Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages

Sam Tilsen[a)]
Department of Linguistics, Cornell University, 203 Morrill Hall, Ithaca, New York 14853-4701

Amalia Arvaniti[b)]
Department of Linguistics, University of California, San Diego, 9500 Gilman Drive, Number 0108, La Jolla, California 92093-0108

(Received 30 May 2012; revised 20 December 2012; accepted 29 April 2013)

This study presents a method for analyzing speech rhythm using empirical mode decomposition of the speech amplitude envelope, which allows for extraction and quantification of syllabic- and supra-syllabic time-scale components of the envelope. The method of empirical mode decomposition of a vocalic energy amplitude envelope is illustrated in detail, and several types of rhythm metrics derived from this method are presented. Spontaneous speech extracted from the Buckeye Corpus is used to assess the effect of utterance length on metrics, and it is shown how metrics representing variability in the supra-syllabic time-scale components of the envelope can be used to identify stretches of speech with targeted rhythmic characteristics. Furthermore, the envelope-based metrics are used to characterize cross-linguistic differences in speech rhythm in the UC San Diego Speech Lab corpus of English, German, Greek, Italian, Korean, and Spanish speech elicited in read sentences, read passages, and spontaneous speech. The envelope-based metrics exhibit significant effects of language and elicitation method that argue for a nuanced view of cross-linguistic rhythm patterns.
 [http://dx.doi.org/10.1121/1.4807565]

PACS number(s): 43.72.Ar [AL] Pages: 628–639

I. INTRODUCTION

The measurement of speech rhythm has been the focus of a great deal of research, but after decades of work, there is no consensus on how rhythm is encoded in speech and how it should be measured. Early efforts to measure rhythm in speech were driven by the assumption that rhythm is encoded in duration. More specifically these early efforts were driven by the intuition that languages differ in whether speakers attempt to coordinate the timing of only stressed syllables—stress-timing—or of all syllables—syllable-timing (Lloyd James, 1940; Pike, 1945; Abercrombie, 1967). However, predictions of isochrony of inter-stress intervals in stress-timed languages and syllable durations in syllable-timed languages have not been supported by empirical investigations that examined the durational characteristics of these units in a variety of languages (see Bertinetto, 1989; Kohler, 2009a,b; Arvaniti, 2012 for reviews).

Moving away from this view, Dauer (1987) attributed perceived rhythmic differences between languages to a mixture of phonetic and phonological variables—such as the relation between tone and stress and the effect of stress on syllable duration—that in her view made stressed syllables more or less salient, thereby shifting the emphasis from the measuring of duration to a parametric view of stress salience. Two of the parameters proposed by Dauer were selected by Ramus, Nespor, and Mehler (1999) as settings for stress- and syllable-timed languages: A typical stress-timed language would exhibit vowel reduction and would have a high degree of phonotactic complexity, i.e., onset and coda consonant clusters; in contrast, a typical syllable-timed language would not exhibit vowel reduction and would have relatively simple syllable structure. Building on this idea, Ramus *et al.* (1999) took a new approach to measuring rhythm that analyzed the variability of consonantal and vocalic interval durations and their relative proportions (thus shifting the emphasis back to the measuring of duration). The approach of Ramus *et al.* sparked a proliferation of related measures derived from the durations of (linguistically defined) units in the speech stream (e.g., Grabe and Low, 2002; Dellwo, 2006; White and Mattys, 2007; Nolan and Asu, 2009)—we refer to these as "interval-based" measures. However, instead of agreeing on the classification of languages into rhythm classes (the primary motivation behind interval-based metrics), the variety of metrics that have been proposed exhibit substantial disagreement due to data collection methods (e.g., Arvaniti, 2012), inter-speaker variability (Wiget *et al.*, 2010; Arvaniti, 2012), and the syllabic composition of utterances (Wiget *et al.*, 2010; Arvaniti, 2012; Prieto *et al.*, 2012).

An alternative class of approaches to characterizing rhythm are event based. One such approach utilizes an auditory primal sketch (Lee and Todd, 2004) in which an auditory representation of the acoustic signal is used to segment

a)Author to whom correspondence should be addressed. Electronic mail: tilsen@cornell.edu
b)Current address: Department of English Language and Linguistics, University of Kent, Cornwallis North West, Canterbury, Kent CT2 7NF, UK.

 0001-4966/2013/134(1)/628/12/$30.00

the signal and assign prominence to acoustic events. The prominence of these events was found to be more variable in read sentences of English than in French, suggesting that stress-timed languages may generally exhibit greater variance in the auditory prominence of phonetic events. Cummins and Port (1998) used p-centers—beat-like events associated with syllables—to characterize rhythmic timing in a phrase repetition paradigm. P-centers were first described as production centers associated with finger taps to stressed syllables (Allen, 1972, 1975) and subsequently redefined as perceptual moments of occurrence (Morton *et al.*, 1976; Pompino-Marschall, 1989). Notably, the temporal location of p-centers can be approximated by the rise of energy in the envelope of the speech signal (Howell, 1988), and amplitude envelope onsets have been argued to be important for rhythm perception (Goswami *et al.*, 2002).

However, recent work suggests that rhythm is a complex multidimensional percept that is not amenable to a simple analysis along a particular dimension of the speech signal (Kohler, 2008; Cummins, 2009). Focusing on interval durations, timing of events, or event prominences alone may be misleading as percepts of these dimensions have been shown to interact with other variables in the speech stream (e.g., Kohler, 2008, on the interactions of amplitude and duration and their effects on the percept of prominence; Yu, 2010, on the effects of F0 modulation on percepts of duration).

To redress this problem, here two alternative approaches to rhythm measurement are presented: Envelope spectral analysis and envelope empirical mode decomposition. Both methods involve analysis of amplitude envelopes derived from filtered speech waveforms, and hence we refer to these as "envelope-based" measures. In envelope analysis, rhythm is conceptualized as periodicity in the envelope, and greater stability of that periodicity corresponds to greater rhythmicity. Hence all utterances exhibit rhythmicity to a greater or lesser degree.

The envelope normally exhibits relatively slow fluctuations of acoustic signal energy that tend to arise from alternations between vowels and consonants but do not correspond precisely to vocalic and consonantal intervals. This observation as well as the aforementioned research on automatic location of p-centers motivated the development of an envelope-based approach to characterizing speech rhythm, which utilized a spectral analysis of the envelope of the speech waveform (Tilsen and Johnson, 2008; Tilsen, 2008).

