Analysis and Synthesis of Emotional Voice Using Fundamental and Formant Frequecies

Harry Chung

MUMT 307
Music & Audio Computing II

Class Project

1 Motivation and Objective

Human beings group together and bond with each as a natural instinct, and it is widely accepted that the ability to express emotion plays a key role in this social behavior. [5] The challenge of understanding how one portrays emotion is that the parameters are complexly intertwined in both realms of nature and nurture. For instance, in his book “The Expression of Emotions in Man and Animals” by Charles Darwin, he claimed that the expression of emotion involves media, such as facial expression, and behavioral and physical responses. Furthermore, humans often utilize “white” lies as a form of social politeness, which is naturally deceptive in communication. Study by Giles, Rothermich, and Pell showed cross cultural complexity of emotion perception[1], and thus further proves that deciphering the speaker’s emotion via parameters stemmed from human consciousness is too complex.

However, what people have in common is the involuntary physical responses, such as dry throat, increase heart rate, and increased muscle tension, induced from a stressful environment. All of these physiological responses influence the muscular structural of the vocal cord, which indicates that there has to be acoustic signatures that can be used to distinguish types of emotion.

The goal of this project is to study acoustic parameters, especially the ones stemmed from involuntary physical responses, that are influential in portraying an angry voice, and attempt a simple and crude alteration to explore whether it’s possible to morph a neutral voice into an angry voice. In section 2, we will review the mechanics of the voice production and parameters involved in portraying emotional voice. In section 3, we will discuss a simple and crude method to apply a selection from those parameters and the result.

2 Acoustics Parameters

2.1 Voice Production

The voice production process is an intricate sequence of physiological activities. However, it can be reduced down to a sequence of variations that are based on a simple resonance model, which resembles a modal synthesis. As the air travels from the lungs through the glottis, the air flow gets modulated at various frequencies depending on the velocity of the air. This is the source of the sound that the body produces, and one could interpret this as a continuous impulse response. This sound wave then gets resonated by a complex, and profoundly personal, topologies of the vocal tract, the oral cavity, and the nasal cavity. [4]The final resonation from the system is the basis of all the vowels, and each specific vowels then gets further specified by modulations of the lips and the oral cavity.

2.2 Stress

The human voice production system is surround by a complex network of nerves. The physiological mechanism is integrated with the central nervous system (CNS), which controls the physical movements, and autonomic nervous system (ANS), which controls the functions of the internal organs, such as blood pressure, heart rate, and swallowing. It’s been well studied that a person under stress experiences various responses from both the CNS and ANS, such as increased muscle tension, increased heart and breathing rates, and dry mouth and throat. [7] Since the shape and texture of the resonator undergo significant changes due to the aforementioned involuntary physical reaction, it is a plausible argument that any type of stress will induce a change in vocal characteristics, which we then could detect through signal processing.

2.3 Parameters

Previous studies have shown that acoustic cues, such as fundamental frequencies, formants, formant bandwidth, intensities, spectral envelopes, along with many other parameters form distinct patterns depending on the speaker's emotion. [4, 6, 7] Interestingly, Mongia found that the vowel and nasal phoneme classes are much more effective than fricative and plosive classes at portraying the emotion. [4] Considering that the goal of this project is to study the effect of emotion in voice, I decided to focus on the fundamental frequencies and formants, as those are the main parameters for the vowel and nasal phoneme classes.

3 Project Design

As mentioned previously, the second goal of the project is to explore if the selected acoustic parameters, the fundamental frequency (F_0) and the formant frequencies (F_1, F_2, F_3, etc), can be inversely applied to a neutral voice to an emotional voice. I chose an angry voice, and I will explain this decision later on, and how it plays a role in this project.

3.1 Fundamental Frequency Estimation

Fundamental frequency estimation has been well studied. However, considering that the project is just a proof of concept, I decided to go with the simplest method. For this application, I utilized a built-in Matlab function called *pitch*. In short, this estimates the fundamental frequency in a given window of time, using a technique called "Pitch Estimation Filter". [3] In order to alter the fundamental frequency for each analysis window, I matched the function's window and overlap length specifications to those of the STFT process I used for the final synthesis of the emotional voice.

