Exercise #1
Manual speech transcription, analysis & synthesis, basic signal representations

İlknur Baş
151226814
20.03.2023

Task 1: Manual segmentation and labeling of words and phones

Q1.1:

In my opinion, segmenting phones was considerably more challenging than segmenting words. In this speech sample that is analyzed, the speaker's pronunciation was clear, and he spoke at a moderate pace, allowing sufficient pauses between words which ease the segmentation process. Yet, certain word combinations (that follow each other) such "hear about" are difficult to segment since there is very little pause between them, and when spoken, they are perceived as a single word rather two separate words. But still segmenting phones was significantly more challenging. For instance, I had difficulty identifying a clear boundary between the first two phones in words such as "public", [P] and [AH], or in "Terran", [T] and [EH]. These two phones seemed as combined. I would not encounter much difficulty in segmentation of word since words consist of many phones.

Q1.2:

I definitely had hard time defining segments for some phones. As I stated above, certain phones in words were particularly difficult to segment, such as the first two phones in words "public", [P] and [AH], or in "Terran", [T] and [EH]. I believe the reason of this difficulty is that there is no pause or silence between phones (as it could be in between words). Also, it is very hard to detect when a certain phone ends and the other one starts because, to me, the first phone is a consonant and the second one is vowel. While pronouncing the consonant, towards the end we would get a vowel sound and it makes hard to clearly segment them. Lastly, as it explained in Hint#3 of the assignment, plosives consist of a closure, and it is hard to detect the beginning of the closure. This could also be a reason for not segmenting clearly.

Task 2: Duration estimation

Q2.1:

Table 1: Manual acoustical measurements of speech segments.

Phone	Duration (ms)	F1 (Hz)	F2 (Hz)	F0 (Hz)
[DH]	314.278			
[AH]	45			
[T]	101.499			
[EH]	70.233			
[R]	56			
[AH]	49			
[N]	103.964			

Figure 1: Measure of durations of some phone segments in *ms* (Note: The rest of the table is completed in further sections.)

During the lectures, we have learnt that the phone durations typically fall between 20-80 ms. Hence, the phone durations are not uniform, and they can vary between those numbers, based on the type of phone and its neighboring phone as well. Based on my measurements, more than half of them align with this statement. It is also worth noting that the durations differ between different phones. Another thing that I have notice was the durations for phone [AH] is very similar, but not same. I think, the duration of a phone can be affected by how the speaker pronounced (longer, shorter) that particular phone during speech. The [DH] phone seemed as an outlier at first, but the pronunciation could also be the reason it got higher duration. Also, I have noticed that there is silence at the beginning of this phone, so this silence is also included in this duration. According to the speech sample, there are 12 phones in the first 1.040 sec of the sample. In other words, the average duration for each phone would be 83.3 ms, which again close to range given in the lectures. Hence, we could say that there would be around 12 phones per second in the sample speech.

Task 3: Formant estimation

Q3.1:

Table 1: Manual acoustical measurements of speech segments.

Phone	Duration (ms)	F1 (Hz)	F2 (Hz)	F0 (Hz)
[DH]	314.278			
[AH]	45	454	1600	
[T]	101.499			
[EH]	70.233	560	1700	
[R]	56	500	1200	
[AH]	49	560	1550	
[N]	103.964	300	1600	

Figure 2: Values of first two formants F1 and F2 (Note: The rest of the table is completed in further sections.)

The estimated frequencies for F1 for phones given in Figure 2 falls between 300-500 Hz, whereas for F2 it ranges between 1200-1700 Hz. When it is compared with the table given in the Appendix of the assignment, it has seen that the frequencies are highly similar. While comparing the estimations, corresponding row for men is checked since the speaker is assumed to be men. For phone [AH], the difference between the real and estimated F1 and F2 has found around 200 Hz. For phone [EH], [AH] has found around 100 Hz, 200 Hz respectively. In general, it could be said that the estimated and the real values of F1 and F2 were very similar.

Task 4: Fundamental frequency estimation

Q4.1:

Table 1: Manual acoustical measurements of speech segments.

Phone	Duration (ms)	F1 (Hz)	F2 (Hz)	F0 (Hz)
[DH]	314.278			
[AH]	45	454	1600	160
[T]	101.499			
[EH]	70.233	560	1700	230
[R]	56	500	1200	247
[AH]	49	560	1550	240
[N]	103.964	300	1600	210

Figure 3: Values of formant F0

As it can be seen from above, F0 values are very close to each other, ranging from 160 to 247 Hz. We could say that during that part of speech, there is not much difference in F0 values. When I made observations for the whole speech sample, I have encountered with the same interpretation, there was no drastical chance in F0 values throughout the speech. I believe, the reason for seeing not many changes is that the speaker speaks at a certain tone. We see some increment in F0 values when the speaker stresses a certain phone (which is located at the start of a word). Hence, we could say that tonation could affect F0 values. Other than that, speaker's consistent tone is the reason for lack of significant changes.

I cannot say F0 values perfectly align with phone and/or word segments, however, as I explained above it can give insights about where the speaker's tone has changed and so on. According to this tonation changes, we can make assumptions about when the certain phone and/or word has started. For example, in the sample speech, the speaker stresses the phones in the middle of "Terran" word and "wanted", for that reason, increment in F0 values around the

middle can be seen in Praat as well. Also, when there is a silence (which indicates the starting a new word most of the time, not always), the F0 value is not shown. Hence, F0 could be an indication of a starting a new segment.

