Low-frequency Fourier analysis of speech rhythm

Sam Tilsen
Keith Johnson

Abstract

We present a new methodology for studying speech rhythm, based upon low-frequency Fourier analysis of the amplitude envelope of bandpass-filtered speech. Rather than quantifying rhythm with time-domain measurements of interval durations, we use frequency-domain representations to characterize speech rhythm with a metric based on a *rhythm spectrum*. In this paper we describe our method in detail, and using the Buckeye corpus, we illustrate and discuss approaches to characterizing rhythm with low-frequency spectral information.

1. Introduction

Most studies of speech rhythm use interval durations to describe temporal patterns. Defining an interval duration requires the selection of endpoints, which are moments in time. A variety of rationales exist for how to choose relevant moments, and these rationales depend upon what information in the acoustic signal is considered most appropriate for identifying rhythmic patterns. A guiding principle in the selection of moments is to define intervals that are psychologically important, i.e. subject to cognitive control or perceptually relevant.

One of the most prevalent uses of interval durations is found in cross-linguistic studies of speech rhythm, which have generally been concerned with the hypothesis that languages belong to distinct rhythmic classes of stress-timed, syllable-timed, and mora-timed languages (Pike 1945; Abercrombie 1967). This hypothesis holds that speakers of a given language organize their utterances so as to produce relatively isochronous intervals between either stressed-syllables, syllables, or morae. If true, evidence of such timing should be found in a tendency toward isochrony of intervals. In other words, the durations from the onset of one unit to the next should be less variable for timed intervals than for intervals not used to organize speech. For example, in a stress-timed language, interval durations between stressed syllables should be substantially less variable than those between syllables or morae. In a syllable-timed language, inter-syllable durations should be relatively less variable than inter-moraic and inter-stress interval durations. Consequently, the durations of inter-stress intervals in a stress-timed language should *not* be highly correlated with the number of syllables contained in those intervals; likewise, in a syllable-timed language the durations of syllables should not be highly correlated with the number of morae in the syllables.

Research into these predictions has failed to reveal a greater degree of syllable and stress-interval isochrony in syllable-timed and stress-timed languages, respectively. Bolinger (1965) and later on Dauer (1983) found that interstress interval durations in English were proportional to the number of syllables contained therein and varied according to their constituent syllable shapes. Furthermore, Roach (1982) found similar syllable durations in several languages that are purportedly stress-timed and several that are purportedly syllable-timed.

One reason that attempts to identify isochronies have not been successful may be, as Lehiste (1979) suggested, that isochrony is perceptual in nature. Perhaps a cyclic rhythm is

perceptually imposed on only vaguely isochronous productions. Hence relative isochrony would not necessarily be observed in speech interval durations. Along these lines, the perceptual system may compensate for rhythmic perturbations due to factors such as syllable shape or number of syllables in a foot. Just as perceptual mechanisms correct for contextual variation in the signal due to the overlap of articulations, so might rhythmic variation due to these factors be corrected, leading to the perception of isochrony not present in the acoustic signal itself.

A different interpretation of the failure to identify rhythmic classes with interval measures is that the choice of salient moments used in these studies did not define the intervals most appropriate for uncovering patterns of isochrony. That is, the intervals between onsets of syllables or onsets of stressed-syllables do not correspond to consciously controlled or perceptually relevant durations, and thus do not reveal a cross-linguistic pattern that correlates with perceived rhythmic differences between languages.

There are other reasonable possibilities for the choice of moments to define cognitively important intervals. One is the beginning of a syllable vowel nucleus, another is the amplitude peak associated with a syllable, yet another is the center of integrated amplitude (analogous to a center of gravity). A more complicated but more cognitively-motivated construct is the "p-center" of a syllable. This concept arose from *finger tap alignment* studies by Allen (1972, 1975), in which subjects tapped their index finger along with stressed syllables. Allen found that subjects tapped their fingers somewhere close to vowel onsets, and he called the temporal location of the finger tap the "production-center" of a syllable. The concept was then refined using a *dynamic rhythm setting task*, in which the timing of a syllable relative to reference beats is manipulated with a knob. These experiments have shown that a perceptually salient center of a syllable also exists near the vowel onset (Morton, Marcus, & Frankish 1976), and that the exact location of the center is influenced by the presence and duration of onset and coda clusters.

These findings have led a number of researchers to attempt to develop algorithms to approximate p-centers from acoustic signals. Howell (1988) used the amplitude envelope of a syllable to predict the location of its p-center. Pompino-Marschall (1989) used a gammatone filterbank (which approximates auditory nerve responses) and a nonlinear function of energy events defined by thresholds in syllable constituents; Scott (1993) used the energy in a specific band of the spectrum, and Cummins and Port (1998) used a variation on this method where the band was 700-1300 Hz—this frequency-band is useful in identifying energy associated with vowels and filtering out energy associated with obstruents.

