CS489 Computational Sound P1: Formant Analysis

Jerome Couture // UWID 20436951 // March 17th, 2016

**Introduction**

CS489 thus far has been largely theoretical in nature, outlining a variety of techniques for audio analysis. And yet, audio is an incredibly broad topic to cover – as discussed in lecture, the range of frequencies that can be detected by a human being range from 20Hz to 20khz. Such content could include music, sinusoids, human speech, animal cries, construction work… anything that a given person has heard before. The human auditory system is significantly more sensitive to frequencies in the 1kHz to 6kHz range, as discussed during the very first lecture of the course.



Source: Wikipedia

These curves go a long way as to explaining how people perceive certain sounds. For instance, it is nearly impossible to ignore the fact that a baby is crying – the voice measures between 1 and 5kHz, centered at around 3.5kHz. And many people have difficulty concentrating when there are others around them talking – a good component of the energy in human speech lies within the frequency band that the human auditory system is the most sensitive to.

**The Short-Time Fourier Transform**

All of the investigative work done on this project was done in the discrete domain, though the techniques do have foundations in analog signals. The Fourier Transform (FT) takes a time-based pattern, measures all possible “cycles”, and returns the coefficients of each cycle that a given time-based pattern is comprised of. It can also be though of as expressing an arbitrary waveform as a linear combination of sinusoids of a myriad of frequencies. Indeed, the Fourier transform of a pure sine wave is represented as a single value, containing information as to its amplitude and frequency.

The formulation for the Fourier Transform is described as:

Each cycle is a complex exponential; Euler’s formula provides some insights as to how these complex exponentials, each with a different frequency, can be summed to generate an arbitrary waveform.

This transformation also contains the negative frequencies, typically in frequency domain analysis this information can be discarded. Each complex exponential will have a real and an imaginary component, information can be extracted from them by calculating their **magnitude** and **phase**. The traditional way in which speech is analyzed involves taking a long signal, breaking it up into short chunks, and analyzing those small chunks using the Fourier (and other) transforms. This procedure is called the **short-time Fourier transform**, and it displays the time-frequency domain. Analysis of the power spectrum in the time-frequency domain is a core concept in the field of linguistics. Linguists and speech therapists used to examine spectrograms to describe the quality and contents of human speech.

In the spectrogram, there are a number of dark bands which represent areas of high intensity. Depending on the window size and sample rate used for the short-time Fourier transform, sometimes these bands will blend together. A longer “chunk” of signal will allow for better resolution on the frequency axis at the expensive of computational cost.

