Low-frequency Fourier analysis of speech rhythm

Sam Tilsen and Keith Johnson
Department of Linguistics, University of California, Berkeley, 1203 Dwinelle Hall, Berkeley, California 94720
tilsen@berkeley.edu, keithjohnson@berkeley.edu

Abstract: A method for studying speech rhythm is presented, using Fourier analysis of the amplitude envelope of bandpass-filtered speech. Rather than quantifying rhythm with time-domain measurements of interval durations, a frequency-domain representation is used—the *rhythm spectrum*. This paper describes the method in detail, and discusses approaches to characterizing rhythm with low-frequency spectral information.

PACS numbers: 43.70.Jt, 43.70.Kv, 43.70.Fq [AL]
Date Received: January 30, 2008 **Date Accepted:** April 23, 2008

1. Introduction

The most successful methods of characterizing cross-linguistic differences in speech rhythm (syllable timing versus stress timing) use interval durations to describe the temporal patterns of speech (Cummins and Port, 1998; Dauer, 1983; Lehiste, 1977; Port *et al.*, 1987; Ramus *et al.*, 1999; Roach, 1982). In Pike's (1945) and Abercrombie's (1967) approach, speech rhythm is defined in terms of the intervals between the onsets of linguistic units—syllables, moras, or feet. The failure to find regularity in these interval durations (Bolinger, 1968; Lehiste, 1977) led to a reconsideration of speech rhythm in terms of temporal properties of consonantal and vocalic intervals (Dauer, 1983; Ramus *et al.*, 1999).

Investigations of the beat of a syllable (Allen, 1972, 1975) and its perceptual moment of occurrence (Morton *et al.*, 1976; Howell, 1988; Pompino-Marschall, 1989) have revealed that speech rhythm (defined as the perceived interval between beats) is influenced by characteristics of the amplitude envelope of energy between beat locations. This observation leads one to consider whether the acoustically defined intervals used in prior tests of the isochrony hypothesis (Nakatani *et al.*, 1981) are perceptually relevant. This concern is heightened given that pitch accent placement in ordinary English discourse (e.g., Ladd, 1996) does not give intonational prominence to every stressed syllable. That is, if rhythm in a "stress-timed" language is sometimes governed by timing between intonationally prominent stressed syllables, leaving out lexically "stressed" but nonaccented syllables, then attempts to find isochrony may have failed because they made a false assumption about the units of timing.

This paper describes the use of a spectro-temporal method of rhythmic analysis that makes no prior assumptions about the rhythms that should be found or the linguistic units that might define beats for any particular stretch of speech. Our method finds that while some utterances in English do exhibit stress-based rhythm, others have a clear syllable-based rhythm, and still others exhibit more regular intervals on a phrasal time scale, i.e., between pitch-accented syllables.

2. Method

Duration measurements represent an interval of speech with a single number, thereby neglecting information about the amplitude envelope of the speech signal. From a naive perspective, this omission might seem odd, but it is so common that it is almost never explicitly noted in methodological appraisals. One culprit for this may be the metaphor in which linguistic units are containers. This metaphor structures our theoretical constructs of the syllable and metric foot, encouraging us to reason about them in some of the same ways we reason about containers. Specifically, in some circumstances, the contents of containers are irrelevant and it is their *sizes* which are important. In many approaches to characterizing rhythm, the duration of a syllable or foot (or intersyllabic or interstress interval) is analogous to the size of a container, and

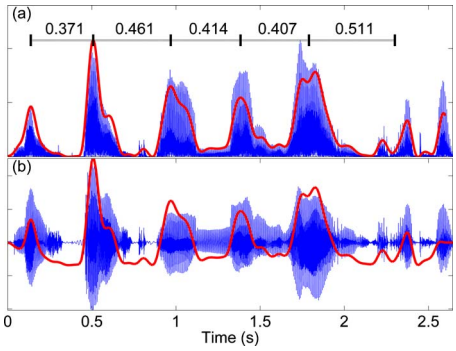

Fig. 1. (Color online) (a) Amplitude envelope superimposed over magnitude of bandpass-filtered signal; intervals between the several most prominent peaks in the amplitude envelope are shown. (b) windowed, mean-subtracted amplitude envelope over original acoustic signal.

its contents are either considered irrelevant, highly abstracted (e.g., labeled *vocalic* or *consonantal*), or thought to consist of other containers (i.e. syllables "within" feet, moras within syllables). Our approach here 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 power spectrum of the slowly undulating amplitude envelope of speech.