An alternative reason why convincing evidence for isochrony has not been found could be that rhythmic differences between languages do not arise from isochrony whatsoever. As Dauer (1987) has argued, there may be no way to choose consistent points in time to define intervals that are relatively isochronous. In other words, the hypothesis that languages fall into distinct rhythm classes, based upon which prosodic units tend to be isochronous, could just be plain wrong. In that case, our impressions of cross-linguistic rhythmic differences could arise from perceptual attention to a different sort of acoustic information.

One intriguing approach along these lines is to attribute perceived rhythmic differences to phonological characteristics of languages. Dauer (1983) observed that stress-timed languages tend to have more syllable types and heavier syllables, along with weight-sensitive stress; additionally, stress-timed language have greater vowel reduction in weak syllables. However, not all languages exhibit all of the characteristics prototypical of a particular class, and so Dauer (1987) has argued that there is a continuum from stress-timed to syllable-timed languages.

To address these observations, Ramus et. al. (1999) operationalized Dauer's hypothesis using measurements of vowel and consonant durations in an utterance. Their rationale for this approach is attributed to Mehler et. al. (1996), who proposed that infant speech perception relies primarily on vowels, partly because consonants tend to have less acoustic energy. Ramus et. al. (1999) used three measures derived from interval durations: %V, the proportion of the duration of a passage taken up by vowels, and ΔC and ΔV, the standard deviations of the durations of consonantal and vocalic stretches in running speech. With these parameters they were able to distinguish syllable and stress-timed languages: stress-timed languages tend to have relatively high ΔC and ΔV, and low %V, reflecting the observation that vowels sometimes reduce and syllable shapes are more diverse in these languages. In contrast, syllable-timed languages tend to have relatively low ΔC and high %V, reflecting a more limited set of syllables and less reduction. Japanese has an even lower ΔC and %V—perhaps indicating a third moraic-timing cluster.

Despite the success of this approach, the question remains as to whether the phonological properties of languages are what give rise to the percept of rhythm, or are merely epiphenomena of something else perceptible in the speech signal. This "something else" may not even be detectable with interval durations. Before we describe a methodology that is designed to overcome the limitations of interval-duration approaches, a few more issues should be pointed out here.

There may be a fundamental problem with the idea that syllable-timing or stress-timing is a property of a language, rather than of a particular utterance. If it can be shown that some utterances in a given language are more syllable-timed and others are more stress-timed, then perhaps the appropriateness of the cross-linguistic classification needs further reexamination. To show this, what is needed is a way of characterizing rhythm that does not rely on statistics of interval durations.

An even deeper problem with the stress/syllable/mora-timing trichotomy is that these so-called prosodic units may not be primitives with respect to speech rhythm. It is conceivable that the groups of speech gestures affiliated with a mora or a syllable or a stressed syllable (i.e. "head of a foot") may sometimes pattern abnormally and adopt timescales characteristic of the other units. For example, (phonologically/lexically) "unstressed" syllables may sometimes act more like stressed syllables, which in turn might sometimes pattern like morae. That is, rhythmic patterns might be dissociable from the prosodic units that are traditionally used to structure speech into hierarchical representations.

2. Method

2.1 Low-frequency spectral analysis

To understand how our method differs from interval duration methods, let's begin by considering all of the information about the speech signal that interval duration measurements ignore. Such measures represent the interval between two points in time with a single number, its duration. This effectively eliminates from the numeric representation of the signal all details about its amplitude envelope. From a naïve perspective, this seems like an egregious omission of detail, yet this dissociation of interval duration from the contents of an interval is so common that it is almost never explicitly noted in methodological appraisals.

The reason that such neglect is so widespread may be related to the very basic metaphors that structure the most popular linguistic theories. Consider a CVCV foot like [sasa], in which the intersyllabic interval (i.e. the duration between syllable onsets) is very clearly defined. Assume that the measurement of such an interval involves negligible error (approx. < 10ms). Our theoretical construct of the SYLLABLE, i.e. σ, encourages us to think of this period of time as a container. By metaphoric extension, this container can be empty (or at least its contents irrelevant), in which case the most pertinent information about the container is its size—the duration of the syllable. The prevalence of this conceptual metaphor, along with the ease with which such durations can be measured, encourages us to ignore the contents of the container. The same metaphor structures our understanding of interstress intervals, i.e. feet, which are understood to contain syllable containers.

The approach in this paper is to give much less attention to where intervals begin and end, and more attention to the acoustic contents of those intervals. We do this by analyzing the frequency spectrum of the slowly undulating amplitude envelope of speech. Consider now the segmented speech waveform in Fig. 1, where the citation form, phonetic transcription, and deletions are shown in tiers below. This utterance is about 2.3 s in duration, and contains 14 syllables.

