

BACKGROUND*

Yuqiang Mu
Adrian Galindo
Gbenga Badipe

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 3.0[†]

Abstract

Offers a basic explanation of several concepts relevant to our voice recognition approach.

1 Background

1.1 Formants

In speech science, formants are defined as the regions of concentrated power in the frequency representation of a vowel. These specific peaks of power are a reflection of a specific configuration of the different parts of the vocal tract; humans produce different kinds of vowels, each one characterized by its pattern of formants, by vibrating their vocal chords and changing the configuration of their articulators (e.g. positioning of the lips, tongue, and jaw). Most vowels across a single person's speech can be completely described by their F1 and F2, the formants with the lowest and second-lowest frequency centers; notable exceptions to this are the "rhotacized" or "r-colored" vowels, which depending on the speaker's vocal tract may often feature overlapping F1 and F2 with other non-rhotacized members of the vowel space and must also be identified by their F3, and a few lip-rounded vowels as distinguished from their unrounded counterparts, which have lowered high-level formants in general but may still overlap with other vowels in F1 and F2.

*Version 1.3: Dec 20, 2011 5:38 pm -0600

[†]<http://creativecommons.org/licenses/by/3.0/>

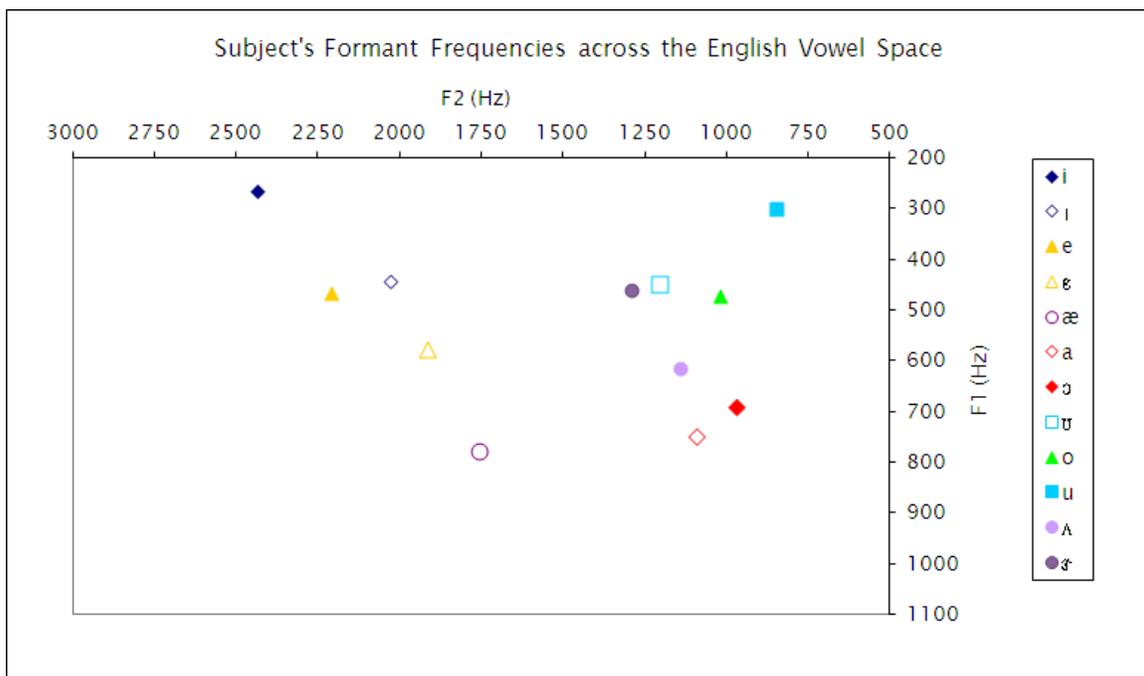


Figure 1: Example formant plot of subject. Note the closeness between [ɪ] and [ʊ].