The current study extends envelope spectrum analysis and presents a new, time-domain method of envelope analysis, which is based upon a technique known as empirical mode decomposition (EMD, Huang *et al.*, 1998). This procedure extracts non-orthogonal basis functions from the envelope, each of which captures oscillations on a different time-scale. The time-scales are empirically determined in that they depend on the frequencies of the oscillations present in the signal itself. Exploratory use of this technique revealed that—with appropriate filtering of the envelope—the first two functions obtained from EMD reflect primarily syllable- and stress-time-scale periodicities in the speech envelope. They are thus quite useful in characterizing rhythmic properties of stretches of speech. A benefit of this approach is that it provides additional dimensions for characterizing utterance rhythmicity. These dimensions include the frequencies of the syllable- and stress-time-scale components, the stability of their oscillations, and their relative contributions to the envelope, all of which in our review represents different facets of rhythm in speech.

The goals of this study are (1) to illustrate the methods along with novel metrics of rhythm that can be derived from them, (2) to assess the effect of utterance length on these metrics, and (3) to demonstrate their utility in two applications: Identification of rhythmic stretches of speech from corpora and comparison of rhythmic patterns across languages. Section II describes the analysis methods in detail. Section III reports analyses applied to the Buckeye corpus of spontaneous American English speech (Kiesling *et al.*, 2006; Pitt *et al.*, 2005) and the cross-linguistic corpus collected at the UC San Diego Speech Lab (henceforth SLab, see Arvaniti, 2012 for further details). The Buckeye corpus is used to characterize the influence of chunk duration on the statistics of various envelope-based rhythm metrics and to demonstrate how they can be used to identify highly rhythmic stretches of speech. The SLab corpus is used to evaluate whether envelope-based metrics can reveal rhythmic differences between languages and elicitation methods. It is shown that although envelope metrics exhibit cross-linguistic differences and are affected by elicitation method, they were not dependent on syllable structure. Section IV discusses the results and potential applications of envelope-based rhythm analysis.

II. METHODS

A. Chunk extraction

Prior to envelope-based analysis, stretches of speech must be selected for the analysis. The duration of these stretches of speech is an important consideration. Very short stretches (e.g., with durations <1 s) are unlikely to yield informative results because they will not contain a sufficient number of syllables or stresses to provide rhythmic information. Very long stretches (e.g., with durations >3 s) are likewise inappropriate because they will often contain a mixture of tempos and a lot of variation in rhythmicity. Using a wide range of durations in a given analysis is problematic because the preceding issues will be confounded. Another consideration is that the stretches of speech should not contain pauses because these presumably interrupt rhythmicity rendering analyses less meaningful. To deal with these issues, the extraction algorithm proceeds through "phrases"–continuous stretches of speech without any substantial pause (a criterion of 100 ms is used when necessary)–and demarcates "chunks," i.e., stretches of speech of target duration, allowing them to overlap by 50%. This overlap increases the number of data points available and reduces the extent to which the statistics of analyses are dependent upon the precise locations of chunk boundaries. To further reduce this dependence, a random variability of ± 100 ms is imposed on the target duration of each extraction. Hence a dataset of 1500 ms chunks will contain a nearly uniform distribution of chunk durations from 1400 to 1600 ms.

B. Envelope extraction

Given a chunk of arbitrary size, the first step in envelope-based rhythm analysis is to extract an "envelope" from the speech waveform. Because our interest is in variation in signal amplitude due to alternation between vocalic nuclei and consonantal margins, the speech signal is bandpass-filtered (fourth-order Butterworth, [400, 4000] Hz) to accomplish two effects. The effect of the low-frequency cutoff is to de-emphasize the contribution of F0 and thereby decrease the extent to which the presence of voicing is directly represented in the signal. This is desirable, for example, to render voiced consonants more similar to the voiceless consonants and hence further differentiate them from vocalic nuclei, for which resonances are preserved in the passband. The high frequency cutoff serves to decrease the representation of sibilant consonants and bursts—this helps avoid associating these events with peaks in the envelope. The exact values of the passband cutoffs are by necessity somewhat arbitrary. The 400 Hz high-pass is not infallible: It may allow some extremely high F0 or low harmonics to be over-represented in the signal, and it somewhat privileges the contribution of low vowels with high F1 relative to high vowels with low F1. However, parametrically varied analyses of envelope-based metrics of the corpora indicate that this value comes close to minimizing variance across datasets. Note that the Butterworth filter used has no passband ripple but relatively gradual decrease in frequency response at the cutoff; this is consistent with our heuristic use of the passband range to extract a signal that represents mostly vocalic energy.

Figures 1(a) and 1(b) show the *vocalic energy* signal, which results from passband filtering the original waveform as discussed in the preceding text. The next step is to low-pass filter the magnitude of the vocalic energy, shown rescaled in Figs. 1(a) and 1(b). A fourth-order Butterworth filter with a 10 Hz cutoff is used, but once again, the precise value of the cutoff is somewhat arbitrary. The purpose of the low-pass filter is to obtain an envelope that varies on the time-scale of alternation between vocalic nuclei and consonantal margins, i.e., a syllable-time-scale. The 10 Hz cutoff implies that the duration of a syllable is expected to be no less than 100 ms (although the relatively gradual roll-off of the Butterworth filter frequency response admits higher frequency oscillations with somewhat reduced energy). The result of the low-pass filtering is shown in Figs. 1(a) and 1(b) and is termed the *vocalic energy amplitude envelope*, "envelope" for short. Note that this is not a true envelope in the sense of a function that passes through positive or negative extrema of the vocalic energy, but we have no reason to believe that this more restrictive sense of an envelope is necessary for current purposes.

Some additional processing is performed to make the signal more suitable for further analysis. Specifically, the amplitude envelope is normalized by subtracting the mean and subsequently rescaling the envelope by its maximum absolute value (so that it will have an extremum at 1 or −1). Next the envelope is downsampled by a factor of 100 to speed subsequent computations. Finally the envelope is windowed using a Tukey window ($r = 0.2$) to aid spectral analyses. A modest value of the Tukey parameter was chosen to avoid influencing a large time span of the signal and to avoid creating very rapid fluctuations near the edges. The resulting signal is the processed vocalic energy amplitude envelope.

C. Envelope spectral analysis

Figure 1 also shows envelope power spectra obtained from the example envelopes. The spectrum is calculated by taking the squared magnitude of the fast Fourier transform of a zero-padded processed envelope. To smooth and normalize the spectrum, it is divided by the length of the input to the fast Fourier transform and multiplied by a factor of two, then smoothed across negative and positive frequencies by averaging over a 1 Hz frequency band centered upon each frequency bin (Chatfield, 1975). This smoothing procedure results in non-zero spectral power at 0 Hz (evident in Fig. 2), but because our spectral-based analyses generally do not incorporate spectral information below 1 Hz, this is not problematic.