To illustrate, we use a 2.6 s stretch of speech, in which a male speaker says "at least based on money raised it looks like…" (to listen, click on the link to Mm1 below). First, to capture mainly vocalic energy and filter out glottal energy and obstruent noise, we apply a first-order Butterworth filter with a passband of 700–1300 Hz. Note that this filter has been used to detect *p*-centers, which are salient moments near the onsets of vowels (Cummins and Port 1998). Next we lowpass filter the magnitude of the signal using a fourth-order Butterworth filter with a 10 Hz cutoff, downsample to 80 Hz, and apply a correction for the phase delays of the filters (45 ms, i.e., the sum of the mean phase delays of the filters in their passbands). The resulting signal represents slow changes in vocalic energy. Figure 1(a) shows the lowpass-filtered magnitude of vocalic energy (henceforth *amplitude envelope*) superimposed over the magnitude of the bandpass-filtered waveform. Next we window the amplitude envelope using a Tukey window ($r=0.1$) and subtract the mean, as shown in Fig. 1(b) superimposed over the original waveform. Before performing the spectral analysis, we zero-pad the amplitude envelope to produce a 2048-sample window, and then we normalize to unit variance.

To derive a frequency-domain representation from the time-domain amplitude envelope, we apply a Fourier transform, which partitions the variance of the time series into components of differing amplitude at each of N Fourier analysis frequencies, where N is the number of samples in the zero-padded amplitude envelope. The normalization to unit variance imposed upon the envelope is retained in the sum of the magnitude of the Fourier coefficients, a fact which follows from Parseval's Theorem (cf. Chatfield, 1975; Jenkins and Watts, 1968). We then analyze the power spectrum (the squared magnitude of the complex Fourier coefficients), which shows the contribution of each frequency component to the amplitude envelope.

The power spectrum of the amplitude envelope is arguably more appropriate for measuring rhythm than interval durations are. The spectral representation derives from a sort of wisdom of the crowd: each otherwise insignificant datapoint within all of the intervals in the entire signal contributes to the spectral 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 interval-based approaches, here no intervals whatsoever need be defined, only frequency components with corresponding phases and amplitudes.

Mm. 1. A stretch of speech in which a female speaker says "at least based on money raised it looks like…." This is a "wav" file (83 Kb).

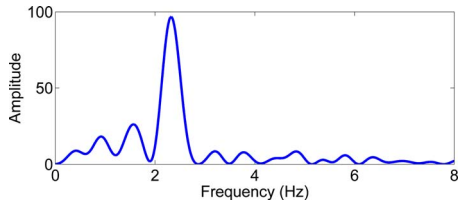

Fig. 2. (Color online) Power spectrum of the amplitude envelope that was shown in Fig. 1(b).

To relate the spectrum in Fig. 2 to the amplitude envelope in Fig. 1, observe that the average duration of the several most prominent peak-to-peak intervals in the amplitude envelope is about 430 ms, and as would be expected, there is a corresponding peak in the spectrum at approximately 2.3 Hz. Note that the duration of a chunk of speech defines a minimal frequency corresponding to the lowest frequency sinusoid that can fit within that duration. Any spectral peaks occurring below twice the minimal frequency do not reflect the presence of a periodicity within the signal; rather, these peaks indicate an imbalance in the distribution of energy in the signal, which can be manifested as substantially louder speech in one part of the utterance, perhaps arising from focus accent, lengthened fillers, etc.

3. Corpus analysis of rhythm in conversational speech

For current purposes, we are using speech from the Buckeye corpus (Pitt *et al.*, 2005), which is a collection of approximately 300 000 words of conversational speech between interviewers and 40 native central Ohio English speakers from a balanced set of ages and genders. 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 nonspeech vocalization. Basic variables associated with each chunk include: chunk duration, syllable count, and speech rate (syllables per second).

For illustrative purposes in this report, we have analyzed chunks in a duration range of $\tau=[2,3]$ s, because we suspect that this range is useful for studying syllable- and foot-timed rhythms. In general, the choice of subset duration range depends upon the time scales of the rhythms being investigated; longer chunk durations are more appropriate for studying rhythms on phrasal time scales. For the present analysis, we divided chunks longer than 3 s into smaller chunks in the desired range, randomly perturbing their durations to provide a more uniform distribution of durations over the $[2,3]$ s range. Figure 3(a) shows the waveform and amplitude envelope of a chunk of speech with a high-amplitude periodicity near 4 Hz in the rhythm spectrum; also shown are citation, transcription, and deleted phones. In this chunk the speaker says "…category of Forrest Gump because Forrest Gump was great guy" (to listen click on the link below). Figure 3(b) shows the power spectrum of this chunk compared to the mean and 2.5 standard deviation region (shaded) for all 2–3 s chunk spectra. The vertical line represents twice the lowest frequency corresponding to the duration of the chunk—all frequencies lower than this correspond to less than two cycles in the amplitude envelope. Larger chunk durations should be used for analyses of lower-frequency, phrasal rhythms; however, larger chunks introduce more variability on syllabic time scales and thus tend to blur the rhythm spectrum at higher frequencies.

