CS571 Project Report

Vinay Visanji Faria , T21006

23-11-2021

Formant Estimation from Speech Signals

1 Summary

The project deal with finding formant from speech signal which are resonances that occur for certain frequency in speech signals.

2 Introduction

Sound wave have five characteristics such as Amplitude, Frequency, Time-Period, Velocity or Speed and Wavelength. The speech signal that are present in the real world are of continuous form. To analyse the contiuous signal we convert that signal to discrete by sampling. Sampling rate or sampling frequency is the number of samples per second. The samplling frequency can be find by Nyquist criteria which says sampling frequency should be atleast twice the maximum frequency present in a signal. A sound consist of one or more frequency component. To know about frequency components present in it, one such technique Discrete Fourier Transform (DFT) is used in digital signal processing. DFT is used to calculate frequency spectrum of signal. Frequency spectrum gives us information about frequency, phase, and amplitude of the component sinusoids. But with DFT we can only find frequency component and we can't know in which interval they are present. So, we use STFT (Short Time Fourier Transform) which is done by dividing the complete signal into smaller signal and applying DFT to smaller signal. STFT formula is given by

$$STFT\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{i=-\infty}^{+\infty} x[n]w[n-m]e^{-j\omega n}$$

Input signal $x[n]$ and window $w[n]$ which can be rectangular, triangular, hamming, etc. In this case, we are using fft for STFT on signal therefore m is discrete and ω is discrete. LPC determines the coefficients of a forward linear predictor. It has applications in filter design and speech coding. 'The peaks that are observed in the spectrum envelope are called formants'[1]. These formants frequency are resonance frequency of the vocal tract. Speech formants can be used in emotion recognition, sex discrimination, diagnosing different neurological diseases, etc. We can use this to distinguish between voiced and unvoiced in speech signal. It is known that 4 peak frequencies of the vocal tract, called formants, are enough to descriminate most vowels. This is one reason LPC works well with speech.

3 Solution

3.1 Assumptions

The audio assumed to have a single channel so that data returned by read function is simply an array containing the amplitude of each samples. LPC assumes the filter is a p-th order all-pole filter. Though not exact, it provides an extendable method for modeling resonances. This also allows for a tractable solution when estimating h(n) from x(n),

3.2 Algorithms used

1. Reading data from audio file and finding sampling rate of audio.
2. Giving hop length and window size
3. plotting audio signal in time domain
4. Using enframe function to divide audio data into frames.
5. Finding FFT of each frame
6. Calculating LPC (Linear predictive coding) coefficients, you can construct impulse response signal
7. Plotting the LPC spectra as well as the original speech spectrum.
8. Finding formants

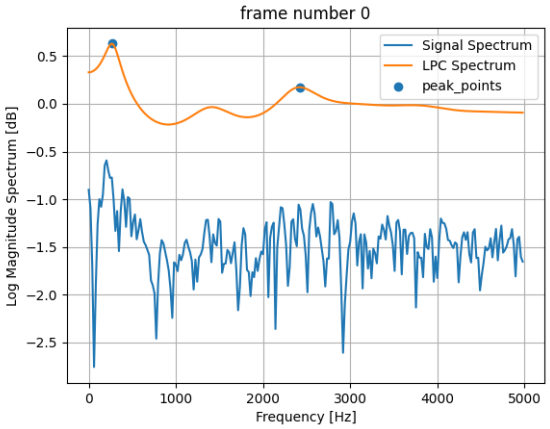
4 Conclusion

The formant frequencies are properties of the vocal tract system and need to be inferred from the speech signal rather than just measured. The spectral shape of the vocal tract excitation strongly influences the observed spectral envelope, such that we cannot guarantee that all vocal tract resonances will cause peaks in the observed spectral envelope, nor that all peaks in the spectral envelope are caused by vocal tract resonances. The problem with root-finding algorithms is that the determination of formant frequencies and bandwidths is only successful for complex-conjugate poles and not for real poles. Peak-picking techniques are vulnerable to merged formants and spurious peaks [2].

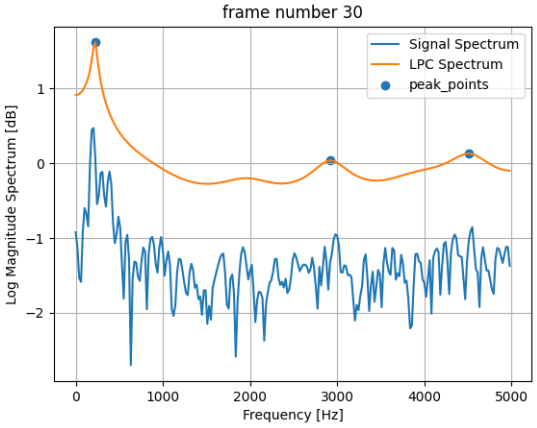
5 Project Github page

`https://github.com/VinayFaria/CS571_PROJECT`

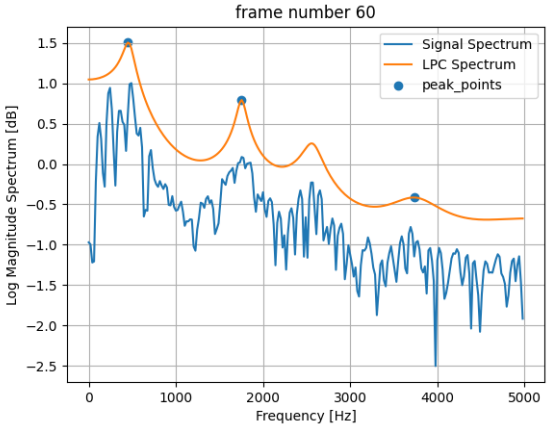
6 Results and analysis




```
[273.3609364146423, 1395.4624596142548, 2389.9745237933107, 3094.8000702009817,
3814.3564959622995, 4622.615730891139]
```




```
[220.58781503562412, 838.1704839756322, 1959.893787055322, 2919.5858714946294,
3880.2453555035822, 4519.211844018118]
```




```
[215.0786426424495, 462.4052118992027, 1754.6611375169568, 2568.7727383131855,
3747.6348043119783, 4886.21691990001]
```

Figure 1: should.wav spectrum and formants respectively

References

[1] A. H. Benade, *Fundamentals of musical acoustics*. New York: Oxford University Press., 1976.

[2] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 36–48, 1998.