3.2 Formant Estimation

For this segment, I utilized a formant estimation method based on linear predictive coding as described in [2]. This also estimates formants in a selected segment of time frame. This method, however, requires a manual selection for each and entire vowel segment. Running *spectrogram* in Matlab gives the following time-frequency representation:

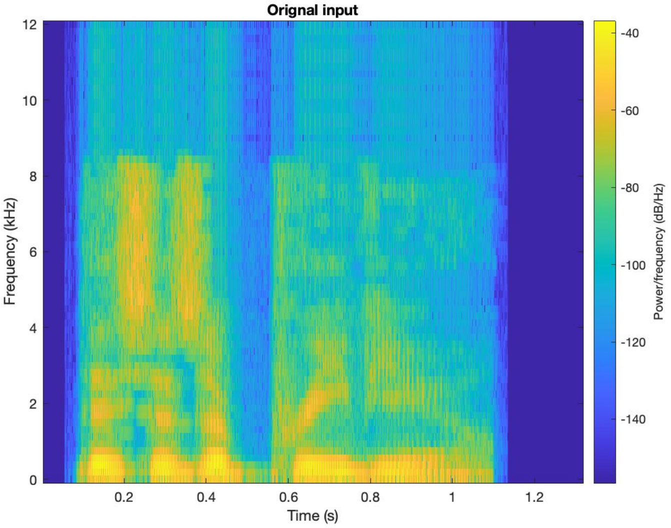


Figure 1: Spectrogram of the input sound

There are five vowels in the example "This is a preview", and those are visible in the domains of [0.12, 0.19], [0.26, 0.32], [0.38, 0.45], [0.62, 0.74], [0.8, 0.95], respectively. For each formant, after applying a hamming and a pre-emphasis filter with coefficient of 0.95, I followed the process described in the referenced Matlab example.

3.3 Anger

During an interview Dr. Marc Pell at McGill's linguistics department, he stated that in a carefully controlled setting in which only the voice is provided, anger and similar strongly negative emotions are the easiest to recognize. In order to maximize the recognition of emotion in a crudely synthesized voice, I decided to mimic an angry voice. According to the study by Sondhi [6], an angry voice 30% higher fundamental frequency while the first two formant frequencies were lower by 5% and the third and fourth formants were remain practically unchanged. This shift was similarly shown in a study by Van Puyvelde [7], in which the researcher found the fundamental frequencies to be higher by 50-100%. For this study, I've adopted the results by Sondhi.

3.4 Matlab Code

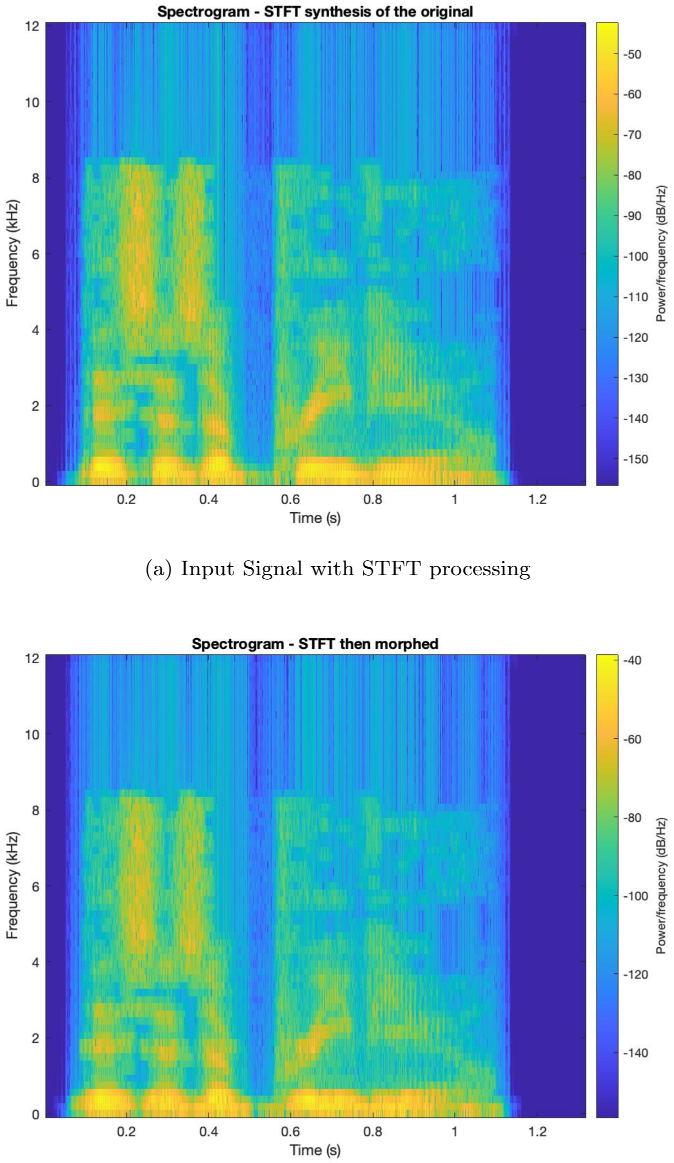
The following a pseudocode of the code used for this project. The STFT portion is based on a sample provided by Prof. Gary Scavone in MUMT307.