Task 5: Manual concatenative synthesis

Segmentation that is used in this task is as follows:

- [EH] [R] [IH] [N] [Z] → ERIN'S
- [R] [EH] [D] →RED
- [M] [UW] [N] →MOON

Q5.1:

The quality of the synthesized speech did not seem good. When the extracted wav file is listened, it seemed like there are some distortions, robotic sounds in the synthesized speech. These distortions are a bit to the point where it is hard for a person to understand what words are said during the speech. These issues could be related to me being able to perform segmentation of phones with less accuracy.

Q5.2:

The first way to make this synthesized speech more natural sound is to correctly segment phones and/or words. The silences between words or at the beginning of the sentence should be consider during this segmentation. In the given speech, some phones and words as well were very close to each other which makes them hard to segment. Hence, more than one speech sample could be used in order to avoid that issue. Even though the given speech is very clear, the quality of it is good and the phones/words clearly are pronounced, we still make sure that these are valid for our samples (if we use more than one) while creating synthesized speech.

Task 6 (MATLAB):

Q6.1:

According to following source [1], to determine the sampling rate for speech, we have to consider the formants that is contained in speech. Frequency range for formants fall between 300 Hz to 3500 Hz. According to Nyquist frequency, in order to capture information accurately, sampling rate should be at least 2 times highest frequency or higher than that number. In other words, sampling rate should be around 7000-8000 Hz. Previously, sampling

rate of 8000 Hz was used in digital speech decoders but was not sufficient enough to capture some information, such as some constants. That is why a sampling rate of 16 kHz is preferred to sample wider range of frequencies and capture much (and quality) information from the speech.

Task 7: Windowing, waveform and spectrum plotting

Q7.1:

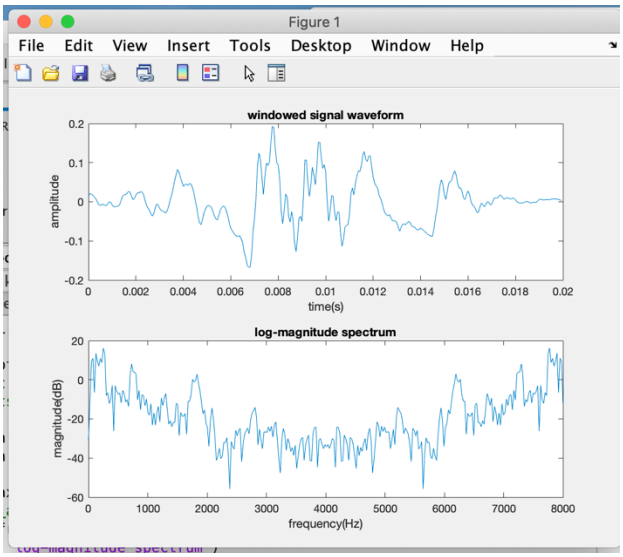

Figure 4: The windowed signal waveform (top panel) and its corresponding untrimmed log-magnitude spectrum (bottom panel) for phone "IH"

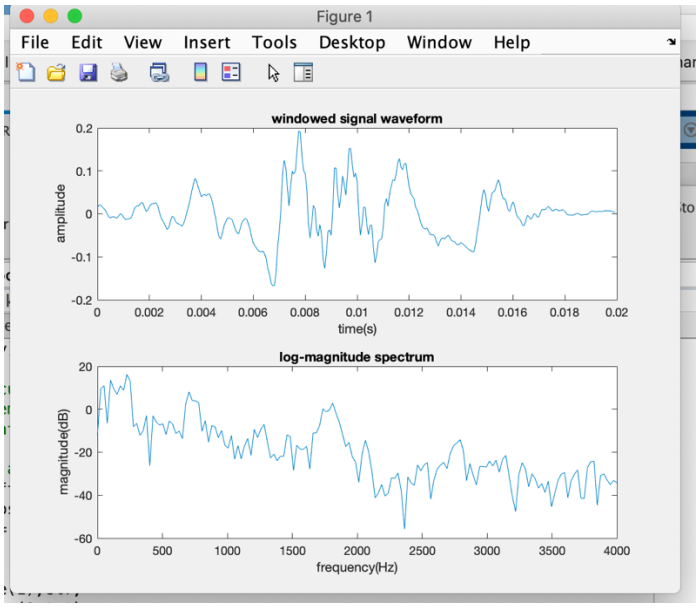

Figure 5: The windowed signal waveform (top panel) and its corresponding trimmed log-magnitude spectrum (bottom panel) for phone “IH”

In the question, it is stated that the dynamic range of high-quality speech is around 60 dB. This means that the difference between the lowest and highest parts of the recording is around 60 dB. When we checked the log-magnitude spectrum both in Figure 4&5 (they are, in fact, the same), we can see that it is around 18-(-56) =74 db. The result a bit derives from 60 dB, but the result could be related to segmentation of “IH” phone.

References:

[1] “Waveform.” *Aalto University Wiki*, https://wiki.aalto.fi/display/ITSP/Waveform.