However, the actual frequencies of the formants are obviously not constant across different people; in fact, the relationship between the formant patterns of different speakers saying the same vowel is highly erratic. This can be attributed to the high degree of variability found in the physical characteristics of the human vocal tract; males often have larger and longer vocal tracts than females, two people may not have the same jaw size or shape, etc. This variability affects the resonant frequencies of the vocal tract's various components, so that even if two people are making the same articulatory gestures their formants will not only be different, but different in a decidedly nonlinear way for each vowel; that is, physical differences between speakers make their formants for multiple vowels differ in different ways. For example, pertaining to the features of rhotacized vowels mentioned above: both F1 and F2 of the rhotacized vowel [ɝ] (as in "herd") overlap with those of either [ɪ] (as in "hood") and/or [ʊ] (as in "who'd) for many speakers; however, short preliminary measurements of our subjects (all speakers of standard American English, though with slight and different dialectal influences of his own) taken for testing purposes before running any actual trials revealed that in one there was only slight F1 overlap between [ɝ] and [ɪ], in another there was F1 overlap between [ɝ] and [ʊ], and in a third there was no clear overlap at all. Larger studies in the realm of Linguistics on this particular kind of overlap do report that it is rather commonly found, but in varying degrees: a testament to the amount of vocal tract diversity that exists in the human population.

This variability makes it highly unlikely that two people will have anything close to a nearly identical set of formant frequencies across a large set of vowels; more topically, this means that a voice-recognition system that knows the average formants of a specific person's vowels would be difficult to dupe for other people solely on the basis of their vowel characteristics. For all intents and purposes, the more vowels used in the trials the more specific and rigorous the trials become (our project ended up utilizing all twelve members of the English vowel space).

1.2 "hVd" words

The "hVd" words are a class of monosyllabic English words with a single vowel nucleus that follow the specific pattern of beginning with an [h] and ending with a [d]. This set is particularly interesting because of the unobtrusive nature of the word-initial and word-final consonants; for the most part they do not interfere with the frequency characteristics of whatever vowel is in between them. In contrast, most of the other consonants in the English language will, in some way or another, alter or bend the formants of the surrounding vowels (e.g. vowels with adjacent nasals like [n] or [m] may become nasalized, changing their formants considerably; formants of vowels next to approximants like [l] will "drift" toward those of the approximant targets instead of remaining steady; etc.). It is for this reason that these hVd words were used in our trials; they provide a clear picture of what a speaker's uncorrupted vowels look like and they represent a middle ground of sorts between saying pure vowels and normal speech with other, more complicated words.