Table 1 gives an indication of how often rhythmic speech occurs in the dataset by showing the percentages of chunks with a spectral peak exceeding 50 amplitude units for each of several frequency ranges. These data indicate that (>1 Hz) high-amplitude periodicity occurs in approximately 23.2% of the 2–3 s chunks in the corpus. The presence of periodicity in a variety of frequency ranges shows that speech is rhythmic on stress and syllabic time scales. Analyses conducted with longer duration chunks (not shown) have revealed phrasal (0.33–1 Hz) rhythms as well.

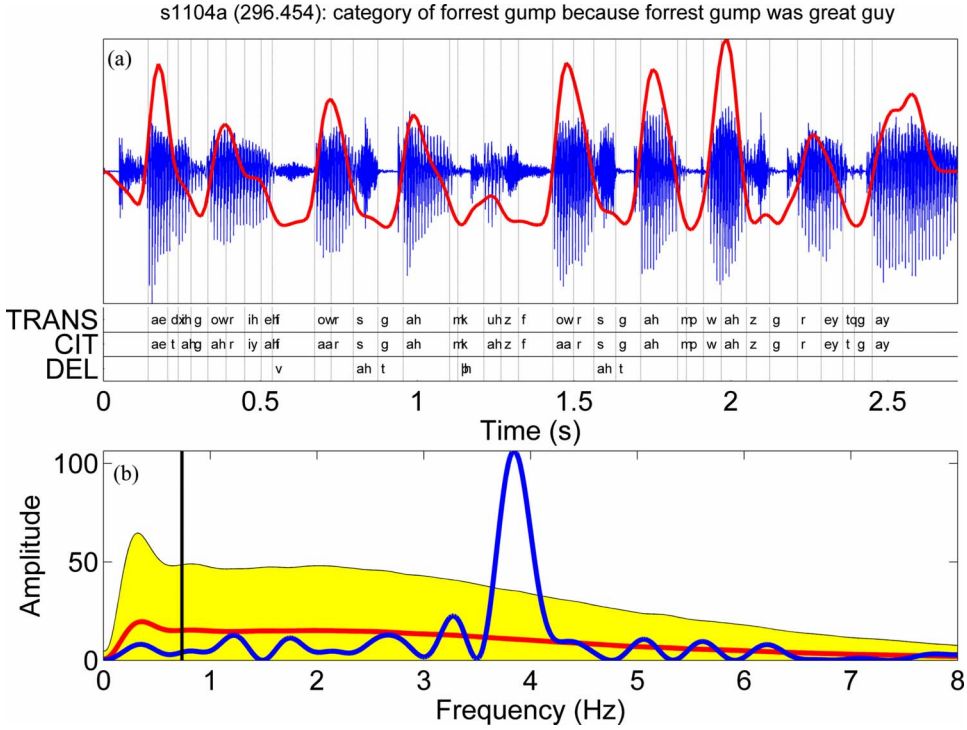

Fig. 3. (Color online) (a) Waveform and amplitude envelope of a chunk of speech, along with citation, transcription, and deleted phones; (b) power spectrum (peaked line) compared to average power spectrum (relatively flat line) and ±2.5 s.d. region for the set of 2–3 s chunks (shaded). Vertical line at twice the minimal frequency corresponding to the duration of the chunk.

Mm. 2. A stretch of speech in which a male speakers says "…category of Forrest Gump because Forrest Gump was great guy." This is a "wav" file (86 Kb).

We visualize the variability in a set of spectra by examining the distribution of peak frequencies and amplitudes, as in Fig. 4(a). For each spectrum, we locate one or more (but in this case one) of the highest peaks within a range of frequencies and then construct a two-dimensional Gaussian kernel density plot. To illustrate an appropriate level of detail, we use an amplitude range from the 0.1 percentile to the 99.9 percentile of amplitude values, an amplitude bandwidth of 5% of this range, and a frequency kernel bandwidth of 0.25 Hz. The most common low-frequency peak in this dataset is at about 1.6 Hz (i.e., a period of 625 ms).

Density plots also offer a useful way to compare datasets by inspecting the difference between density matrices. Figures 4(b) and 4(c) show peak frequency/amplitude density differences between subsets of data consisting of chunks with and without consonant and vowel deletions, where speech rate has been controlled by excluding chunks further than 1 s.d. from the mean speech rate. Deletions are identified by comparing the phonetic transcriptions with citation forms; deletions that occurred less than 80% of the time in their respective words were excluded in order to avoid artifacts due to overly specified citation forms. Rhythms tending to

Table 1. Counts of rhythmic chunks in several frequency ranges.