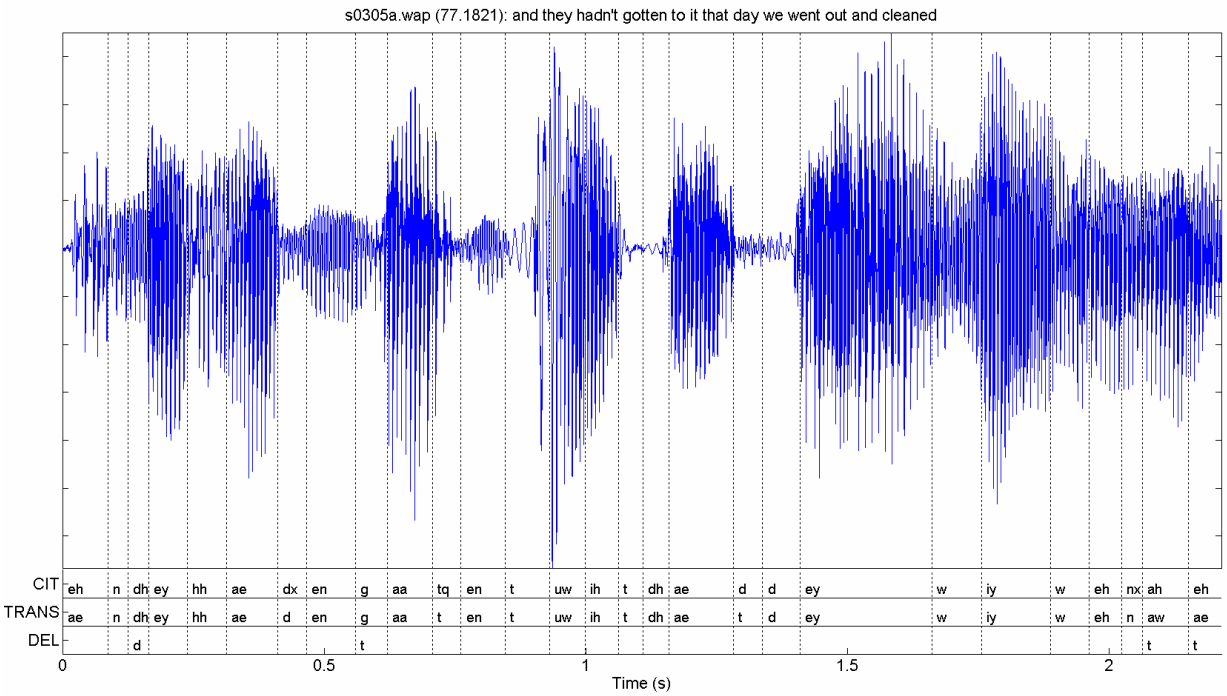

Fig. 1. Acoustic waveform and segmentation of a 2.3 s utterance from the VIC corpus.

One of the first things to notice about the speech signal is that in the long term, its integrated amplitude tends to zero (assuming a decent microphone), since the negative and positive pressure variations for voiced speech cancel each other out on small timescales. It is the magnitude (absolute value) of those variations that we are really interested in, because the magnitude varies more slowly and does not cancel itself out when integrated on small or large timescales.

Fig. 2 below shows the signal from Fig. 1 in panel (a) and its magnitude in panel (c). Panel (b) shows the same speech signal after it has been filtered using a 1st-order Butterworth filter with a passband of 700-1300 Hz, which is the same filter that Cummins & Port (1998) used to detect p-centers. The frequency-response of this filter decreases gradually from the cutoffs, so a fair amount of energy outside of the band remains in the signal after the filtering. The bandpass-filtered signal reflects primarily vocalic energy, because it filters out glottal energy and most of the noise occurring in obstruents, especially sibilant energy. However, it also responds better to low and back vowels than high-front vowels with low F1 and high F2; hence the bandpass-filtered signal does not represent the energy of all vowels equally. Panel (d) shows the magnitude of the bandpass-filtered signal.

