INVESTIGATION ON SPEAKING AND SINGING TIMBRE DIFFERENCE

Haiqin Yin & Shuizening Li

University of Rochester

ABSTRACT

This project explores the sonic difference between spoken and singing voice in various aspects such as power, harmonic content, periodicity, pitch contour and consonant length, etc. by using functions in Matlab to visualize the sonic difference.

1. INTRODUCTION

The goal of this project is to investigate how the timbre of singing and speaking voices differ from each other. Timbre is the quality give to a sound by its overtones such as the resonance by which the ear recognizes and identifies a voiced speech sound and the quality of tone distinctive of a particular singing voice or musical instrument (Merriam-Webster). The timbre of one's voice, along with other properties, are what makes one's voice disguisable from others'.

Singing and speaking voices are closely related to each other as one's voice in different situations have very similar movements of articulation. However, they are also distinct from each other from features of pitch movement(contour), timbre, phoneme duration, rhythm, power, etc. We planned to analyze the production, acoustic properties and perception of speaking and singing voices respectively and compare the features of each voice from the result retrieved from research and experiments.

At the end of the project, we should be getting a detailed conclusion of the difference of the singing and speaking voices and gives a brief report of the application of utilizing these differences in audio technologies.

2. CONCEPT

The following are the five main properties of speaking and singing voices and the methods we used to analyze of our research:

2.1. CONSONANT LENGTH

Consonant is "a speech sound produced by occluding with or without releasing (p, b; t, d; k, g), diverting (m, n, ng), or obstructing (f, v; s, z, etc.) the flow of air from the lungs (opposed to vowel) [1]." Because consonants often occur at the beginning and the end of a syllable and contributes to the power of the sound in forms of plosives, we can visualize the waveform of the voices and analyze the difference in length of consonants in speaking and singing voices.

2.2. HARMONIC CONTENT/FORMANT

One of the most important attributes of one's voice that contributes to the timbre of one's voice. Age, gender and body condition all have an impact on the harmonic content in one's voice. Comparing harmonic content in speech voice and singing voice is effective because the fundamental frequency won't be the factor since the fundamental frequency in speech voice doesn't match up with the singing voice most of the time.

2.3. PITCH CONTOUR

Singing and speaking voices both have a underlying pitch contour, which outline the intonation and the melody of the voices. The melody of singing is divided into discrete pitches defined by different temperament. It is most common to use equal temperament which divide an octave into 12 pitches with equal intervals [2]. The climax of the melody usually matches the keyword that the songwriter want to emphasize on. In contrast, the pitch contour of a natural spoken utterance does not follow any temperament and glides rather continuously [3].

2.4. RHYTHM/PERIODICITY

The main difference in rhythm between speaking and singing voice is periodicity. "Meter describes the grouping of beats and their accentuation." as Julia Merrill, et al. stated in their research "Perception of Words and Pitch Patterns in Song and Speech". The rhythm of a singing voice is stricter than speaking voices as it sets boundaries to where the words need to be spoken and it could also break the pronunciation of a word into syllables. Speaking voices, on the other hand, only have restrictions on the word and phrase.

2.5. POWER

Power, which is also referred as volume, indicates the amplitude and energy of voices. It is assumed that the changes in power between singing voices and speaking are different. Singing voices are assumed to change according to

the music score, while speaking voices are overall more continuously.

3. METHOD

LPC (Linear Predictive Coding):
LPC, which is also referred as the linear predictive coding, is an audio signal processing technique that are used to analyze speech. It is the prediction of current sample as a linear combination of past samples [4].

$$\tilde{S}(n) = -\sum_{k=1}^{P} a_k \, S(n-k)$$

YIN Algorithm:
We used YIN Algorithm, a method of autocorrelation with modification, to find the pitch contour, periodicity and consonant length of the voices [5]. By observing the F0 contour graph generate using YIN Algorithm, we will be able to analyze the pitch contour over time by directly analyzing the F0 contour, the periodicity by determining the duration of each vowel and gaps between pitches and the consonant length by identifying the parts in the audio that don't have stable pitches as consonants and the parts with stable pitches and harmonics as vowels.

Here are the steps we used to achieve YIN Algorithm:

- Calculate difference function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

- Cumulative mean normalized difference function

$$P = \lim_{T\to\infty} \frac{1}{T} \int_0^T |x(t)|^2 \, dt$$

- Absolute threshold
- Parabolic interpolation

Power Spectral Density:
Power can be analyzed with the power spectral density (PSD), which illustrates the strength of energy as a function of frequency. We used power spectral density to show the location of the strength of frequencies variations. It has a unit of energy per frequency and we are able to obtain the energy easily by integrating over a specific frequency range [6].

