Exercise #4:
Statistical parametric phone synthesis GMMs and LP

İlknur Baş
151226814
26.04.2023

Task 1: Implement excitation signal generation

Q1.1

According to the attached *excitation_impulse.wav* and *excitation_glottal.wav* sounds, it is hard to identify vowel identities (by listening). As the phones to be synthesized is known, in those *.wav* files, we can identify the vowel placements however identify phones accurately can be challenging. In short, the excitation signal could give us an idea whether the phones are voiced or unvoiced sounds. In fact, while creating the excitation signal, we apply certain rules for certain cases (white noise when phone is unvoiced etc.) For F0, we learned that each phone has different fundamental frequency, which could be beneficial in terms of identifying vowels and their identities. However, as I said before we should take into account for other factors in order identify vowel identities more accurately. For example, in previous exercises we have learned that vocal track gives information about the spectral envelope of the signal which also gives information about formants. Formants are the resonant of the vocal track and they are useful for identifying differences in vowel identities. Also, coarticulation has an effect on vowel identities, for that reason we can say that person's lips, shape of mouth etc. has an effect on vowel identity as well. In short, these factors also should take into account, not only excitation signal and F0 alone.

Q1.2

In the current type of generation of excitation signal, we are assuming that each phone is independent from each other. Meaning that, we are generating the excitation signal for each phone separately. However, we should have taken into account of coarticulation where each phone is affected by the previous and following phones in continuous speech. Not having this could also affect the quality of resulting synthesized speech signal at the end in a negative way and it may not sound as natural. For that reason, we can say that the phone-wise excitation signal generation is suboptimal. A way to improve this generation, we could try to use biphones or triphones, which can capture the coarticulation between the adjacent phones.

Q1.3

A strict division in voiced and unvoiced excitation signals is not an ideal solution as in some cases, the excitation can be mixture of those two. As I explain above, coarticulation is an important term when it comes to natural speech. A phone is affected by its previous and following phones. We can especially observe this behavior (mixture of voiced and unvoiced excitation signals) at the boundaries where there is transition from one phone to another. Assume a phoneme is unvoiced and the previous phone is voiced. This voiced phone can make unvoiced phone partially voiced. Duration of the voiced phone is one of the factors why this case is happening. The voiced phone can cause the vocal folds to be open and let's say it is pronounced very quickly. And while pronouncing the unvoiced phone, the vocal folds may still be open a bit, resulting of making unvoiced phone partially voiced phone as it allows passing some air through it. Also, personal speech patterns (how an individual pronounces certain phones, speaks etc.) could be another reason that we observe this behavior.

Task 2:

Note: Vocal track parameters do not depend on excitation signal but there is randomness while creating MFCC vector. That is why I have attached separate plot for each case.

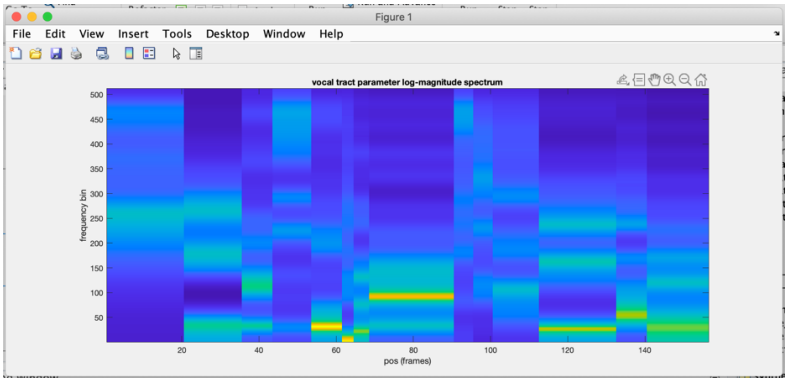

Figure 1: Vocal track parameters while synthesizing speech in which impulse train has used as excitation

