to make this comparison. We will try two of them in this lab. It will probably save you a lot of time to record all the digits once and store them in a `.mat` file for future use (i.e., so you don't have to re-record them every time you run your code). This can be done by either saving the workspace or using the function `save()`. In any case, you can ask your TAs for help on storing the recorded voices. It would also help, if the script for recording voices and storing them is in a separate file (although, it would be useful that *every* script for every exercise is in a separate file).

**2 Comparison with average spectrum.** For each of the training sets define the average spectral magnitudes

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

Further define the inner product $p(X,Y)$ between the spectra of any two signals $X$ and $Y$ as the inner product between their absolute values,

$$p(X,Y) = |X|^T|Y| = \sum_k |X(k)| \cdot |Y(k)|. \tag{2}$$

Compare the inner product $p(X,\bar{Y})$ between the unknown spectrum $X$ and the average spectrum $\bar{Y}$ with the inner product $p(X,\bar{Z})$ between the unknown spectrum $X$ and the average spectrum $\bar{Z}$. Assign the digit of the spectrum with the largest inner product. Estimate your classification accuracy. Explain why we are using the absolute values of the spectra.

The classification accuracy can be estimated by recording your voice several times saying either "one" or "two" and regarding each of these recordings as $x$ (remember to keep track of what number you actually said). Then, we consider a success every time the comparison in 2 yields the right digit. The ratio of the number of successes to total number of attempts is a measure of classification accuracy. Try recording and testing an additional 10 utterances of "one" and "two" (typically known as the *test set*).

**3 Nearest neighbor comparison.** Compute the inner product $p(X,Y_i)$ between the unknown spectrum $X$ and each of the spectra $Y_i$ associated with the word “one.” Do the same for the inner product $p(X, Z_i)$ between the unknown spectrum $X$ and each of the spectra $Z_i$ associated with the word “two.” Assign the digit of the spectrum with the largest inner product. Estimate your classification accuracy.

**4 Larger number of digits.** Try developing a system to identify all 10 digits. We will give you 2 extra points for the effort and 3 more extra points if you succeed.

## Time management

This lab marks a shift with respect to previous labs. You have acquired, or at least I am assuming that you have acquired, all the fundamental concepts on the spectral analysis of one dimensional signals. This lab is just a test of the application of the concepts you have learnt. The pieces are not lengthy. You should be able to solve Part 1 in less than 1 hour and spend two or three hours in each of the other two parts. The extra time you can use it to start preparing for the midterm.