Rhythmic chunks	0–1 Hz	1–2 Hz	2–3 Hz	3–4 Hz	4–5 Hz	5–6 Hz	Total (>1 Hz)
Count	1354	897	871	396	109	27	2303
Percent	13.7%	9.1%	8.8%	4.0%	1.1%	<0.1%	23.2%

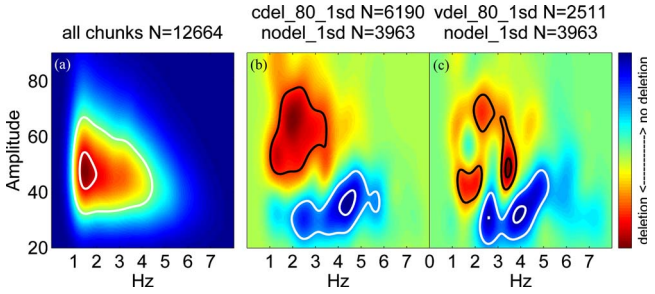

Fig. 4. (Color online) (a) Peak frequency/amplitude densities of 2–3 s chunk dataset where two highest peaks above twice the minimal frequency were taken from each spectrum; darkness corresponds to density and 50% and 90% contours are shown. (b, c) Density difference plots comparing rate-controlled subsets with and without consonant and vowel deletions; 90% and 50% positive and negative density contours are shown. (The information in this figure may not be properly conveyed in black and white.)

occur with more deletions are circled with dark 50% and 90% contour lines, and rhythms tending to occur without deletions are encircled with light contour lines. These figures indicate that very high-amplitude rhythms around 1–2 Hz are associated with consonant deletions, while rhythms around 3–4 Hz are more associated with vowel deletions. The predominance of the absence of deletion at lower amplitude periodicities indicates that when speech is less rhythmic, especially in the 2–3 Hz and 3.5–5 Hz ranges, deletion is less likely. Hence the data show a positive correlation between deletion and speech rhythmicity. Further, consonant and vowel deletions are most strongly correlated with highly rhythmic speech at different frequencies.

4. Conclusion and future directions

This report has presented a method for the quantitative analysis of rhythm that does not rely on interval durations, but rather, uses spectral analysis of the amplitude envelope of vocalic energy in speech. We believe that this "rhythm spectrum" analysis has the potential to augment studies of speech rhythm in a variety of ways. It offers a new approach to cross-linguistic rhythmic typology that involves statistical comparisons between large corpora of conversational speech. It can offer insights into rhythmic styles and characterizations of fluency from sociolinguistic and clinical perspectives. It may also shed light on relations between speech rhythm and intergestural timing, providing a deeper understanding of variation in conversational speech.

References and links

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine, Chicago).
Allen, G. D. (**1972**). "The location of rhythmic stress beats in English: An experimental study, parts I and II." Lang Speech **15**, 72–100,179–195.
Allen, G. D. (**1975**). "Speech rhythm: Its relation to performance and articulatory timing," J. Phonetics **3**, 75–86.
Bolinger, D. (**1968**). *Aspects of Language* (Harcourt, Brace, and World, New York.)
Chatfield, C. (**1975**). *The Analysis of Time Series* (Chapman and Hall, London).
Cummins, F., and Port, R. (**1998**). "Rhythmic constraints on stress timing in English," J. Phonetics **26**, 145–171.
Dauer, R. M. (**1983**). "Stress-timing and syllable-timing reanalyzed," J. Phonetics **11**, 51–62.
Howell, P. (**1988**). "Prediction of P-center location from the distribution of energy in the amplitude envelope," Percept. Psychophys. **43**(1), 90–93.
Jenkins, G. M., and Watts, D. G. (**1968**). *Spectral Analysis and Its Applications* (Holden-Day, San Francisco).
Ladd, D. R. (**1996**). *Intonational Phonology* (Cambridge Studies in Linguistics 79) (Cambridge University Press, Cambridge).
Lehiste, I. (**1977**). "Isochrony reconsidered," J. Phonetics **5**(3), 253–263.
Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual Centers (P-centers)," Psychol. Rev. **83**(5), 405–408.
Nakatani, L. H., O'Connor, K. D., and Aston, C. H. (**1981**). "Prosodic aspects of American English speech rhythm," Phonetica **38**(1–3), 84–106.
Pike, K. L. (**1945**). *The Intonation of American English* (University of Michigan Press, Ann Arbor).
Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (**2005**). "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun. **45**(1), 89–95.
Pompino-Marschall, B. (**1989**). "On the psychoacoustic nature of the P-center phenomenon," J. Phonetics **17**, 175–192.

Port, R. F., Dalby, J., and O'Dell, M. (**1987**). "Evidence for mora-timing in Japanese," J. Acoust. Soc. Am. **81**(5), 1574–1585.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**, 265–292.
Roach, P. (**1982**). "On the distinction between 'stress-timed' and 'syllable-timed' languages," in D. Crystal, *Linguistic Controversies* (Arnold, London).