Two different approaches to quantifying the rhythm of a chunk based upon its envelope spectrum are presented here. One approach is to define relatively low and high frequency bands, integrate spectral power within those bands, and compute the ratio of the power in the lower band to the power in the higher band. The basis for this is the assumption that lower-frequency, longer time-scale periodicity in the envelope corresponds to the "supra-syllabic" influence of stress

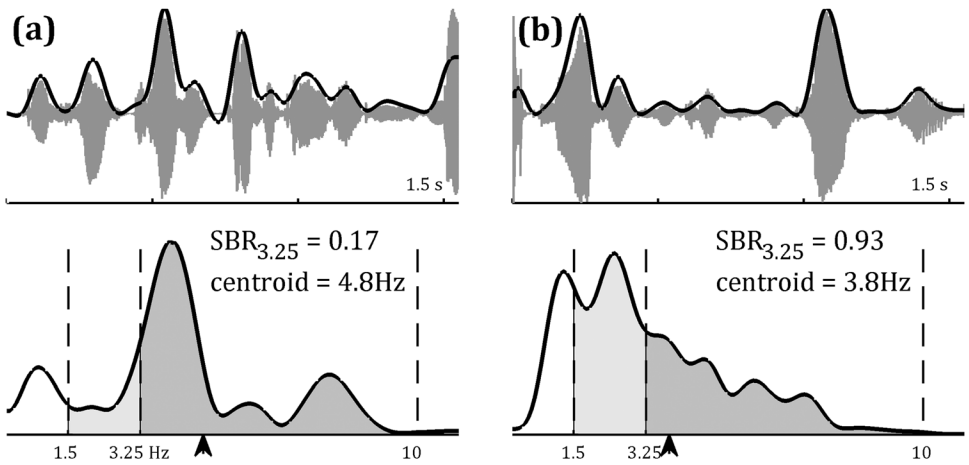

FIG. 1. Top: Example waveforms and vocalic energy amplitude envelopes; bottom: Power spectra corresponding to the example waveforms. Spectral bands used for calculating the spectral band power ratio are shaded.

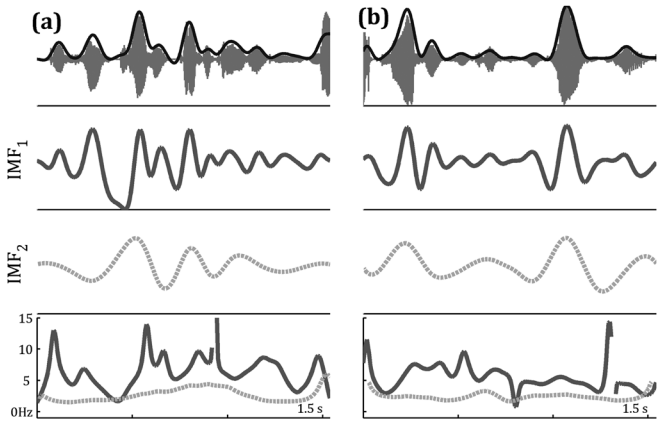

FIG. 2. Examples of intrinsic mode functions obtained from empirical mode decomposition of envelopes with corresponding instantaneous frequencies.

or feet on rhythm, whereas higher-frequency, shorter time-scale periodicity corresponds to an influence of syllables on rhythm. This ratio metric is referred to as the *spectral band power ratio* (SBPr). This approach is illustrated in Fig. 1, where the low- and high-frequency bands are shaded. Notice that the cutoff between the bands in this example is 3.25 Hz; this corresponds to a period of approximately 300 ms. The division between bands is somewhat arbitrary. Generally speaking, the duration of the *typical* syllable should be smaller than the period corresponding to this parameter. Because we expect that most syllables in fluent spontaneous speech will not exceed this value, it is useful in a practical sense. The low-frequency cutoff of the low-frequency band is 1.5 Hz (667 ms). This is derived from the presumption that the interval between stressed syllables (or the typical duration of feet) will not exceed this duration in spontaneous speech.

However, it is clear that the definition of spectral bands is necessarily arbitrary. To obtain a less parameter-dependent metric characterizing the spectrum, the spectral centroid over a range of 1.5-10 Hz was also analyzed. The centroid (or spectral center of gravity) is a weighted mean of frequencies. It is computed by summing all of the frequencies in the selected range multiplied by their associated spectral power and then dividing by the sum of all the spectral power over that range. Although this metric is not sensitive to an arbitrary division between relatively low and high frequencies, it remains sensitive to the low- and high-frequency cutoffs for the range of the spectrum over which it is calculated. In the example here, a 1.5 Hz cutoff was used (for reasons discussed in the preceding text), but in the analyses reported below, this low-frequency cutoff was varied to explore the consequences for analyses.

D. Empirical mode decomposition of the envelope

There are well known shortcomings of Fourier analysis involving non-stationary signals such as the vocalic energy amplitude envelope: By adding many harmonic frequency components to the analysis, such non-stationary signals can be analyzed for practical purposes, but the signal energy ends up being spread over a wide range of frequencies. This shortcoming motivated the development of a method for analyzing non-linear time series called *empirical mode decomposition* (Huang *et al.*, 1998) to facilitate instantaneous frequency analysis of signal components using a Hilbert transform. Like Fourier analysis, EMD decomposes a signal into a number of basis functions. Yet in EMD, these bases are not orthogonal; rather, they are empirically determined from the signal using an algorithmic sifting procedure. The empirically obtained basis functions have the property of having a well-defined instantaneous frequency: They have zero local mean and the same number of zero crossings and extrema (Huang *et al.*, 1998, p. 915)—which entails that all peaks and troughs will be separated by zero crossings, and the average of the negative and positive envelopes is zero at any point. Functions meeting these criteria are termed *intrinsic mode functions* (IMFs) by Huang *et al.*

The output of EMD is thus a set of IMFs, each of which represents an oscillation in the input signal but on time-scales that depend on the signal itself. The relation between the input and output of EMD is illustrated in Fig. 2, which shows the first two IMFs for an EMD of the chunks shown in Fig. 1 in the preceding text. The first intrinsic mode function (IMF_1) represents the fastest time-scale of oscillation in the envelope, IMF_2 represents the next fastest oscillation, and so on. The processed envelope can be reconstructed by adding together the IMFs. In principle, there is no limit to how many IMFs can be constructed, but because the signal is of finite duration and higher-order IMFs represent progressively slower oscillations, eventually there will be no power for higher-order IMFs to capture. Another important property of the IMFs is that they cross zero between all extrema; this not only makes them suitable for further instantaneous frequency analysis using a Hilbert transform but is also a criterion in the sifting process, which we illustrate in the following text.