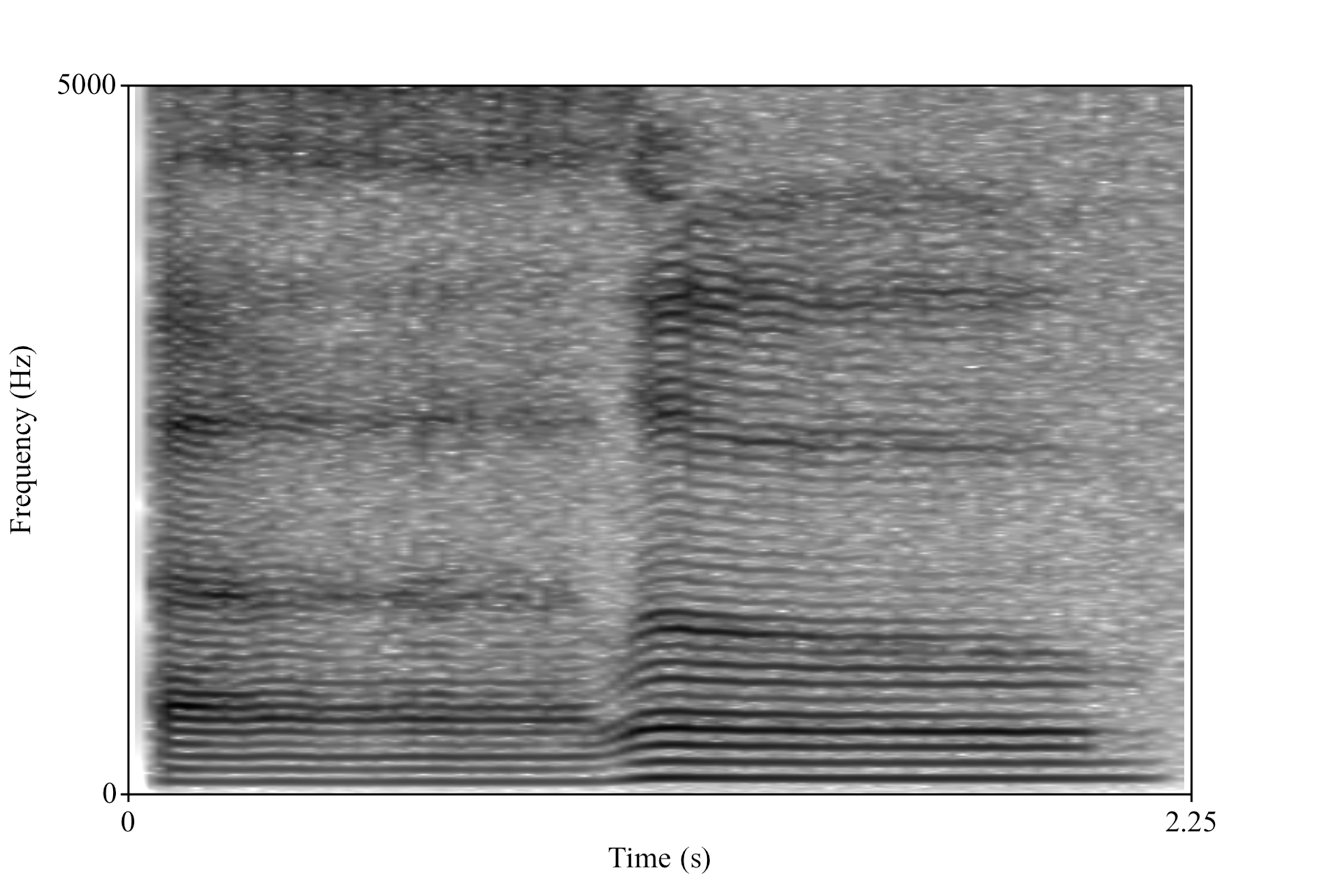


Figure 1: Spectrogram of a sung “Hello”, going from approximately 88Hz (E2) to 112Hz (A2).

In figure 1 above, there are more widely varying “trends” for the two syllables that were spoken in “Hello”. There is a specific way of representing phonetics – “hello” is written as həˈlō in this notation. The larger, darker bands at approximately 400, 1400, and 2600Hz are what are called “formants”. Each vowel has unique values for each formant, and the transition from consonant to vowel has a unique shape. These formant shapes are what result in the perception of speech! To illustrate, the spectrogram above will have its formants marked using the audio analysis tool “Praat”.

It is important to note that the peaks in the spectrogram (dark lines) that are lower are not formants. Rather, they are the fundamental frequency (F0) and harmonics that build upon it. Because the sample is sung rather than spoken, these harmonics are emphasized more than a typical speech signal. The first, second, and third formant are most commonly used as features to identify vowels, and their trajectory through time is used to identify some consonants. Identification of consonants tends to be much more difficult than vowels – a myriad of categories exist. Some examples are nasal, stop, sibilant and non-sibilant affricate, sibilant and non-sibilant fricative, approximant, flap, trill, and many others. English is not comprised of all possible sounds a human can make, however – there are many guttural, sibilant, and plosive noises that exist in other languages.