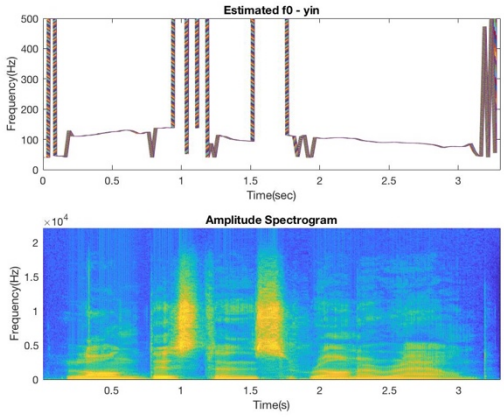
$$P = \lim_{T\to\infty} \frac{1}{T} \int_0^T |x(t)|^2 \, dt$$

4. EXPERIMENT

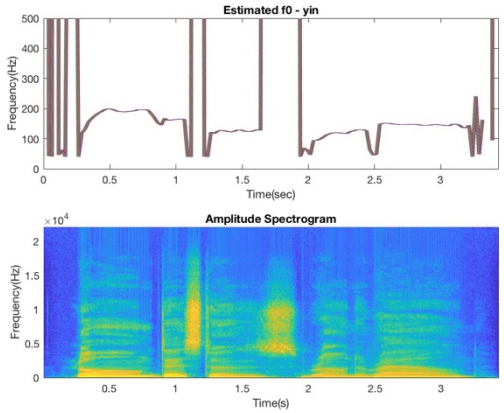
The subject, male, 22 years-old, read and sang the first line of the following 6 songs: *Don't Take Me Alive, Hey Jude, Layla, Long Distance Runaround, Patience, Pennyroyal Tea.*

5. RESULT & DISCUSSION

5.1. CONSONANT LENGTH:



Song 4 Speaking

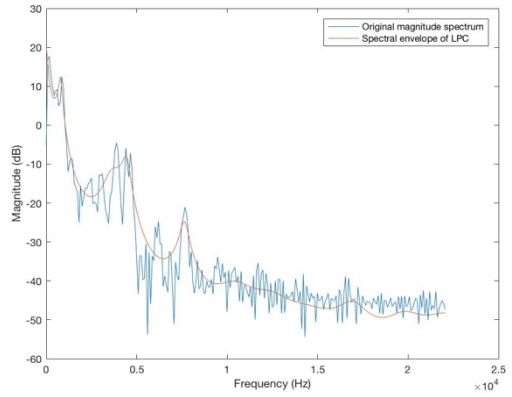


Song 4 Singing

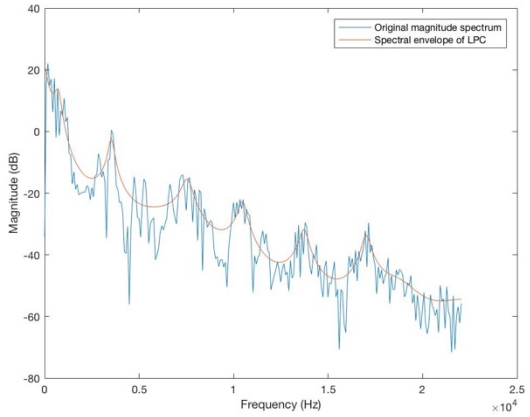
Because consonant has lower correlation with pitch, we use YIN algorithm to identify the parts in the audio that don't have stable pitches as consonants. We plotted the audio of one set of examples and measure the length of the part where the voice doesn't have a stable pitch. We also generated the spectrogram of the audio and found that there are no harmonics of consonants that can be seen on the graph and the region matches up with the part where the fundamental frequency of the audio can't be determined per YIN Algorithm. From the raw audio, on average, the consonants at the beginning of each syllable in singing voices are 50% longer than the consonants in speaking voices. However, after time-stretching the audio of the spoken examples and matching the total length of words with the sung examples,

we found that the consonants in speaking voices are 30% longer than the consonants in singing voices on average.

5.2. HARMONIC CONTENT/FORMANT



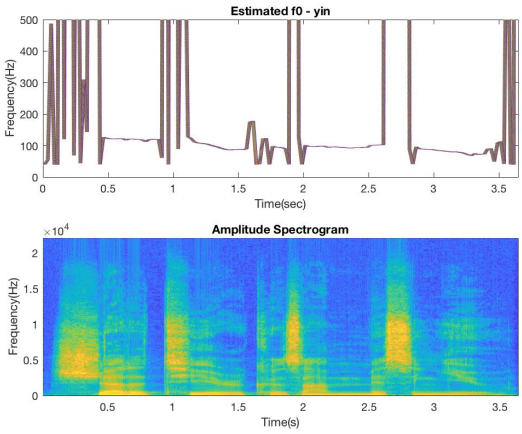
Song 4 Speaking



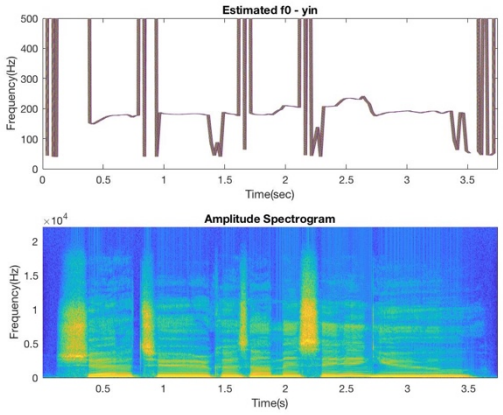
Song 4 Singing

LPC gives a representation of the spectral envelope of a digital speech. We use the graph to show the compressed spectral information of the audio [4]. The LPC graph shows that in speaking voices, the power of harmonics decreases in a linear manner as the order of the harmonics goes up. On the other hand, in singing voices, the power of harmonics decreases faster in lower orders (i.e. 2nd and 3rd harmonic) and decreases slower in higher orders so there is more higher order harmonics present in singing voices than in speaking voices. In practice, these differences make the speaking voice sound fuller and more masculine and singing voice sound thinner and less masculine.