The critical observation to make regarding the IMFs is that the first two IMFs generally reflect syllable- and stress-driven fluctuations in the envelope, respectively. We do not have an *a priori* proof of this assertion, but rather, have observed it to hold true in the majority of cases we have examined by inspection. These associations likely follow from two speech-related characteristics inherent to the envelope. First, having low-pass-filtered the signal at 10 Hz the fastest time-scale oscillations are expected to derive from alternation of vocalic nuclei and consonantal margins. Second, because stressed syllables can imbue the envelope with additional amplitude fluctuations, these additional fluctuations will be represented in an IMF, and because there is no linguistic source of envelope modulation that intervenes between the time-scale of the syllable and foot, the second IMF will represent these modulations. Furthermore, energy in higher IMFs appears to relate to less periodic linguistic sources of envelope energy, namely, phrasal prominence and other forms of emphasis. The observation of an association between syllables/stress and the first/second IMFs provides the basis for the rhythm metrics developed in the following text.

Figure 3 illustrates the decomposition method for the first IMF of the example envelope from Fig. 1(a). Each IMF is obtained through a series of sifting processes. The original

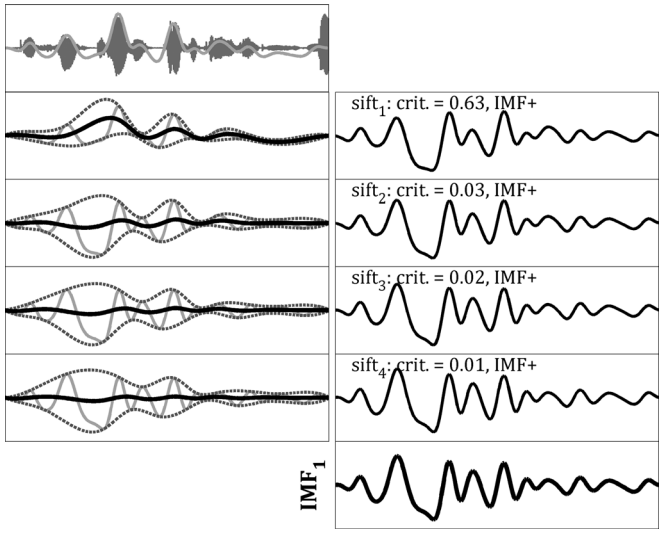

FIG. 3. Illustration of empirical mode decomposition sifting processes to obtain the first IMF of an amplitude envelope. Top left panel: Input waveform and envelope; middle left panels: Sifts (solid lines), sift-envelopes (dashed lines), and average of sift-envelopes (bold lines); middle right panels: New sifts, i.e. the difference between the preceding sift and average of sift-envelopes; bottom right panel: Output intrinsic mode function.

envelope can be considered a zero sift, i.e., sift_0. The local maxima and minima of this signal are connected by a cubic spline providing upper and lower envelopes (dashed lines). The first sift (sift_1) is then obtained by subtracting the mean of the envelopes (bold line) from the preceding sift (sift_0). Sift_1 is assessed according to two criteria: First, whether it meets the definition of an IMF, with regard to having a zero-crossing between all extrema, and second, a standard deviation criterion, which is the ratio of the sum of squared differences between it and the preceding sift to the sum of the preceding sift squared. If this value is not below a threshold, e.g., 0.01, the sifting process is repeated: Envelopes of sift_1 are obtained and their mean is subtracted from sift_1 to obtain sift_2. This new sift is once again assessed by the criteria, and the process is repeated until a sift meets both criteria at which point it is taken as the first IMF. The standard deviation criterion serves to preserve both amplitude and frequency modulations in the IMF because otherwise too many siftings would render it with constant amplitude (Huang *et al.*, 1998, p. 920). The first IMF will contain the shortest time-scale component of the envelope. Next, the first IMF is subtracted from the original data to obtain a residue, which will contain longer time-scale components of the envelope. A new round of siftings is performed on this residue until a second IMF is obtained. This second IMF is then subtracted from the preceding residue and another round of sifting is performed, and so on. Further IMFs can be computed as long as the preceding residue contains two local extrema, but if those are of relatively low amplitude, the resulting IMF is not very informative. For the original and more complete exposition of the method, the reader should consult Huang *et al.* (1998).

A key property of the resulting set of IMFs is that they can be added together, in combination with the final residue, to reconstruct the original envelope. It follows that the amplitudes of the IMFs thus have a consistent physical meaning with regard to the envelope: They contain energy from the envelope on a variety of time-scales inherent to the envelope itself. This property provides the basis for a new metric of rhythmicity, which characterizes the relative power in the first and second IMFs. It was observed that the first IMF generally contains syllable-time-scale oscillations and the second IMF contains stress/foot-time-scale oscillations. To quantify the relative power of these oscillations, we propose the intrinsic mode function power ratio (IMFr_{12}), which is simply the power of the second IMF (i.e., the sum of its squared values) divided by the power of the first IMF. The value of this metric expresses the contribution of lower-frequency, stress-related energy in the signal relative to higher-frequency syllable-related energy. It may be informative to analyze other ratios involving the third IMF (possibly related to phrasal energy), but this possibility is deferred for exploration in subsequent studies.

E. Instantaneous frequency of envelope components

Having obtained IMFs from EMD, the IMFs can be analyzed using a Hilbert transform to obtain an instantaneous phase, from which the instantaneous frequency (ω) can be defined as the time derivative of phase (Huang *et al.*, 1998, p. 912). It should be noted that the instantaneous frequency derived from the Hilbert transform in this way is not very similar conceptually to frequency in the Fourier domain, for a variety of reasons (Huang *et al.*, 1998, p. 930). Because IMFs do not have a constant frequency or amplitude, changes in instantaneous phase can be quite rapid, resulting in large jumps of instantaneous frequency. Furthermore, when an IMF oscillates around zero with very low amplitude, the instantaneous phase can exhibit discontinuities in instantaneous frequency. To mitigate against these effects, the phase is unwrapped where jumps occur, and each datapoint is smoothed by averaging over nearest neighbors; then, frequency values greater than 3 standard deviations from the mean are excluded from subsequent analyses. The first and last 100 ms of the instantaneous frequencies are also excluded to avoid the influence of window-related edge effects.

The instantaneous frequencies for the first two IMFs of the two example envelopes are shown in Fig. 2. The instantaneous frequency of the first IMF is generally quite variable, but this should be expected given the nonstationarity of the speech envelope. The instantaneous frequency of the second IMF typically changes more slowly, disregarding edge effects. Variability in the instantaneous frequency over time is diagnostic of the stationarity of an oscillation. Hence the variance of the instantaneous frequency from each IMF can be used to assess the stability of its frequency of oscillation. In Sect. III, these stability measures are used to identify highly rhythmic utterances in a spontaneous speech corpus.

Another possibility not explored here is that IMF instantaneous frequencies from a given chunk may be related via frequency locking. There are several models of speech production that have posited that utterances are organized by a hierarchy of harmonically related oscillatory cycles (O'Dell and Nieminen, 1999; Cummins and Port, 1998; Tilsen,

2009). For example, syllables may be associated with oscillations that exhibit a frequency that is an integer multiple of foot oscillations. There are numerous factors that may confound the detection of these harmonically organized oscillations in spontaneous speech, but in some utterances, such relations may be more or less evident. Future studies may profit from investigating the relation between the instantaneous frequencies of the IMFs.