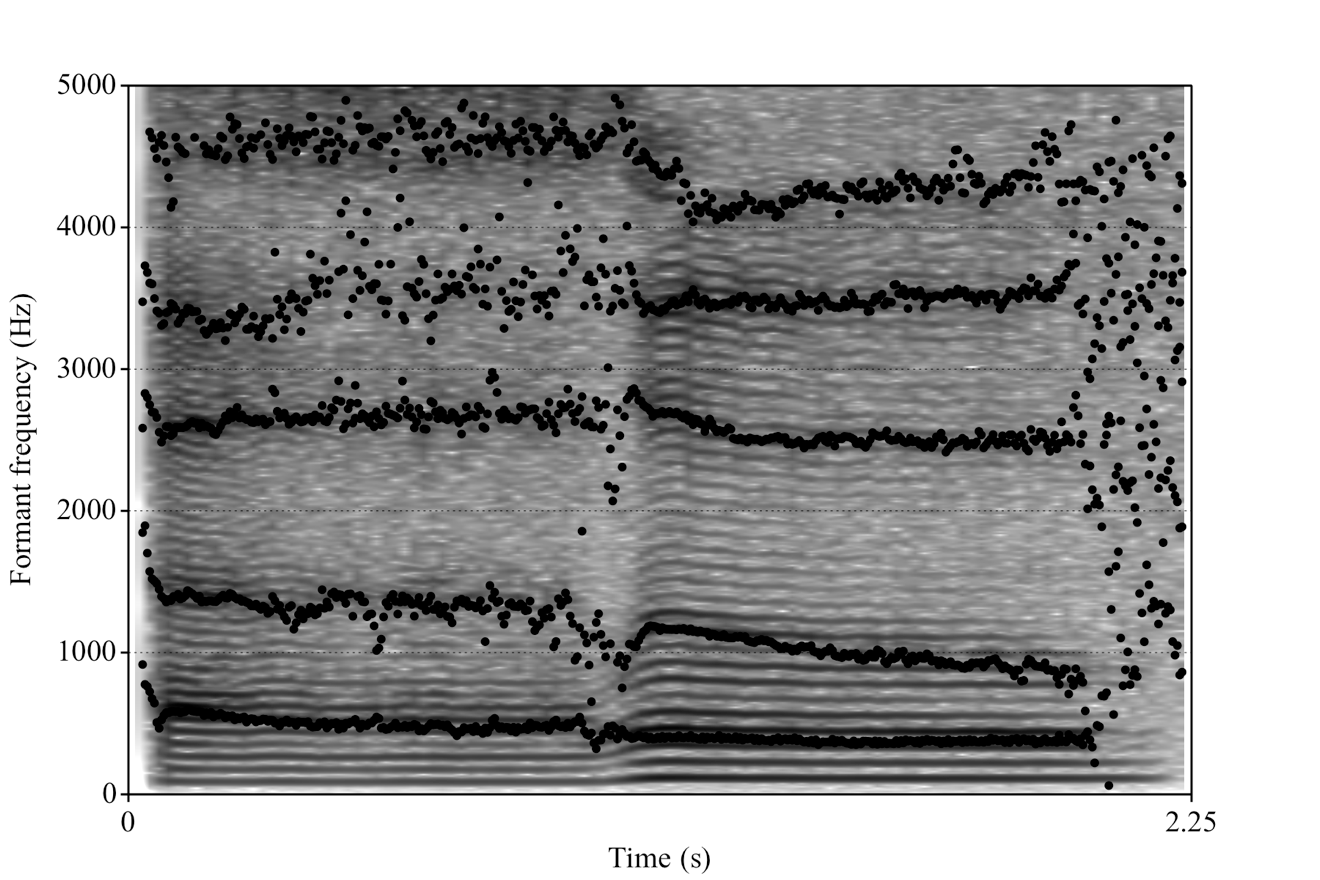


Figure 2: Spectrogram of a sung “Hello” with the first five formants (F1-F5) marked.

Somewhat counterintuitively, the note that is being sung does not influence formant location very much. This can be seen by having a subject change the pitch of a continuously enunciated vowel in one voice clip, and having them say numerous vowels at the same pitch. This can be seen clearly in the images below. The formants are most commonly calculated using Linear Predictive Coding (LPC), a method which models the human vocal tract as a series of impulses and attempts to design a filter which can transform those impulses into a given voice sample.