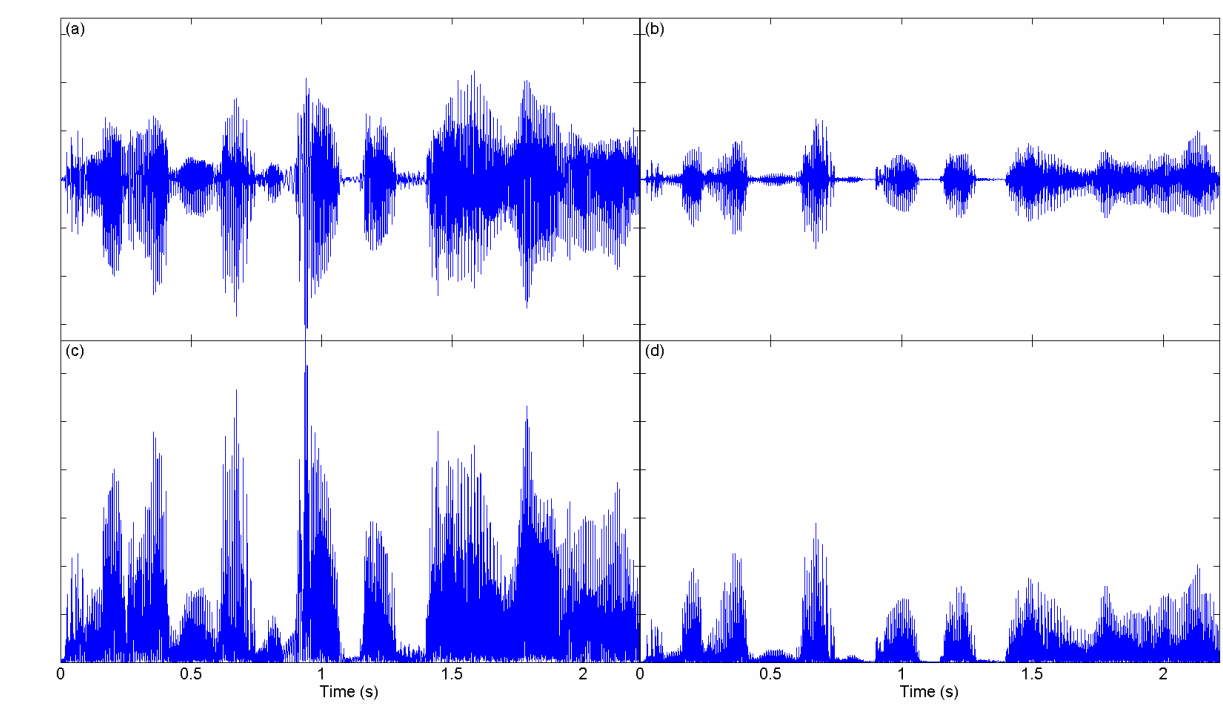

Fig. 2. Unfiltered (a) and bandpass-filtered (b) acoustic signals, along with corresponding magnitudes, (c) and (d).

To eliminate the rapid amplitude fluctuations of vocal fold vibration, we lowpass filter using a 4th-order Butterworth filter with a 10 Hz cutoff. The resulting signal is extremely redundant, so we downsample from 16000 Hz to 80 Hz to get rid of some—but not all—of the redundancy. This corresponds to an increase in sampling period from 0.0000625 s to 0.0125 s. Downsampling eliminates a fair amount of information, but plenty has been retained. Fig. 3(a) shows the downsampled lowpass-filtered magnitude (superimposed over the magnitude of the bandpass-filtered waveform) after a 45 ms correction has been made for the phase-delays of the filters (this correction is the sum of the mean phase delays of the bandpass filter in the 700 to 1300 Hz range and the lowpass filter in the 0 to 10 Hz range). Fig. 3(b) shows this same magnitude over the original waveform, after the magnitude has been windowed (using a Tukey window (r = 0.1) and the mean has been subtracted.