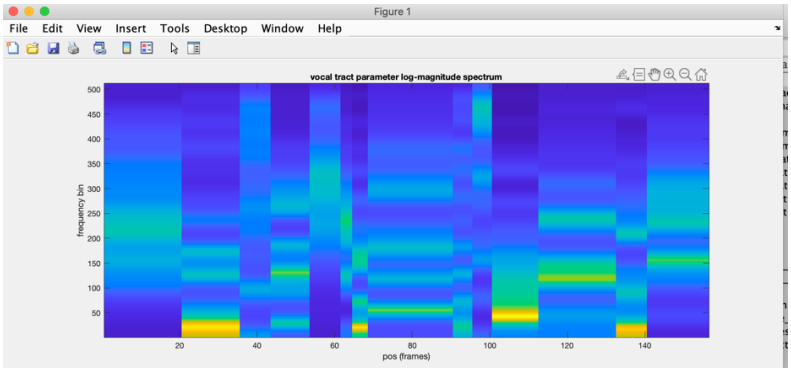

Figure 2: Vocal track parameters while synthesizing speech in which glottis has used as excitation

Q2.1

In the case where we use fixed mean values of the model, we would assign same MFCC vector to each instance of the same phone in the sequence of the phones to be synthesized. However, different phones might have different properties depending on individuals, context of speech etc. Hence, in that case, the vocal track parameters of the same phone would not vary and as a result we wouldn't be capturing variability. On the other hand, by stochastically sampling vocal tract values from GMM, we would introduce variability to the system. Specifically, we have used *"mvnrd"* in the code which allows randomly sampling of MFCC vector for each instance of the same phone. That can be counted as a potential advantage.

Since, stochastically sampling introduces some randomness (during the implementation, one random component is selected in the creation of MFCC vector), in each run we might produce different results. That can be counted as a potential disadvantage as we wouldn't have stabile output/synthesis. Also, assume we don't have the variability in the data that we used while training the GMM, in that case even if we use stochastically sampling, we still might not capture the variability. It is not directly a disadvantage but another point to consider.

Q2.2

We could consider the effect of neighboring phones when it comes to vocal track modeling. For that reason, we could try to use biphones or triphones, which can capture and model the coarticulation between the adjacent phones. In the current implementation we are using each phone independently to model vocal track. Another approach could be observing the measurements or movements of speech articulators, such as lips, tongue etc. During the lectures we have learned about synthesis and one of the examples that is given was how we can estimate the position of the tongue from images and after that it can be mapped into a model. Hence, those speech articulators have an effect on coarticulation, such as the position of the tongue in one phone can affect the position of the tongue in the following one. For that reason, we could model speech articulators in order to capture coarticulation. Another way to model coarticulation is that taking the duration of each phone into account. As I explain in Task 1's questions, the duration of a phone can affect the following phone, results in giving information about the degree of coarticulation. Another approach is that in the Appendix B of this assignment, Gaussian Mixture Model-Hidden-Markov Models (GMM-HMMs) is mentioned briefly. It is stated that HMMs capture the transition between states meaning that

transitions from sounds to other, however, in the current implementation we have used GMMs. Hence using GMM-HMMs models together could be a good way to model coarticulation in the system. Lastly, using neural networks that shows the temporal dependencies such as RNN etc., can be another way to take coarticulation into account.

Task 4:

Q4.1

The synthesized sound is more natural, human-like and less robotic when the glottal excitation is used compared to when the impulse train excitation is used. The sounds are somehow understandable, but in the synthesized speech, some phones sound like they are mispronounced. For instance, in *"test_synthesis_glottial.wav",* phone 'F' is a bit sounded like 'P'. We get to see same pattern in *"test_synthesis_impulse.wav"*. I believe the GMMs could be a reason why this is happening. Because, we have used GMM model to generate acoustic features, MFCC vector, and it might not capture variations in those phones very well, there might be confusions. But for some phones, there might be no such confusions as we perceive them as pronounced well. Also, 'F' is an unvoiced sound, which does not perform the vibration of vocal cords. For that reason, it might be a hard to perceive unvoiced sounds compared to voiced. Hence, the quality of the synthesized sound definitely depends on specific phones/sounds in the sequence of phones to be synthesized.