The sample speech signal is to be replicated by the difference equation:

Where x[n] is the impulse train with a frequency of F0, y[n] is the intended speech signal, G is a “history parameter”, and ai are the filter coefficients of an all-pole filter. In MATLAB, the design of such a filter is trivial: the “lpc” function simply returns the model estimate, and the prediction error variances. Conveniently, the model estimate corresponds to the poles of the filter LPC is supposed to be implemented by, so plotting the frequency response should give what is called the LPC spectrum. The frequencies at which the LPC spectrum for a short window in time are maximized are the locations of formants! The LPC spectrum can also include the fundamental frequency (F0), so the peaks located below a frequency threshold should be ignored.

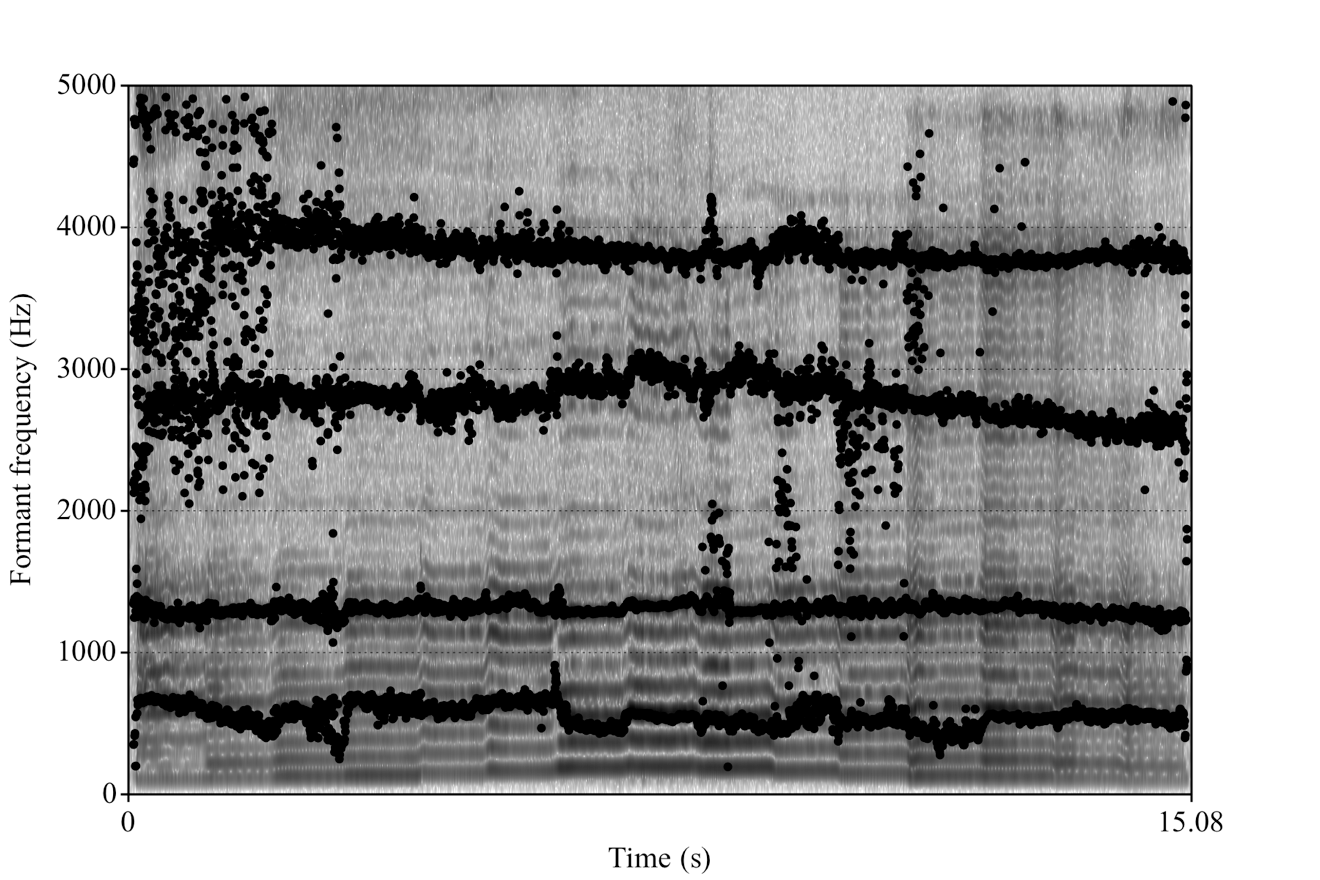


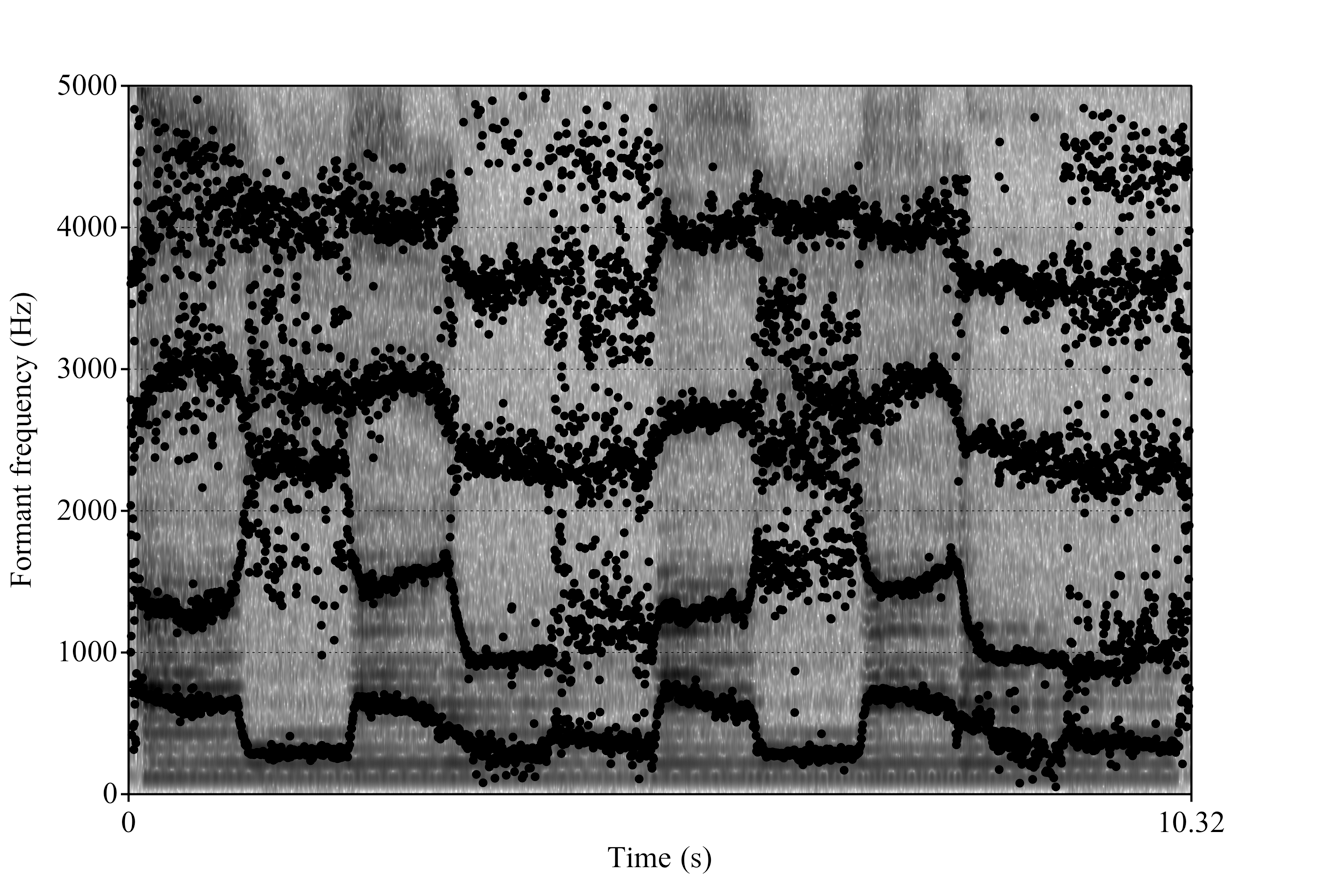
Figure 3: Spectrogram and formants for singing the same vowel, but many pitches  


Figure 4: Spectrogram and formants for singing many vowels at the same pitch