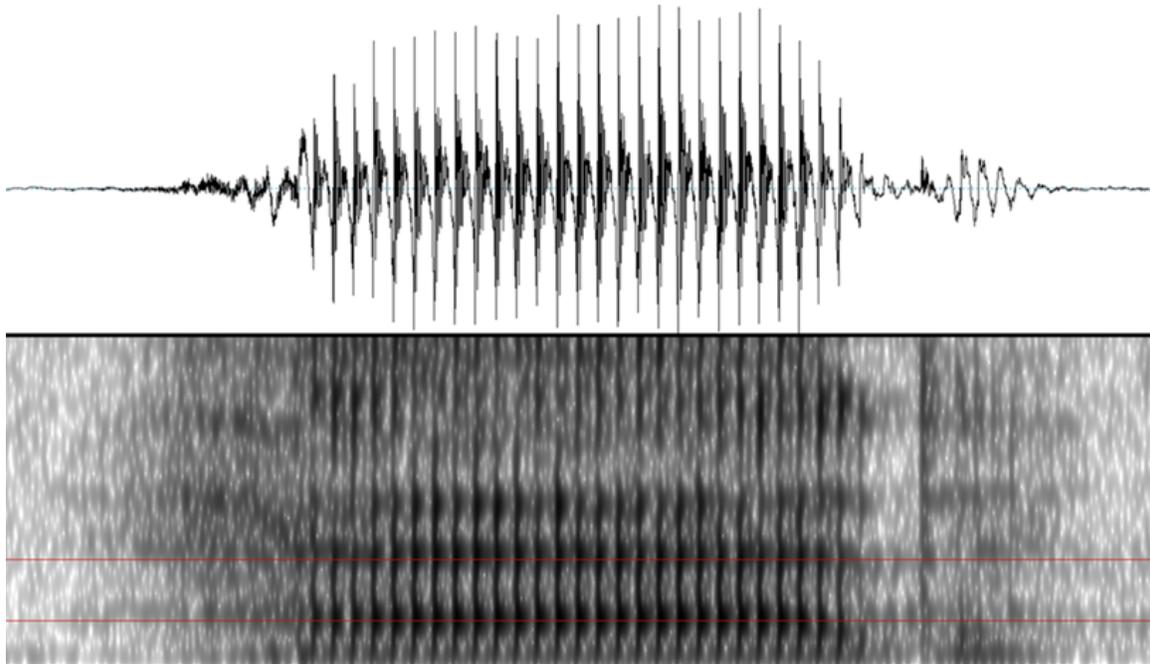


Figure 2: Example waveform and spectrogram of the word "had." Note the stability of the formants (centers approximated by red lines).

1.3 Windowing

It is quite commonplace in the realm of Fourier analysis (especially spectrum analysis of sound waveforms) to obtain a time-domain-based view of the frequency content of a signal. This is addressed by "windowing" the data in question in order to obtain a series of segments of the original waveform, each representing a particular segment in time of the (discrete-time, naturally) signal; the process of creating each of these segments is, in short, taking the product of a windowing function of specific length and the signal at that particular time.

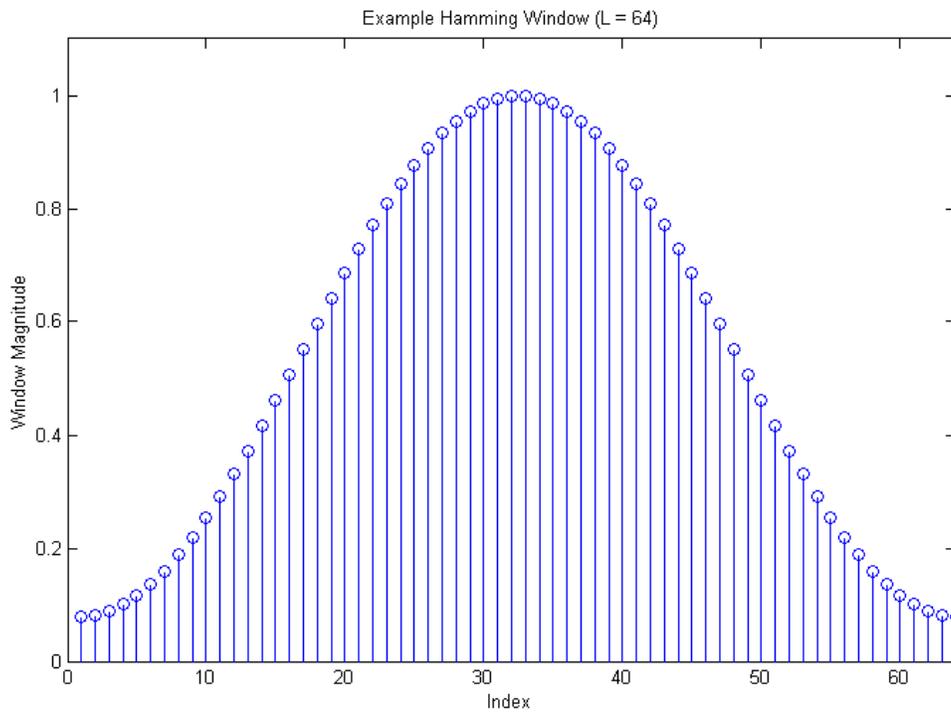


Figure 3: Example plot of a length 64 Hamming window (magnitude vs. discrete-time index).