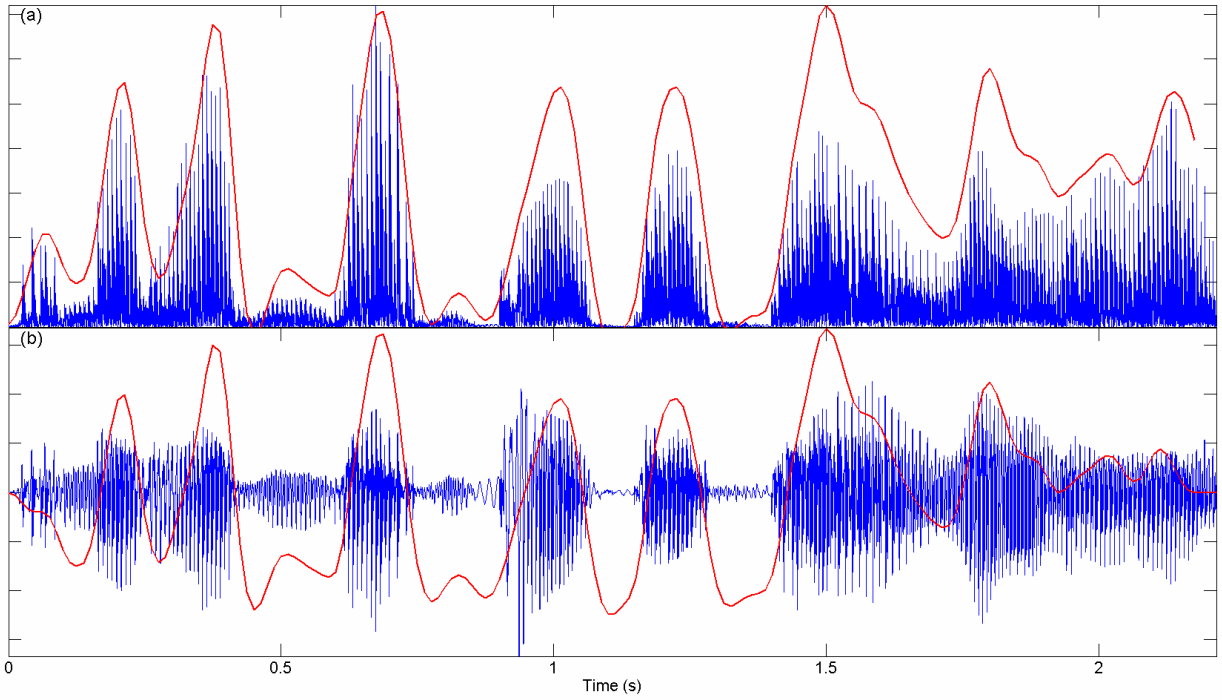

Fig. 3. Lowpass-filtered magnitude of bandpass-filtered signal superimposed over bandpass-filtered signal magnitude (a), and superimposed over original acoustic signal (b).

The low-pass filtered magnitude (or "processed magnitude") represents slow changes in the amplitude envelope of vocalic energy in the original waveform. The last step before the spectral analysis is performed is to pad the processed magnitude with zeros and normalize so that its variance is unity, as represented in equation (1).

(1)

$$\frac{1}{N}\sum_{t=1}^{N}(x_t - \overline{x})^2 = \sigma^2 = 1$$

Next we apply a Fourier transform (FT) to derive a frequency domain representation from the time-domain amplitude envelope signal. The FT preserves all of the same information that was contained in the processed magnitude, in the sense that the FT can be inverted to reconstruct the original signal. The Fourier transform is based upon the Fourier series (eq. 2), which can be used to approximate any real function.

for $\omega_j = 2\pi j/N$ and $t = 1...N$, the finite Fourier series is:

(2)

$$x_t = a_0 + \sum_{j=1}^{N/2-1} (a_j \cos(\omega_j t) + b_j \sin(\omega_j t)) + a_{N/2} \cos(\pi t)$$

where

$$a_0 = \overline{x}$$

and for $j = 1...N/2 - 1$

$$a_j = \frac{1}{N} \sum_{t=1}^{N} x_t \cos(\omega_j t)$$

$$b_j = \frac{1}{N} \sum_{t=1}^{N} x_t \sin(\omega_j t)$$

Re-expressing (eq. 2) in polar form (eq. 3) gives us a direct representation the phase and amplitude of sinusoidal components of the signal.

for $j \neq N/2$, the j^{th} harmonic in polar form:

(3)

$$a_j \cos(\omega_j t) + b_j \sin(\omega_j t) = R_j \cos(\omega_j t + \phi_j)$$

where

$$R_j = \sqrt{a_j^2 + b_j^2}$$

$$\phi_j = \tan^{-1}\left(\frac{-b_j}{a_j}\right)$$

One of the nice aspects of the information expressed by the FT is that it still bears a meaningful relation to our intuitive understanding of "rhythm"; arguably, it is even more relevant to measuring rhythm than interval durations are. One way to think about why the FT representation of the signal provides a good view of rhythm is to see it as the wisdom of the crowd. Each otherwise insignificant datapoint within all of the intervals in the entire signal contributes to the representation of the signal—as if polling a bunch of people has given us a more accurate idea of the overall inclinations across the population. Indeed, in profound contrast to the interval-based approaches, here no intervals whatsoever are defined, only frequency components with associated phases and amplitudes.

Note that the normalization to unit variance imposed upon the time series (eq. 1) is retained in the sum of Fourier amplitudes (eq. 4), a fact which follows from Parseval's Theorem (c.f. Chatfield 1975; Jenkins 1968; Anderson 1971). The Fourier Transform thus partitions the variance of the time series into components of differing amplitude at each of the Fourier analysis frequencies.

(4)

$$\sigma^2 = 1 = 2 \sum_{j=1}^{N/2-1} R_j^2 + R_{N/2}^2$$

The Fourier coefficients still contain far more information than we can use. For starters, we will discard information about the phase of each frequency component—not because this information is entirely irrelevant to rhythmic analysis, but because we assume that the phases of high-energy periodicities nearby in the frequency spectrum cluster together, and hence there is a fair degree of redundancy between the phase and amplitude information on local frequency-scales. The raw power spectrum, or periodogram, (eq. 5; Fig. 4) indicates how much each frequency component contributes to the waveform.