5.3. PITCH CONTOUR & PERIODICITY



Song 5 Speaking

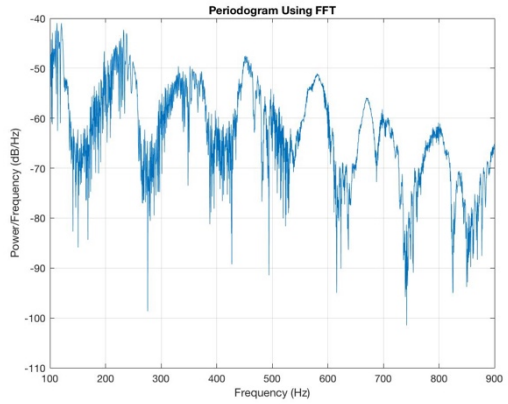


Song 5 Singing

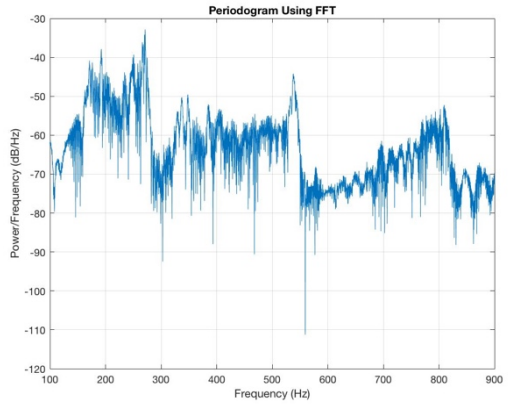
The diagram shows that the singing voices changes more successively over time, while speaking voices have some noticeable pauses in between pitches. The mean F0 of speaking voices are about 100Hz lower than singing voices and stabilized at a certain frequency, which is 100Hz for this subject. The speaking voice also has a noticeable low frequency at the beginning and end of each utterance while singing voices does not necessarily have this low frequency drop.

As for periodicity, there are more gaps between pitches in speaking voices than singing voice. The graph reflects the periodicity of singing voices and speaking voices. The duration of each vowel in singing is generally longer than speaking voices. These differences can be used to stretch or shrink the phoneme when trying to align the rhythm of speaking and singing voices.

5.4. POWER (POWER SPECTRAL DENSITY)



Song 3 Speaking



Song 3 Singing

Since the fundamental frequency range of a male voice is from 100-900Hz) [7], we only obtained the power spectral density in this range.

In general, the power spectral density of speaking voices both follow a similar pattern. Singing voices, comparing to speaking voices, vary rather smooth and indicate an excess of lower frequencies. Different than singing voices, speaking voices vary wiggly, which means there is an excess of higher frequencies. Based on all the periodogram we generated, we found out that there is an obvious peak at around 200 Hz and some other peaks at multiples of this frequency in the singing voices. The first frequency is the fundamental frequency of the signal and others are its harmonics. The peaks imply that the signal is strong at those frequencies. On the other hand, the periodogram of speaking voice have some peaks around same level and the distribution of the waveform is much more evenly than singing voices.

6. CONCLUSION & APPLICATION

The results from the experiment match the expectation and the estimation from listening to the subject. The LPC shows that speaking voice is more prominent in second and third harmonics while singing voice excites more higher harmonics. The PSD shows that speaking voice is more constant in power while singing voice is more dynamic. YIN Algorithm shows that the pitches shift more smoothly in singing voices and speaking voices have more sudden changes and pauses. YIN Algorithm and the Amplitude Spectrum Analysis show that the consonants are usually longer in speaking voices than in singing voices when matching the length.

This project finds several qualitative and quantitative difference between speaking and singing voices. These different features can be used in voice command detection, singing detection in music applications, speaking to singing synthesis, etc.

7. REFERENCES

[1] Dictionary.com, “Consonant” http://www.dictionary.com/browse/consonant?s=t

[2] The Editors of Encyclopaedia Britannica (2009), Equal Temperment, *Encyclopædia Britannica, inc.* https://www.britannica.com/art/equal-temperament

[3] Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012), Perception of Words and Pitch Patterns in Song and Speech. *Frontiers in Psychology*, *3*, 76. http://doi.org/10.3389/fpsyg.2012.00076

[4] Singh, L., and Garg, R. K. (2015), LPC Analysis of Speech Signal, *International Journal of Electrical and Electronic Engineering & Telecommunications,* 1, 2. https://pdfs.semanticscholar.org/bc0c/f32b08eda4a8cabfa283e70dabfe66391b8a.pdf

[5] Cheveigne´, A., and Kawahara, H. (2002), YIN, a fundamental frequency estimator for speech and music, *2002 Acoustical Society of America* http://audition.ens.fr/adc/pdf/2002_JASA_YIN.pdf