F. Corpora

Two corpora were used in the analyses presented in the following text. The first is the freely available Buckeye corpus of spontaneous speech (Kiesling *et al.*, 2006; Pitt *et al.*, 2005), which contains approximately 300 000 words of conversational speech from 40 native central Ohio speakers, balanced for gender and age. The corpus is derived from interviews with everyday topics such as politics, sports, traffic, and schools. One speaker from this corpus was excluded from analyses presented in the following text due to issues with recording quality. The second corpus used here was developed by the UC San Diego Speech Laboratory specifically for research on rhythm. It is composed of speech from eight speakers from each of six languages (for a total of 96 min of speech evenly divided among the languages): English, German, Greek, Italian, Korean, and Spanish. Each speaker participated in three elicitation tasks: Reading the North Wind and the Sun story, reading a corpus of 15 sentences, and providing 1 min of spontaneous speech on a topic provided by the experimenters (see Arvaniti, 2012 for details). All 48 speakers from the San Diego corpus were included in the analyses presented in the following text.

III. RESULTS

A. Influence of chunk duration on metrics

This section reports the effects of chunk duration on envelope metrics in the Buckeye corpus. The parameters used for these analyses were described in Secs. II B to II E. Figure 4 shows the relation between chunk duration and the mean, standard deviation, and detectable effect size (assuming 80% power, expressed as a percentage of the mean) for each metric. The values are shown for both individual speakers (light lines) and the average across speakers (bold lines). In general, it can be seen that standard deviations of metrics decrease when longer chunks are used, presumably because longer chunks more closely approximate long time-scale trends. However, because using longer chunks results in fewer datapoints, the resulting statistical power diminishes in analyses based upon longer chunks. This can be observed from the increase in the minimum detectable effect size for each metric. Chunks longer than 2.6 s were not examined because further increases in chunk size drastically limit the number of datapoints available. To assess whether effects of chunk duration on metric values were significant, metric values were grouped by chunk duration into four bins: Short (1–1.3 s), short/moderate (1.4–1.7 s), moderate/long (1.8–2.1), and long (2.2–2.6 s). An analysis of variance (ANOVAs) with repeated measures over speaker means was conducted on each metric with the factor duration.

The first three metrics shown in Fig. 4 will be referred to as *power distribution metrics*, and are listed in Table I. These metrics represent the relative amount of power in the envelope on supra-syllabic vs syllabic time-scales. Two of the power distribution metrics are derived from the envelope spectrum. The spectral band power ratio (SBPr, Sec. II C) is shown in the leftmost column of Fig. 4. This metric represents the amount of spectral power in the 1.5–3 Hz band (333–667 ms time-scale) relative to power in the 3–10 Hz band (100–333 ms time-scale). The average value across speakers did not change significantly with chunk duration [$F(3,620) = 0.33$, $p = 0.81$], although within speakers there is duration-dependent variation attributable to specific chunk locations, i.e., the points at which chunks were extracted for a given chunk duration. Several speakers obtain quite high values in the range of 0.7–0.9, suggesting a predominance of utterances in which there is low-frequency, stress-time-scale periodicity in the envelope, whereas others fall in the 0.5–0.6 range, suggestive of a relative absence of utterances with stress-time-scale periodicity. Without further analysis,

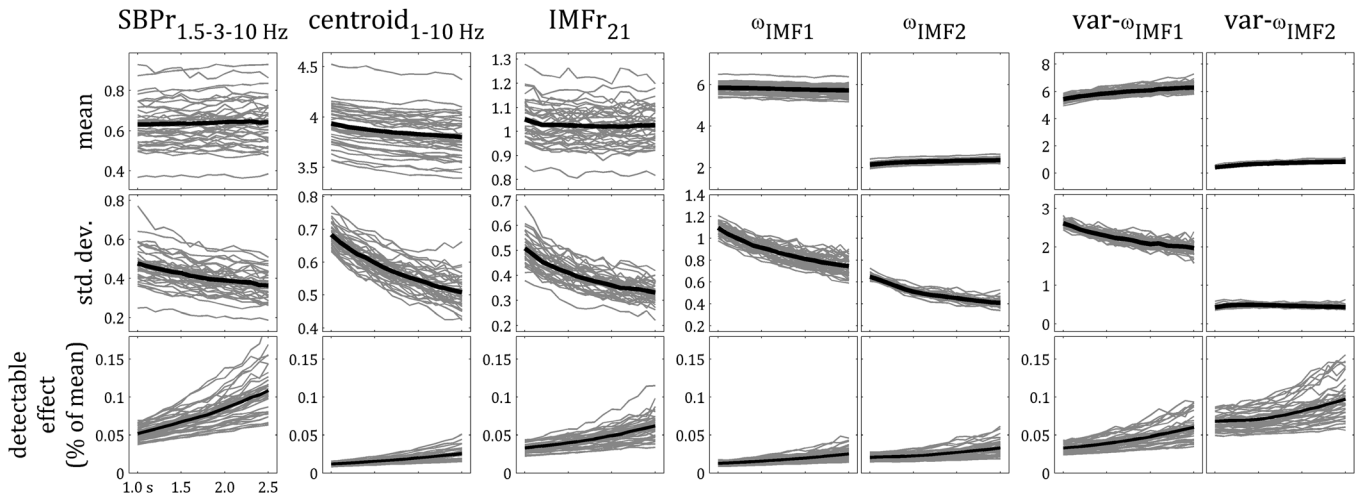

FIG. 4. Effects of chunk duration on envelope-based metrics: Values for individual speakers (light lines) and across-speaker means (bold lines). Top panels: Mean values; middle panels: Standard deviation; bottom panels: Detectable effect as a percentage of mean.

we can only speculate on the sources of these inter-speaker differences, which may arise from variation in tendencies to produce more rhythmic speech, phonetic manifestation of stress, or the prevalence of short, prosodically strong utterances. Further investigation can help elucidate such speaker specific differences in degree of rhythmicity in speech.

The other spectrum-derived measure shown in Fig. 4 is the envelope spectral centroid, calculated over 1–10 Hz. This measure exhibits a significant dependence on chunk duration [$F(3,620) = 7.2$, $p < 0.001$], likely because short chunks will necessarily contain fewer long time-scale periodicities and hence the distribution of spectral energy will be biased toward the upper end of the spectrum.