(5)

$$X(\omega_j) = \frac{N}{2\pi}(R_j^2/2)$$

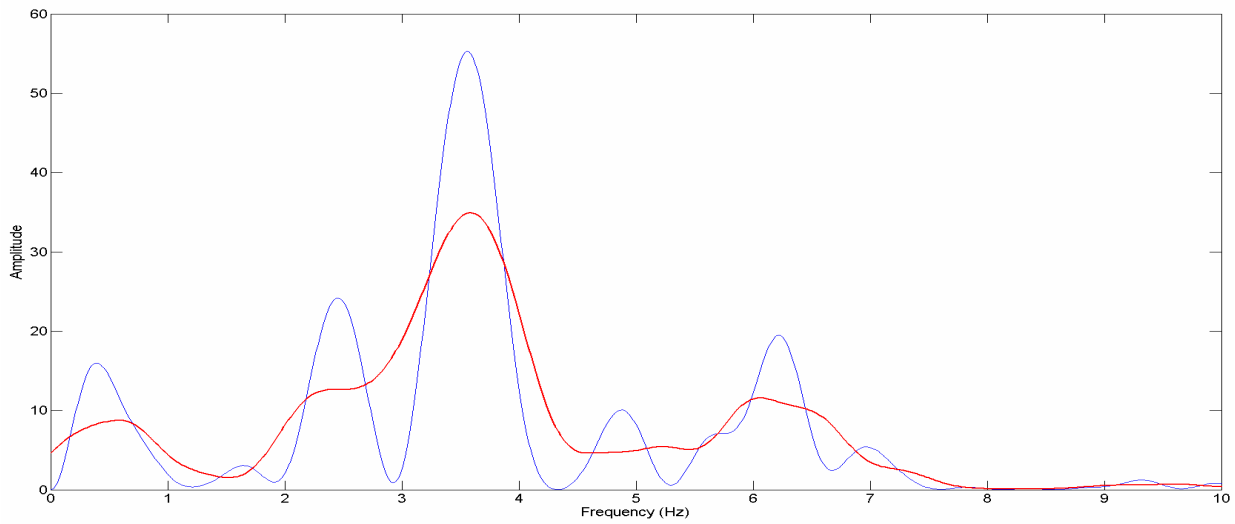

Fig. 4. Raw (blue) and smoothed (red) spectrum. L = 31 points. N = 2048.

The last step in this process is to smooth the power spectrum with bandwidth L, as shown by the red line in Fig. 4. Smoothing is accomplished using a moving average filter (eq. 6), and it is common to treat the spectrum as symmetric about 0 for this purpose (cf. Chatfield 1975, who describes the use of a Daniell filter for this purpose). The redundancy added by smoothing makes hypothesis testing easier, because the confidence intervals associated with smoothed spectra are narrower, owing to the fact that each smoothed amplitude value represents the average of values from a number of nearby frequencies. However, because of the nonlinear relation between frequency and period, this form of smoothing introduces greater spectral blurring for periodicities corresponding to longer intervals than for ones corresponding to shorter intervals—hence in some circumstances we will want to analyze the raw (unsmoothed) power spectra.

(6)

$$X_{sm}(\omega_j) = \frac{1}{L}\sum_{j-m}^{j+m} X(\omega_j)$$

where

$$m = (L-1)/2$$

It is instructive to compare the smoothed spectrum in Fig. 4 to the processed magnitude and waveform from Fig. 3(b), repeated in Fig. 5 below. The most dominant peak in the spectrum has a frequency of about 3.7 Hz, corresponding to a period of about 270 ms. Looking at just the seven highest peaks in the magnitude in Fig. 5, one can see that a number of them are separated by around 250-300 ms (3-4 Hz), which accords well with the spectral representation.

There are also smaller peaks at 6 Hz (167 ms), 2.2 Hz (450 ms), and 0.5 Hz (2 s). The presence of high-amplitude components at these frequencies implies that peaks in the magnitude signal should recur approximately at intervals corresponding to those frequencies. Fig. 5 additionally shows all of the intervals between all peaks in the signal.