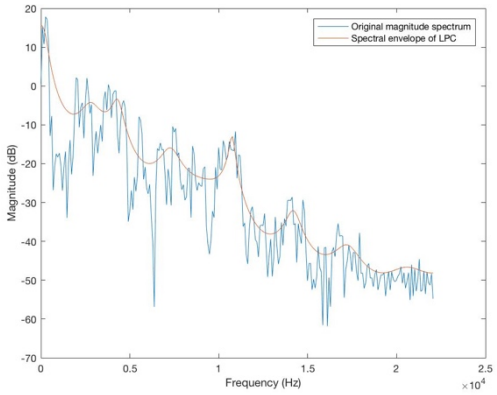
[6] Cygnus Research International, Power Spectral Density Function https://www.cygres.com/OcnPageE/Glosry/SpecE.html

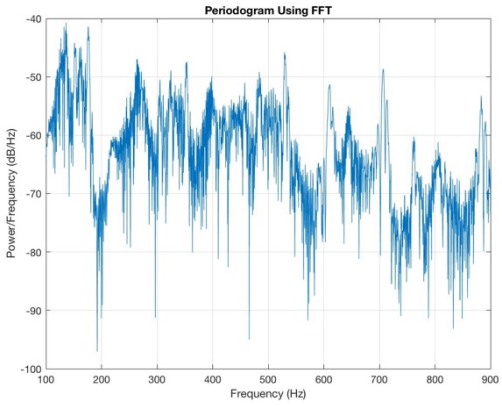
[7] Sound Engineering Academy (2017), Human Voice Frequency Range, *Sound Engineering Academy Blog* http://www.seaindia.in/blog/human-voice-frequency-range/

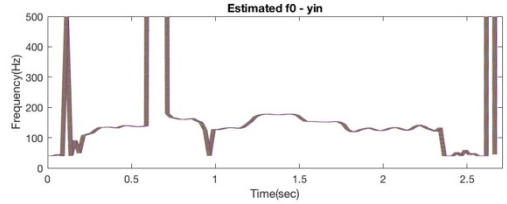
8. APPENDIX

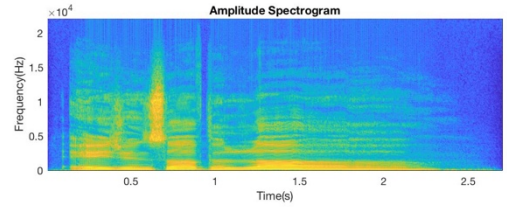
Song 1: Don’t Take Me Alive

Singing:

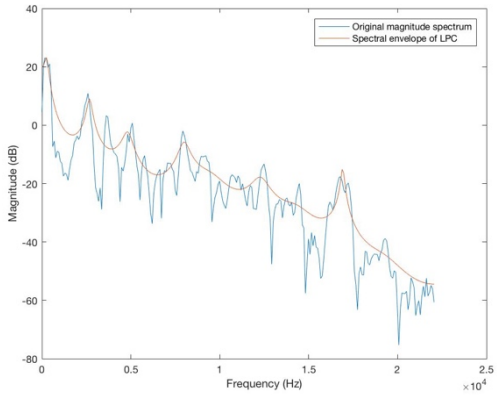


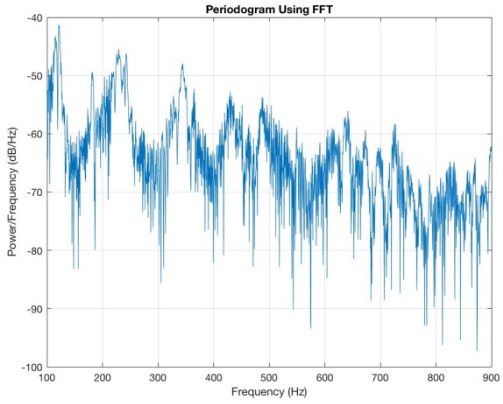


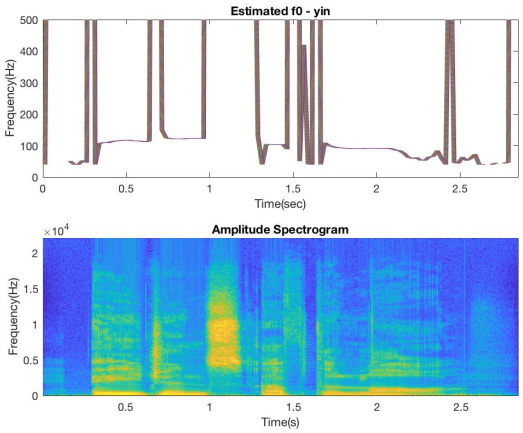


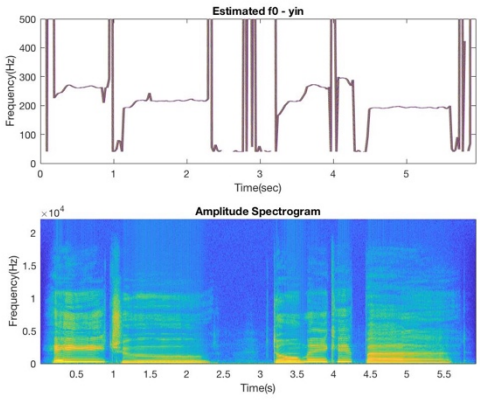


Speaking:

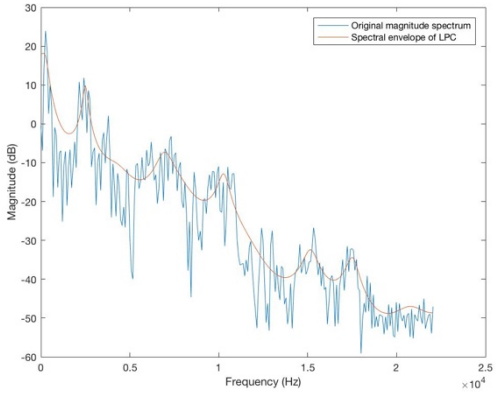




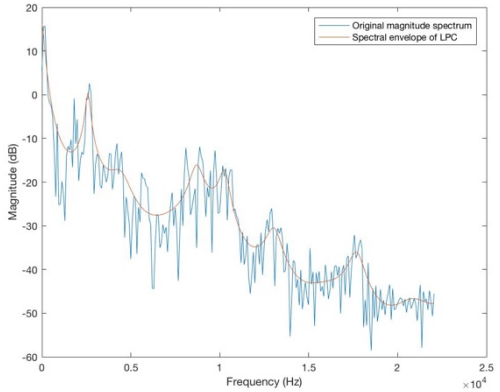


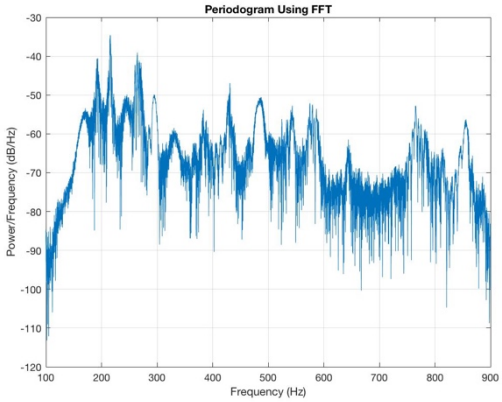


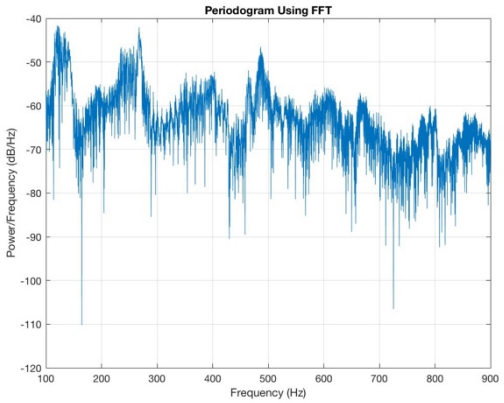
Song 2: Hey Jude
Singing:

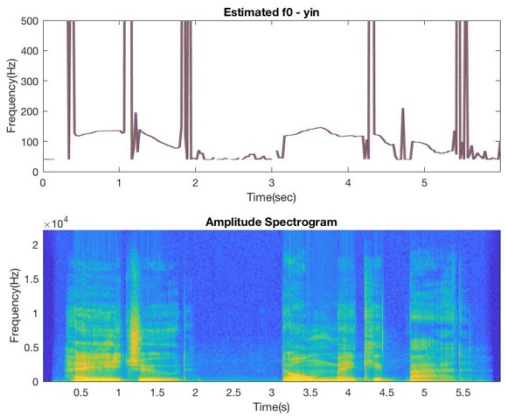


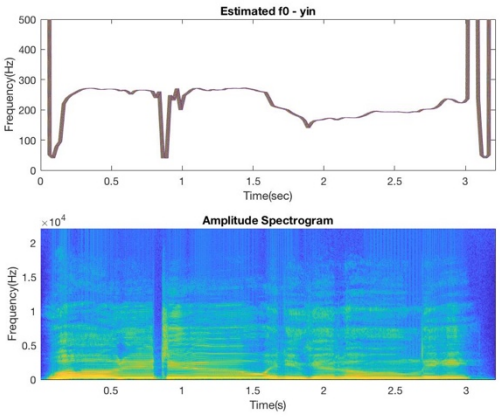
Speaking:



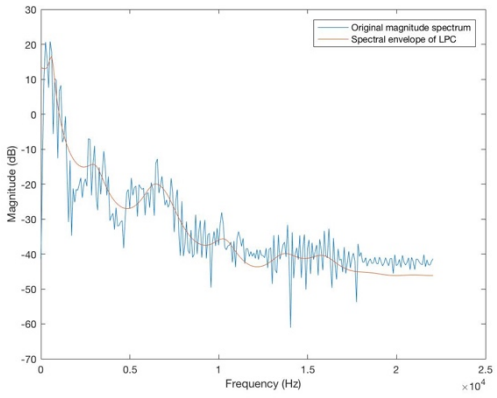




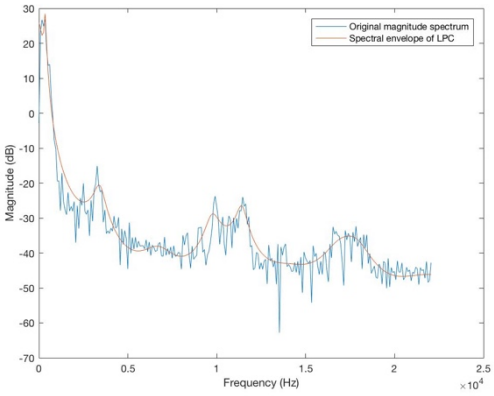


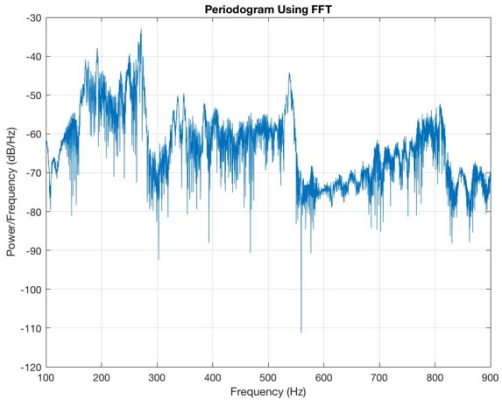


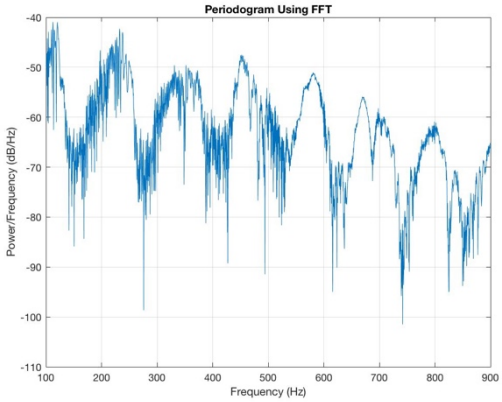
Song 3 Layla
Singing:

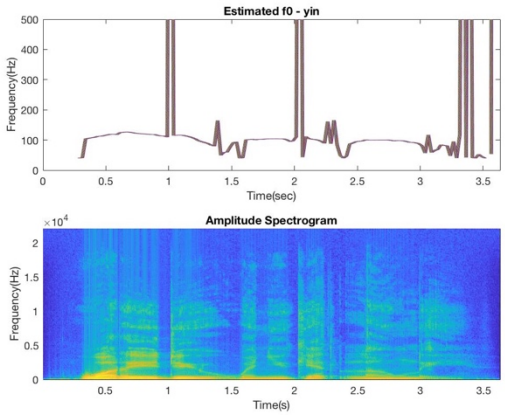


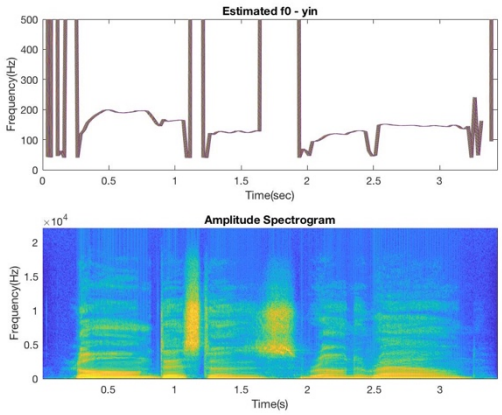
Speaking:





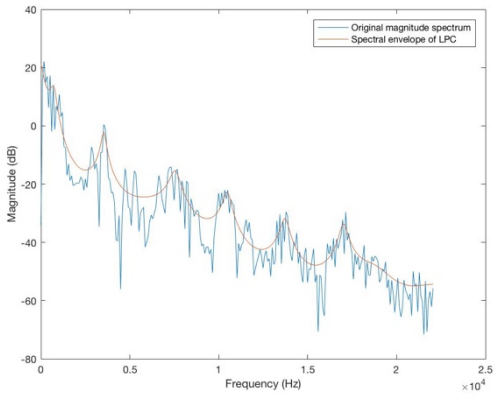


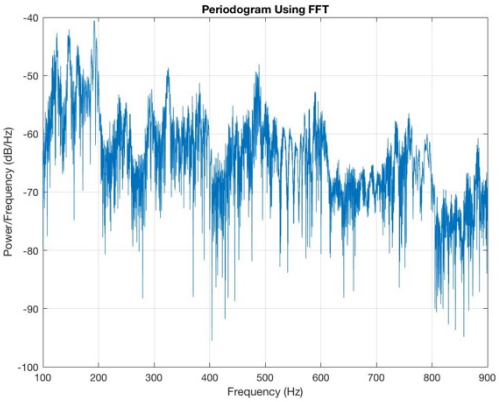




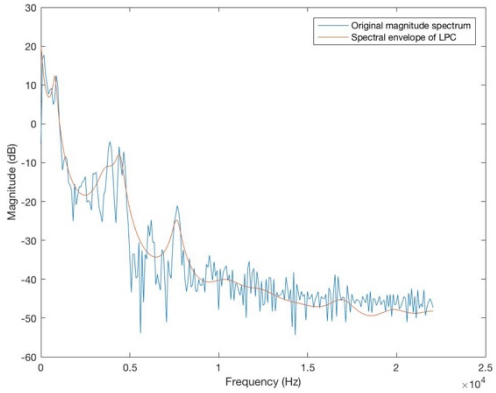
Song 4: Long Distance Runaround

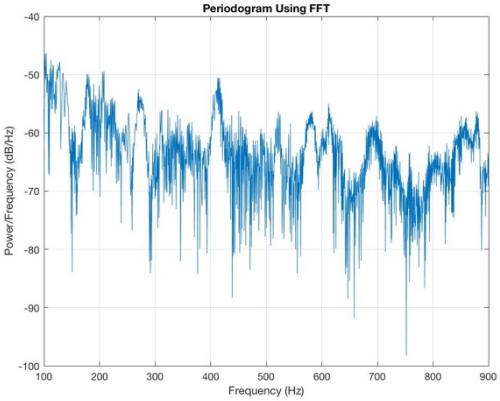
Singing:

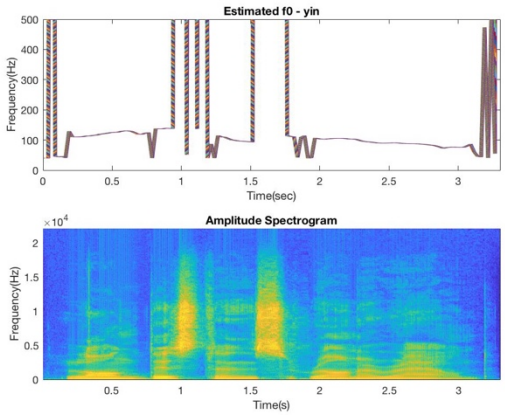


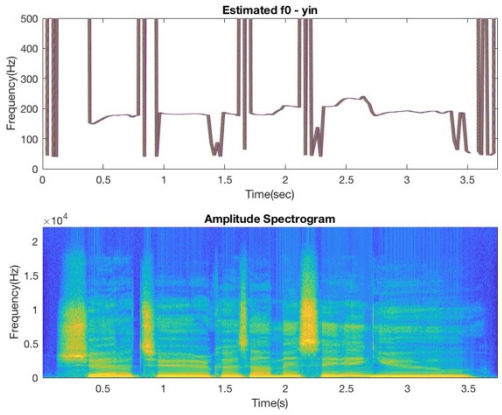


Speaking:



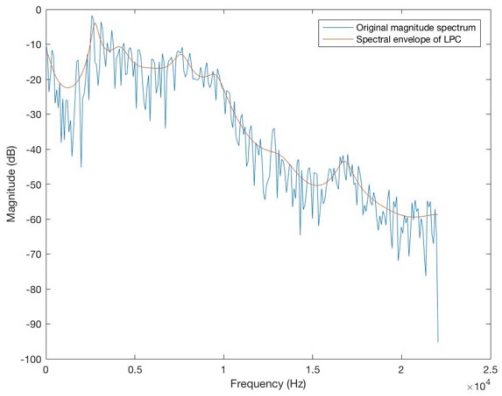




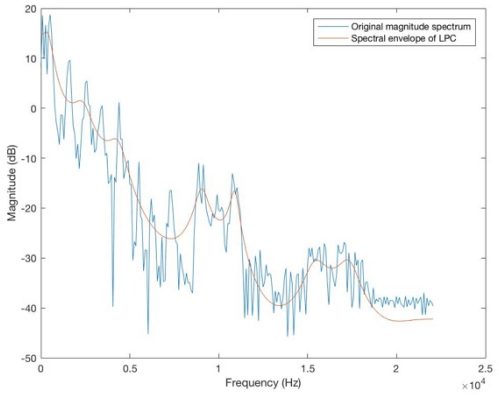


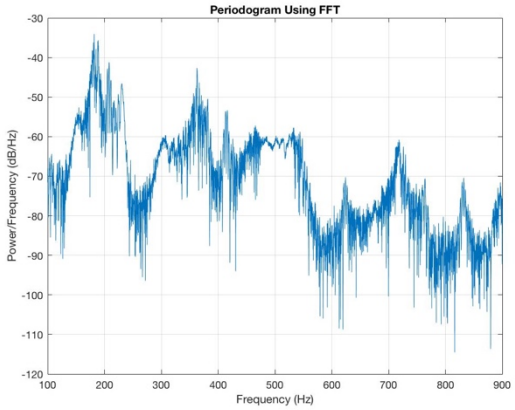
Song 5: Patience

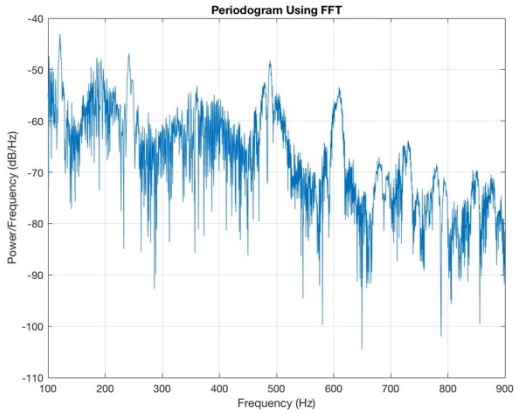
Singing:

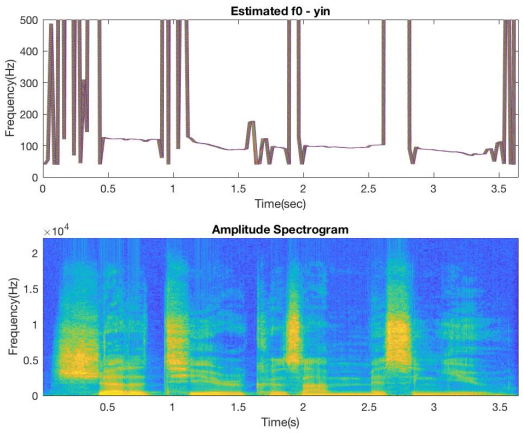


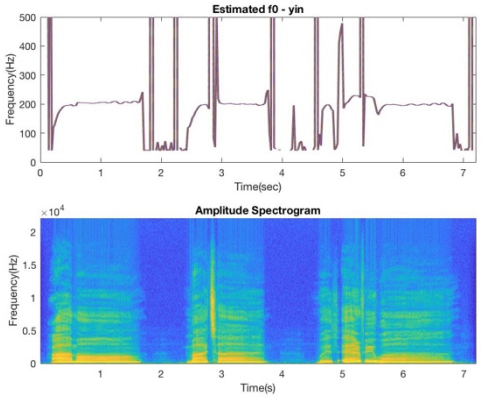
Speaking:





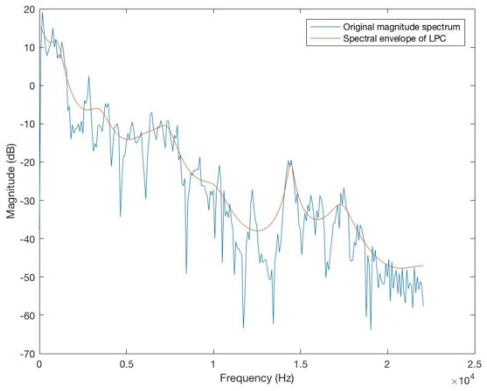




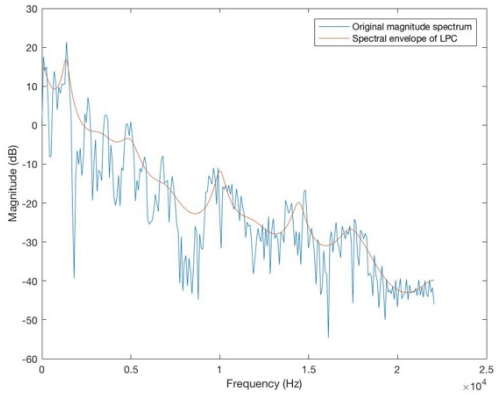


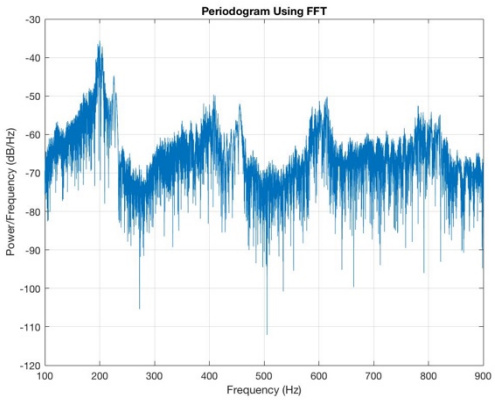
Song 6: Pennyroyal Tea.

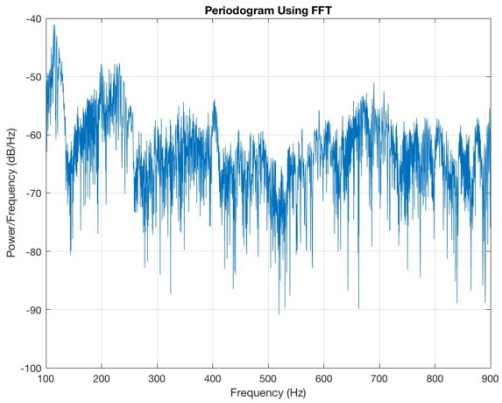
Singing:

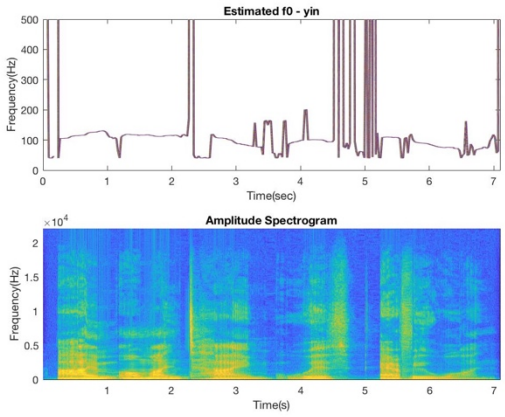


Speaking:








Estimated f0 - yin
Frequency(Hz)
Time(sec)
0
100
200
300
400
500
0
1
2
3
4
5
6
7

Amplitude Spectrogram
×10^4
Frequency(Hz)
Time(s)
0
0.5
1
1.5
2
1
2
3
4
5
6
7