The remaining metrics shown in Fig. 4 are derived from the first two EMD intrinsic mode functions. The IMFr is a power distribution metric, corresponding to the ratio of power in IMF_2 to IMF_1; hence this represents the relative influence of lower frequency, stress-time-scale periodicity (Sec. II D). With the exception of the shortest chunk durations, the mean IMFr across subjects is fairly stable: chunk duration did not have a significant effect on the mean value of this metric [$F(3,620) = 0.99$, $p = 0.40$]. Within subjects. it can vary to some extent depending upon chunking locations, much like the SBPr.

A second class of EMD-derived metrics will be referred to as *rate metrics* and are listed in Table I. These are the average IMF instantaneous frequencies (ω, Sec. II E). Both rate metrics exhibit a significant dependence on chunk duration [ω_1: $F(1,620) = 6.3$, $p < 0.001$; ω_2: $F(1,620) = 70.8$, $p < 0.001$], which notably differed in direction between metrics. The average value of the instantaneous frequency of IMF_1 (ω_1) is slightly less than 6 Hz, suggesting that the period of fluctuation in IMF_1 tends to be about 180 ms. This is consistent with what one would expect for the average syllable duration in English spontaneous speech, although somewhat shorter than is typically reported (cf. Clopper and Smiljanic, 2011, who report speaking rates ranging from 5.1 to 5.5 syllables/s depending on gender and accent). The within-subject mean ω_1 values fall in a range of 5.3–6.2 Hz (approximately 160–190 ms periods); variation in these values is likely diagnostic of interspeaker differences in speech rate. The mean ω_2 is near 2.2 Hz for most speakers, and this value increases with chunk duration. As noted in Sec. II E, ω_1 fluctuates quite drastically within a given utterance due to abrupt changes in the amplitude of syllable-time-scale oscillations. This can be seen by contrasting the standard deviations of ω_1 with the standard deviations of ω_2—the latter are substantially less variable, ranging from 2.0 to 2.4 Hz. However, because ω_2 is expressed in the frequency domain, the variation in time-scale is somewhat obfuscated as this range of variation corresponds to wavelengths from approximately 415–500 ms. Indeed, the normalized detectable effect sizes for ω_1 and ω_2 are quite comparable, falling in the range of 1% (in shorter utterances) to 5% (in longer utterances).

A third class of metrics investigated here are the *rhythmic stability metrics*; these correspond to the within-chunk variance of ω_1 and ω_2, shown in the rightmost two columns of Fig. 4. As described in Sec. II E, these measures represent the instability of the instantaneous frequencies of IMF_1 and IMF_2, respectively. In other words, larger values indicate that the instantaneous frequency of an IMF changed more over the course of a chunk, suggesting a lower degree of rhythmicity. Both rhythmic stability metrics exhibited a significant effect of chunk duration [var-ω_1: $F(1,620) = 156.7$, $p < 0.001$; var-ω_2: $F(1,620) = 418.9$, $p < 0.001$], whereby the instantaneous frequencies became more variable in longer chunks.

In sum, our analysis of the effect of chunk duration on envelope-based metrics shows the following: two of the three power distribution metrics—SBPr and IMFr—exhibit stable mean values across a range of chunk durations from 1.0 to 2.6 s; in contrast, the centroid tends to decrease with increasing chunk duration. Rhythmic stability metrics (var-ω_1 and var-ω_2) are likewise influenced by chunk durations, indicating that IMF instantaneous frequencies vary more in longer chunks and that rhythmic characteristics are dynamic in any stretch of speech. Rate metrics show contrasting effects of duration: The rate of syllable-time-scale envelope oscillations decreases in longer chunks, whereas the rate of supra-syllabic time-scale oscillations increases. Furthermore, all metrics exhibit a decrease in variance as chunk durations are increased; however, due to diminishing sample sizes when larger chunks are used, there is a decrease in detectable effect size (and hence decrease in statistical power).

B. Identification of highly rhythmic speech

Envelope-based metrics can be used to identify chunks of speech that exhibit a high degree of envelope periodicity on syllabic and/or supra-syllabic time-scales; moreover, this identification can target utterances with relatively fast or relatively slow envelope periodicities. Section IV discusses why targeted identification of rhythmic utterances is a useful tool

TABLE I. Types of envelope metrics (EMs), definitions, and interpretations.

Type	Metric	Description	Interpretation
Power distribution metrics	$IMFr_{12}$	Ratio between power of IMF_2 and IMF_1	Relative amount of power in supra-syllabic and syllabic-time-scale oscillations
	$SBPr_{3.5}$	Ratio between power in envelope spectrum bands (1/3.5/10 Hz)	Relative amount of spectral power in supra-syllabic and syllabic time-scale oscillations
	$CNTR_{1\text{-}10}$	Envelope spectrum centroid calculated over 1-10 Hz band	
Rate metrics	ω_1	Mean within-utterance instantaneous freq. of IMF_1	Rate of syllabic oscillations
	ω_2	Mean within-utterance instantaneous freq. of IMF_2	Rate of supra-syllabic oscillations
Rhythmic stability metrics	var. ω_1	Variance of within-utterance instantaneous freq. of IMF_1	Stability of syllabic oscillations
	var. ω_2	Variance of within-utterance instantaneous freq. of IMF_2	Stability of syllabic oscillations

for research. Here we briefly describe the approach to identifying highly rhythmic utterances and present an example of chunk with a high degree of supra-syllabic time-scale periodicity.

To identify highly rhythmic stretches of speech, two factors in line with the definition of rhythm provided in Sec. I are taken into account: Rhythmic stability and instantaneous frequency. However, one must first decide what time-scale to consider and hence which IMF to use. Here an example is presented based on IMF_2, which corresponds to supra-syllabic periodicity in the envelope. However, there may be circumstances in which syllable-time-scale periodicity is of interest or in which multiple time-scales may be taken into consideration. In the present examples, rhythmicity is taken to be reflected by rhythmic stability and hence indexed by var. ω_2. In other words, when the instantaneous frequency of IMF_2 is fairly constant over a chunk, var. ω_2 is relatively low and this indicates a high degree of rhythmicity. Hence one can set a criterion z score of var. ω_2 for a given duration to identify chunks with stable rhythm. Furthermore, one can incorporate criteria for the instantaneous frequency of the IMF_2 periodicity (ω_2), which indexes whether the oscillation is relatively fast or slow. If ω_2 is very low, the IMF_2 periodicity may represent energy associated with phrasal emphasis patterns. Figure 5 illustrates the envelope, first and second IMFs, and envelope spectrum of an example highly rhythmic chunk from the Buckeye corpus. In Sec. IV, we discuss potential applications of this approach to identifying rhythmic utterances.