Algorithm 1: Angry voice synthesis

```
Result: Synthesized sound
Source audio import;
Set up variables for STFT;
Run Spectrogram to estimate the vowel locations;
for first vowel : last vowel do
    Estimate Formants (input, sampling rate, start of the vowel, end of the vowel);
    Find the windows where the formant is located;
end
Run Pitch to find the fundamental frequency for each specified window size and locations;
for first window : last window do
    Find the index (location) of the $F_0$;
    Find the index of the new $F_0$, 30% increased;
    Assign the power of the old $F_0$ to the new $F_0$;
    Attenuate the power of the old $F_0$;
    if This window has the first vowel then
        Open the formant morphing (decrease by 5%) process;
        Find the index (location) of the $F_{1,2}$;
        Find the index of the new $F_{1,2}$;
        Assign the power of the old $F_{1,2}$ to the new $F_{1,2}$;
        Attenuate the power of the old $F_{1,2}$;
        Turn off the formant morphing;
    else
        Repeat for the rest of the vowels;
    end
end
Make the conjugate symmetric;
OLA method to synthesize the final sound;
```

4 Result

As shown in Figure 2, the changes in time-frequency representation isn't very noticeable. One could noticed a small changed in the shapes of the darkest yellow blobs on the low frequency region. These regions are mostly composed of the fundamental frequency, and thus reflect the 30% change. In terms of sound, the synthesized sound had too much distortion to notice the change in emotion, while the changes in the fundamental frequencies were noticeable.



(a) Input Signal with STFT processing

(b) Angry Voice Morphing

Figure 2: Voice Morphing

One of the potential reasons for such a result is that I did not apply any changes in amplitudes of the fundamental frequency or formant frequencies for each vowel. As mentioned previously, amplitude plays a role in emotion infliction, but the method used in the project didn't reflect that. A simple reallocation of the amplitude possibly caused the distortion and didn't realistically reflect the change in timbre.

5 Challenges and Rewards

The most challenging part of this project was narrowing down parameters that are played in reflection of human emotion in voice, and figuring out appropriate techniques to extract and reapply those parameters. With enough time to fully digest and customize the techniques and proper code designing, it should be possible to create less distorted result. The code used in the project was probably too simple to reflect the subtle changes in timbre, and thus impossible to conclude if a morphing from neutral voice to an angry voice is possible.

6 Potential Usage

The initial motivation for this topic was a therapy for patients with schizophrenia using an avatar. My intention was to explore ways to morph the input voice of a therapist, for example, into the voice of a hallucination that a patient encounters. This hallucination typically has a voice with timbre that resembles strongly negative emotion, such as anger. I intend to continue this field of study in multiple directions, one of which being a physical modelling of the vocal system and the other being digital signal analysis and processing.

References

[1] R. Giles, K. Rothermich, and M. Pell. Differences in the evaluation of prosocial lies: A cross-cultural study of canadian, chinese, and german adults. 07 2019.

[2] Matlab. Formant estimation with lpc coefficients. https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html.

[3] Matlab. Pitch. https://www.mathworks.com/help/audio/ref/pitch.html.

[4] P. Mongia and R. Sharma. Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. *Journal of Computer Networks and Communications*, 2014, 11 2014.

[5] S. Planalp, J. Fitness, and B. Fehr. *Emotion in Theories of Close Relationships*, page 369–384. Cambridge Handbooks in Psychology. Cambridge University Press, 2006.

[6] S. Sondhi, M. Khan, R. Vijay, A. Salhan, and S. Chouhan. Acoustic analysis of speech under stress. *International Journal of Bioinformatics Research and Applications*, 11:417–432, 01 2015.

[7] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn. Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9:1994, 2018. https://www.frontiersin.org/article/10.3389/fpsyg.2018.01994.