Q4.2

Theoretically, using glottal excitation will result in more natural speech compared to impulse train as it models how the actual vocal folds generate sounds. In this exercise as well, we make use of pre-recorded glottal excitation signal. Also, note that the glottal waveform is extended in time to support much higher F0s as it is written in the assignment. On the other hand, when we defined impulse train excitation as a signal that contains zeros and ones, it is normal that it may have less natural/more robotic sound as it does not capture the complex structure of the vocal folds. When I listened the synthesized speeches, the one with glottal excitation signal sounds more natural and it has a human-like sound. On the other, hand the one with impulse train has sounded like more robotic. But in general, both synthesized speeches are not that good and does not sound as natural as human, which is understandable since we used GMMs.

Q4.3

In the current system, GMMs are trained using speech data from multiple speakers. One way to produce more natural speech could be using large speech data from one speaker. During the lecture about unit selection synthesis, for that type of synthesis, we have learned that high-quality speech from a single speaker could avoid the unnecessary speech variability. And it also provides consistency when it comes to characteristic of speech for that speaker. Hence, we could use that kind of data in order to synthesize more naturalistic speech. Also, the current system uses GMM to model acoustic features of speech sounds. We could use for example neural networks in order to improve accuracy and as a result to produce more natural speech. Lastly, as I mentioned in above sections, we could use biphones or triphones instead of a single phone. By using those we could avoid transition problems that occurs in consecutive phones and as a result, the resulting synthesized speech sounds more natural.

Q4.4

By using higher number of GMM components, we can achieve to capture more variability in the data, however we should be careful about the number as it may introduce overfitting problem. We have used LP order to decide on vocal tract parameters. A higher number of LP order can provide more accurate synthesized signal as we would have more coefficients to represent vocal tract parameters. But similarly, we should be careful about using much higher LP order. In general, using hyperparameters that has optimal number lead to higher quality sound in synthesized speech, meaning that speech sound with more natural, less artifacts.

One way to automatically find the optimal hyperparameters is that adjusting possible parameters iteratively and save the synthesized speech samples. After that, as we learn in the lectures, we could use for example Mean Opinion Score (MOS) metrics where human listeners evaluate the synthesized speech samples. But in this case, humans are involved into the process. We could also make use of machine learning such that a model could be trained to predict the MOS scores. In the lectures, we have seen a prediction-based metric, specifically “Neural models predicting MOS scores”. Even though it is a good approach, we need large amount of training data, speech signal and MOS score pairs to train the model. Or similarly, we could use speech signal and the optimal hyperparameters pairs to train the model as well.

Specifications of synthesized custom utterance

Utterance (with male voice): Erin's red moon

phones_synthesize = {'EH', 'R', 'IH', 'N', 'Z', 'sil', 'R', 'EH', 'D', 'sil', 'M', 'UW', 'N'}
f0 = [127 100 120 120 120 100 100 127 100 100 100 230 120]
voiced = [1 1 1 1 1 0 1 1 1 0 1 1 1]
phone_durations = [100 70 70 50 90 70 70 90 60 90 70 90 80]

While deciding whether a phone is voiced or unvoiced, The International Phonetic Alphabet chart is used [1]. For F0 values, the attached chart in the Appendix of Exercise 2 assignment is used.

As a side note, in the assignment it is written that name the synthesized custom speech as *'custom_synthesis.wav'*. However, I wasn't so sure whether it is asked to used glottis or impulse train as excitation signal, so I have attached both. Hence, the synthesized custom speeches are named as *'custom_synthesis_impulse.wav'* and *'custom_synthesis_glottis.wav'* inside the attached zip file.
Additionally, in both cases, the speech quality was bad, not very understandable.