C. Cross-linguistic variation

Envelope-based metrics can also be used to characterize cross-linguistic differences in speech rhythm. The analysis presented in the following text concerns the effects of language and elicitation method on the mean values of envelope-based metrics (EMs) for each elicitation method included in the SLab corpus of English, German, Greek, Italian, Korean, and Spanish (for details, see Arvaniti, 2012). An ANOVA was conducted for each metric listed in Table I with repeated measures over speaker means. The effects of elicitation method (spontaneous speech, read sentences, and read passages), language (English, Korean, German, Italian, Greek, and Spanish), and their interaction were included. Although the read sentences were controlled to exhibit simplex, complex, and mixed syllable structure, a separate ANOVA showed that the effect of syllable structure was not significant for any metric, and hence these conditions were collapsed in the analyses presented in the following text.

1. Power distribution metrics

Cross-linguistic differences were observed for all power distribution metrics. ANOVAs showed language to be a significant factor [IMFr: $F(5,125) = 8.8$, $p < 0.001$; SBPr: $F(5,125) = 8.3$, $p < 0.001$; centroid: $F(5,125) = 11.3$, $p < 0.001$] with moderate effect sizes. The elicitation method and language ANOVA effect sizes (η^2) for all metrics are shown in Table II. *Post hoc* Tukey HSD comparisons showed that the IMFr and SBPr were significantly greater in English than the other languages (see Fig. 6), and likewise the centroid was significantly lower in English than the other languages—both of these findings point to a relatively high degree of supra-syllabic periodicity in English and minimal differences among the other languages in the corpus. Elicitation method also had a significant effect on the power distribution metrics [IMFr: $F(2,125) = 12.7$, $p < 0.001$; SBPr: $F(2,125) = 16.2$, $p < 0.001$, centroid: $F(2,125) = 29.1$, $p < 0.001$]. *Post hoc* tests showed spontaneous speech IMFr and SBPr were significantly higher, and spontaneous speech centroids significantly lower than those of read speech (sentences and passages); read passages and sentences did not differ.

2. Rate metrics

Analyses showed significant effects of language on both ω_1 and ω_2 [ω_1: $F(5,125) = 7.1$, $p < 0.001$; ω_2: $F(5,125) = 4.3$, $p = 0.001$]. Mean values by language and elicitation are shown in the middle row of Fig. 6. For ω_1, *post hoc* tests showed that English exhibited significantly lower ω_1 (mean

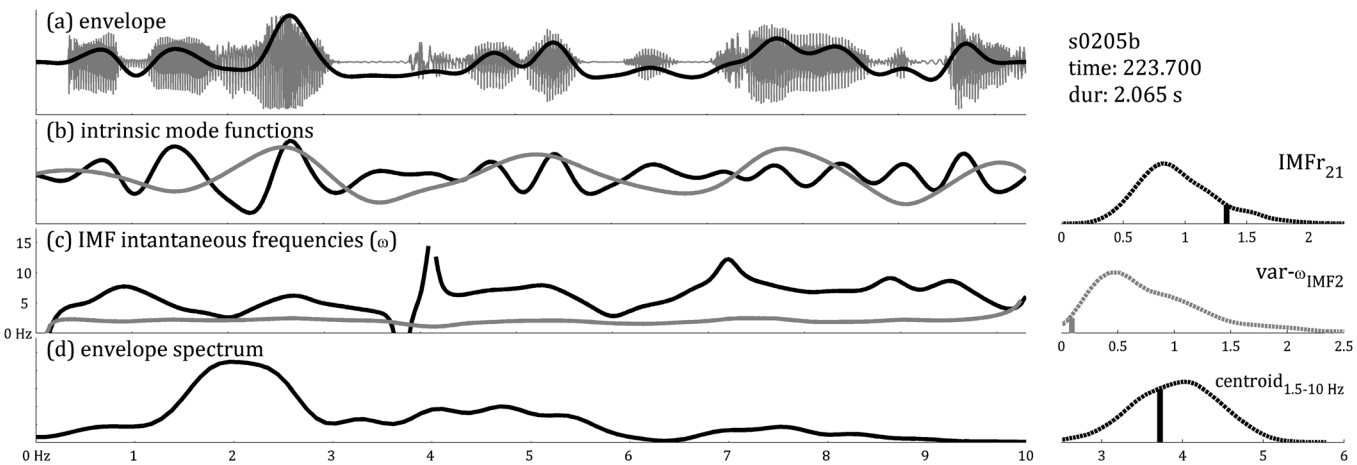

FIG. 5. Example of a highly rhythmic 2.05 s utterance from the Buckeye corpus. Left: (a) Waveform and amplitude envelope; (b) intrinsic mode functions; (c) instantaneous frequencies; (d) envelope spectrum. Right: Distributions of metrics across the Buckeye corpus (vertical lines = values of this example). Transcribed as "ever went to the electric chair that they…."

TABLE II. Elicitation method and language ANOVA effect sizes (η^2) for each metric. An η^2 of 0.10 is considered a small effect, and an η^2 of 0.25 is considered a medium-size effect.

	$IMFr_{12}$	$SBPr_{3.5}$	$CNTR_{1\text{-}10}$	ω_1	ω_2	var. ω_1	var. ω_2
Elicitation method	0.12	0.14	0.23	0.05	0.04	0.10	0.03
Language	0.21	0.18	0.22	0.20	0.13	0.14	0.18

$= 5.9$ Hz) than Greek and Italian (mean ω_1 from 6.2 to 6.4 Hz), suggesting a somewhat lower speaking rate for the former. Likewise, ω_1 was significantly lower in German and Spanish than in Italian. *Post hoc* cross-linguistic comparisons of ω_2 showed that German exhibited higher ω_2 than English, Korean, and Spanish. Effects of elicitation method were significant for both rate measures [ω_1: $F(2,125) = 4.4$, $p = 0.02$; ω_2: $F(2,125) = 3.5$, $p = 0.03$], although the effect sizes are rather small (Table II). *Post hoc* comparisons showed that spontaneous speech had lower ω_1 than the read sentences and lower ω_2 than the read passage. This suggests that spontaneous speech is slower than read speech on both syllabic and supra-syllabic time scales.

3. Rhythmic stability metrics

Rhythmic stability metrics, which index the stability of syllabic and supra-syllabic rhythms within utterances, also reveal effects of both language and elicitation method, illustrated in the bottom row of Fig. 6. Significant language effects were present for both var. ω_1 [$F(5,125) = 5.0$, $p < 0.001$] and var. ω_2 [$F(5,125) = 6.8$, $p < 0.001$]. *Post hoc* tests showed that English and German exhibited significantly higher var. ω_1 than Korean and Spanish, indicative of greater variability in the timing of syllabic rhythms. Greek and Italian showed significantly higher var. ω_2 than Korean and Spanish, and German also showed higher var. ω_2 than Korean, indicating the presence of cross-linguistic differences in the regularity with which stresses appear in speech. Elicitation method also had a significant effect on rhythmic stability [var. ω_1: $F(2,125) = 9.7$, $p < 0.001$; var. ω_1: $F(2,125) = 3.2$, $p = 0.046$] but with quite small effect size for the latter. *Post hoc* comparisons showed that in both spontaneous speech and read sentences var. ω_1 was significantly greater than in the read passages. In contrast, var. ω_2 was significantly greater in sentences than in spontaneous speech.