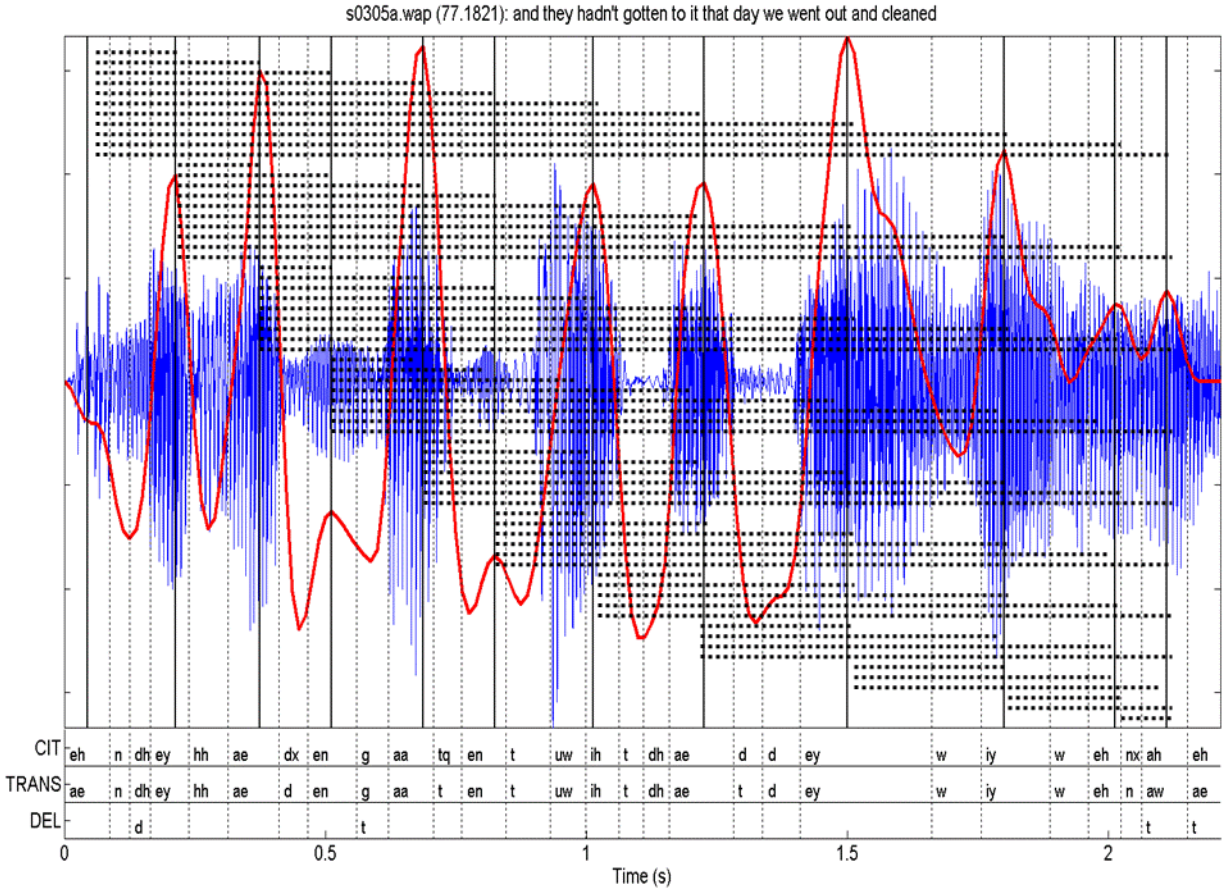

Fig. 5. Processed magnitude and waveform, along with all interpeak intervals.

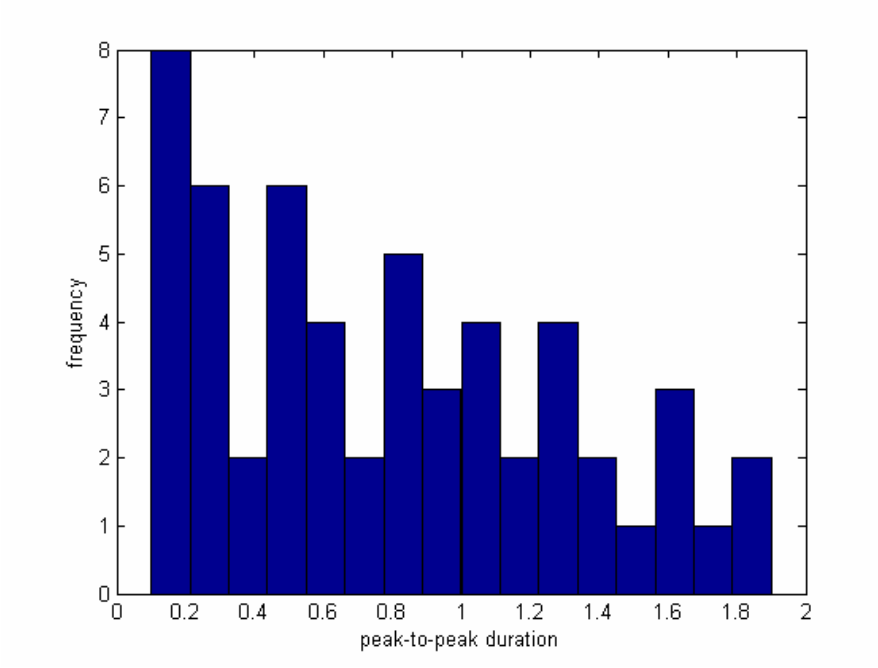

Fig. 6. Histogram of interpeak interval durations (from Fig. 5)

Fig. 6 shows a histogram of all the peak-to-peak intervals from Fig. 5. The highest concentration of durations is in the bins flanking 200 ms. There are also high concentrations in bins flanking 500 ms. These 200 and 500 ms interval duration concentrations are approximately what the spectrum predicts (i.e. 6-3.7 Hz ≈ 167-270 ms and 2.2 Hz ≈ 450 ms), but these correspondences are inexact because the interval duration approach does not take into account the amplitude of the peaks which define each interval.