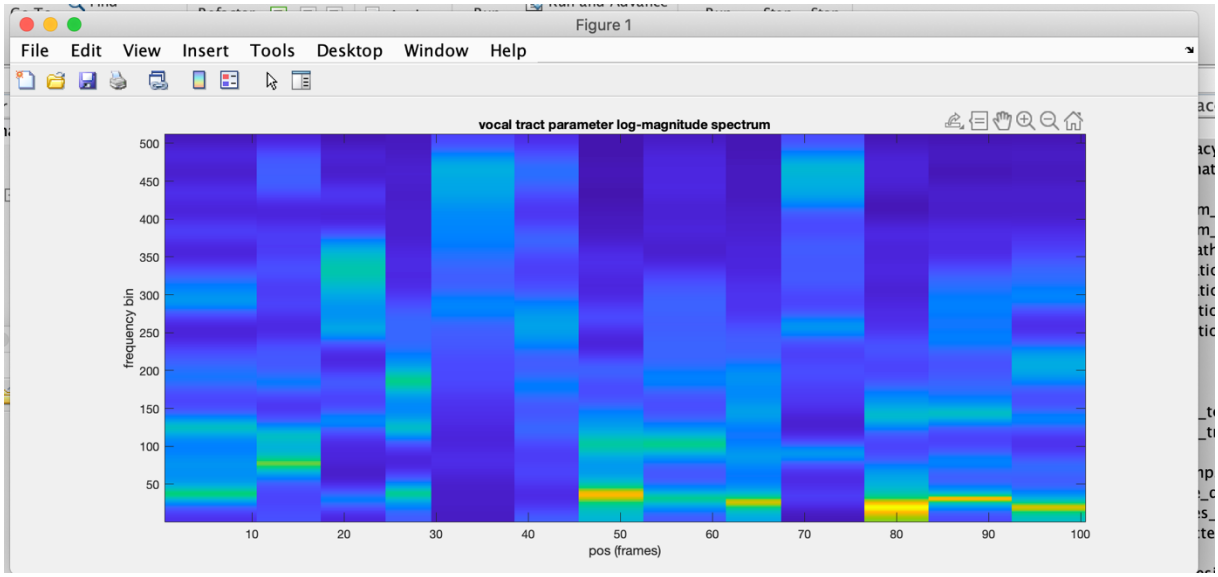

Figure 3: Vocal track parameters while synthesizing speech in which impulse train has used as excitation

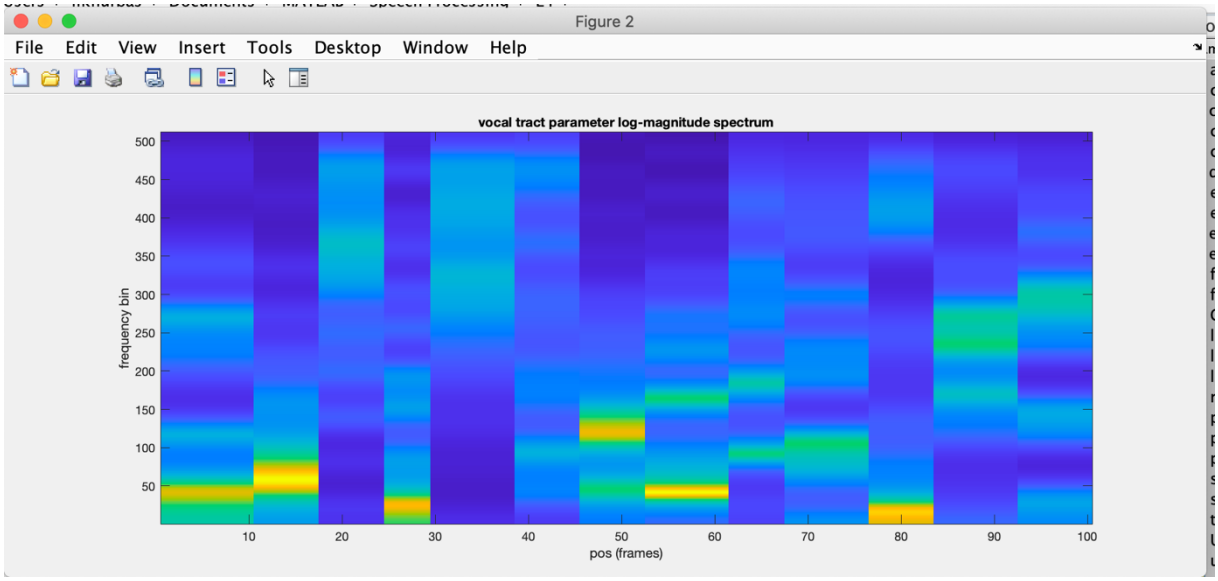

Figure 4: Vocal track parameters while synthesizing speech in which glottis has used as excitation

REFERENCES

[1] *File: IPA Chart 2020.SVG* (no date) *Wikimedia Commons*. Available at: https://upload.wikimedia.org/wikipedia/commons/8/8f/IPA_chart_2020.svg (Accessed: April 30, 2023).