The windowing function is zero outside its interval, and is usually tapered down to zero or near zero to reduce the erroneous high-frequency content that would be introduced by sharp cutoffs at the window ends; in addition, the windows are staggered in steps of half the window length to help mitigate potential information loss due to the attenuated ends of the window function. The window of choice for this project was the Hamming window, a "raised cosine" shape (see the above figure). To be brutally honest, this window function was not the optimal choice for this particular application; however, the window choice is not a large issue for us because the core of this project does not depend much on factors like high side lobe attenuation and falloff for the window function (we are only concerned with locating the positions of peaks within a certain frequency range for each window of the signal, and in fact we bandpass-filter out all the content outside the normal human range for the two formants in question).

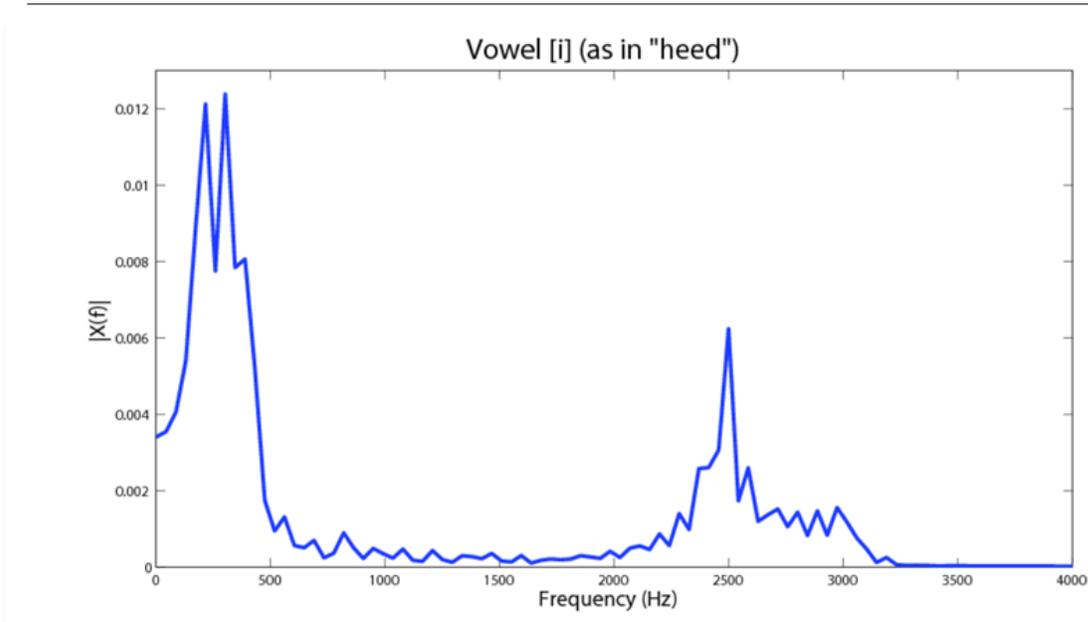


Figure 4: Example plot of the FFT of a length 1024 window of the vowel [i] in "heed," spoken by a test subject. Clear formant spikes observed at ~ 400 Hz and ~ 2500 Hz in this particular window.

A window length of 1024 samples was chosen for this project with consideration given to the nature of the time-frequency tradeoff inherent to the windowing process; with increasing time resolution (i.e. more windows for a given signal, more bins of time information) comes less frequency information, and with increasing frequency resolution comes less temporal information (i.e. less windows for a given signal, less bins of time information). The standard accepted window length used widely in linguistic analysis applications is 0.025 seconds, an interval that balances time and frequency resolutions for the windowed signal reasonably well. With sampling frequency of 44100 Hz for our waveforms, we chose the nearest power of 2 that would give us the necessary time interval (1024 samples is slightly under 0.025 seconds, but the margin is so small that it does not affect our results at all).