Note that the duration of the signal defines a fundamental frequency—or, to avoid confusion, "minimal frequency"—which is the lowest frequency fitting within the signal period. Any peaks in the spectrum which occur below twice the minimal frequency of the signal, do not in any sense indicate the presence of a periodicity within the signal. Rather, these peaks indicate a global imbalance in the distribution of energy in the signal, which can be manifested as substantially louder speech in one part of the utterance, a lengthened filler, or various other forms of disfluency. In the example above, the signal period is 2.2 s, and so the minimal frequency is 0.45 Hz and spectral peaks below 0.9 Hz do not represent true periodicities.

2.2 VIC corpus chunks

For the current investigation, we are using speech from the Buckeye corpus (Pitt, Johnson, Hume, Kiesling, & Raymond 2005), which is a collection of approximately 300,000 words of conversational speech between interviewers and 40 native central Ohio English speakers. Three factors—age of speaker (over 40, under 40), gender of speaker, and gender of interviewer—were balanced across the interviews. The corpus was phonetically transcribed and segmented by transcribers trained to use acoustic and spectrographic information, following a number of conventions to ensure consistency.

To analyze the corpus, we first extract "chunks" of speech with no interruption or non-speech vocalization. In addition, the extraction procedure can separate chunks by all silences, or

a silence duration parameter can be set so that only silences greater than some duration cause chunks to be separated. There are a number of basic variables that are associated with each chunk, these include the duration of the chunk, the number of syllables (here considered equivalent to the number of vowels and syllabic consonants), the number of non-syllabic consonants, and the rate of syllables per second. Additionally, because the corpus is phonetically-segmented, it is possible to infer when segmental deletions and alternations have occurred, by comparing the citation form of a word to its phonetic transcription. Figs. 7-9 show the distributions of some of the basic chunk variables for raw chunks extracted from the corpus. Note that the mean rate of syllables per second is approximately 5 σ/s. Table 1 shows correlations between these basic chunk variables. As one might expect, the number of syllables in a chunk and its duration are highly correlated, and there is also a fair degree of correlation between rate and syllable count. Furthermore, there is some correlation between rate and chunk duration, which reflects the observation that very short chunks are likely to consist entirely of hesitations, fillers, or pitch-accented one-word utterances.

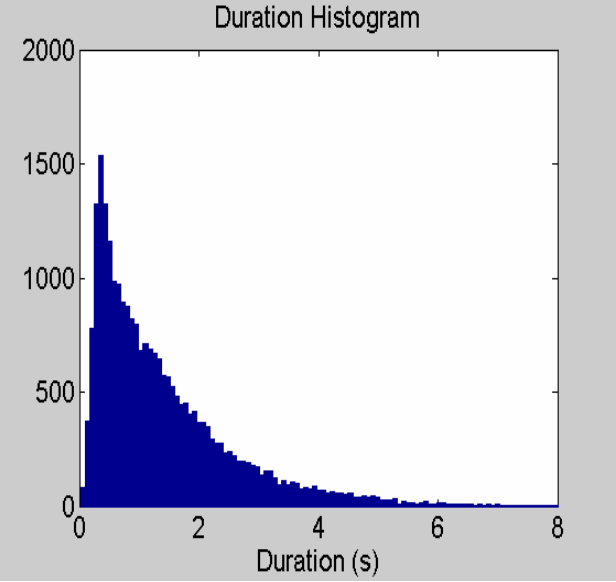

Fig. 7. Chunk durations

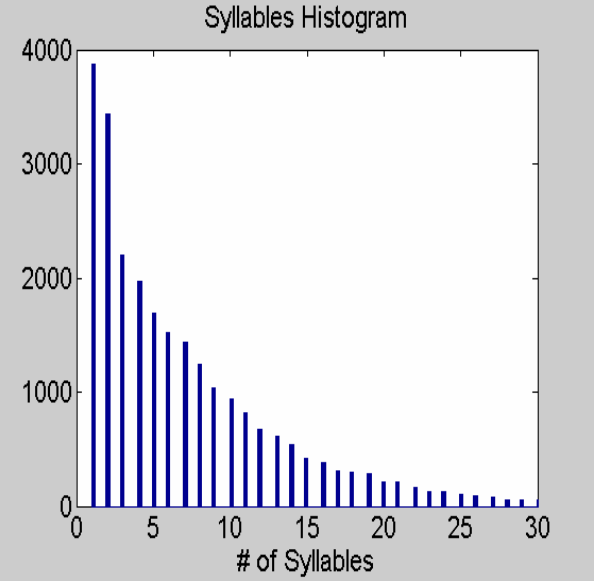

Fig. 8. Syllable counts