IV. DISCUSSION

The results presented in the preceding text evaluated the influence of chunk duration on envelope-based measures and demonstrated two applications of envelope-based rhythm analysis: Targeted identification of highly rhythmic speech and cross-linguistic comparison of aspects of rhythm.

In Sec. III A, the influence of the effects of chunk duration on envelope-based metrics was examined, and two general conclusions were drawn. First, as chunk length increases, the metric variances decrease. This may be due to a tendency for longer chunks to better represent the rhythmicities typically present on utterance time-scales. Second, as chunk length increases, the detectable effect size increases modestly due to the decrease in the number of chunks; the consequence of this is a slightly diminished ability to resolve differences among samples. Hence in practice, there is an inherent trade-off between chunk duration and anticipated statistical power.

On the other hand, no optimal chunk duration that minimizes variance of ω_1 or ω_2—i.e., a duration over which IMF periodicity is most stable—was observed in the range of durations analyzed, 1–2.6 s. This suggests either that variance minima occur in the limit of infinite duration or exist at a longer duration outside of this range. However, speakers typically pause or hesitate before producing chunks that would be longer than this. Hence in practice, the relative scarcity of long continuous utterance stretches in spontaneous speech corpora argues against using the minimum variance as a guide for selecting chunk durations.

An alternative approach involves consideration of the typical period of the highest IMF of interest. For example,

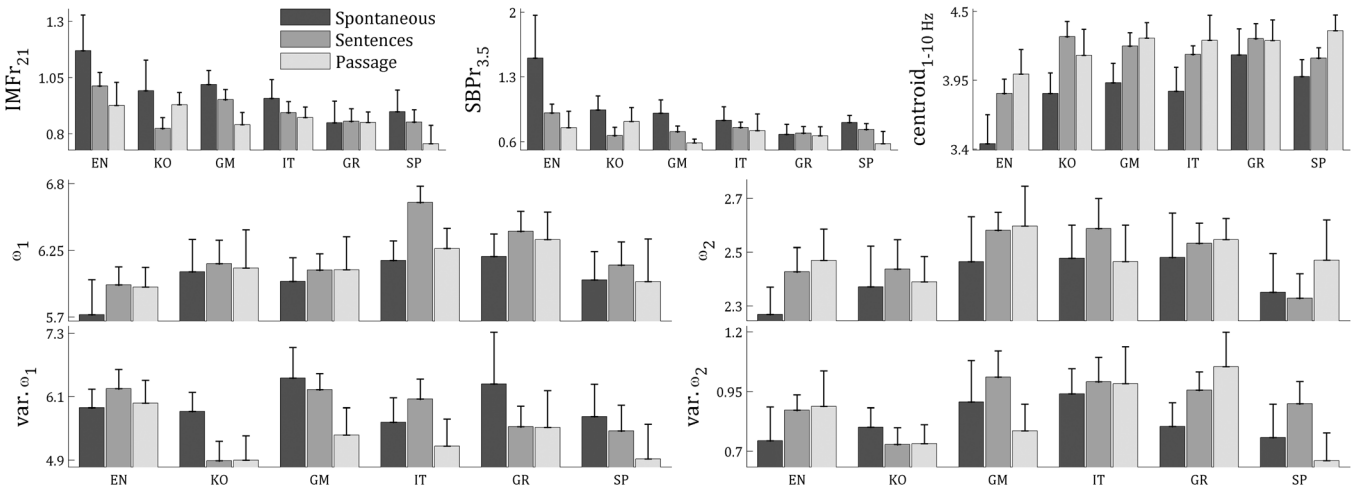

FIG. 6. Across-speaker mean IMFr, SBPr, centroid, instantaneous frequencies (ω) and variances (var. ω) for IMF1 and IMF2 by language and elicitation method (spontaneous, read sentences, read passage). Languages shown are English, Korean, German, Italian, Greek and Spanish. Means are shown with 2.0 standard error intervals.

the average within-speaker ω_2 in the Buckeye corpus fell in the range of approximately 2.0–2.5 Hz (period 500-400 ms). Because four cycles of a 500 ms period can occur in 2.0 s, this chunk duration is appropriate for adequately capturing stress-related low-frequency periodicities and assessing their stability. In any case, broader goals of the analysis must be brought into consideration. If one was interested in lower-frequency phrasal rhythms, a longer duration would be preferable; in contrast, if the interest is only in syllable-time-scale fluctuations, a shorter duration would be more suitable.

In Sec. III B, we demonstrated how to identify highly rhythmic utterances using the variance of ω_2, i.e., the variance of the instantaneous frequency of IMF_2. Although the quantification of rhythm developed here treats all speech as exhibiting a degree of rhythmicity, chunks with low var. ω_2 exhibit a high degree of stability in the frequency of IMF_2 oscillations, indicating particular regularity in the frequency of supra-syllabic time-scale oscillations; in English, these are likely to represent stress-related periodicity. As this illustration of the use of var. ω_2 indicates, envelope metrics can be used to identify utterances that exhibit different degrees of rhythmicity on any time-scale. Furthermore, by considering the average instantaneous frequency of a chunk (ω_2) in combination with its variance (var. ω_2), another dimension of identification can be added: Chunks with fast regularly repeating stress can be distinguished from ones with slow regularly repeating stress; hence degree of rhythmicity can be dissociated from the rate of occurrence of the envelope oscillations that are associated with rhythm.

Targeted identification of highly rhythmic speech in particular has several potential applications. One of the most important is locating data that allow for investigation of factors that give rise to variation in speech rhythmicity. These factors are currently not well understood because it is difficult not only to locate rhythmic speech in naturalistic corpus data but also to quantitatively characterize rhythmicity. The current approach offers a fairly straightforward method for locating and quantifying rhythmic stretches of speech. Although we have not systematically investigated these factors, our observations of the Buckeye corpus suggest two things: First, rhythmicity sometimes arises from identifiable causes, which include behaviors such as counting, repeating words or short phrases, hesitation with filled pauses (i.e., uh, um), and affective and contrastive emphasis; second, that there are plenty of cases in which speakers just happen to produce a stretch of speech with a high degree of periodicity on the stress-time-scale without a clearly identifiable cause. Some of these occurrences may be attributable to the lexical stress patterns of the utterance; another possibility is that the speech motor control system occasionally enters a mode in which stressed syllables are timed quite isochronously, as if strongly governed by an oscillatory cycle. Our understanding of the cognitive mechanisms of speech production should benefit from a better understanding of the factors that give rise to both rhythmic variation and stability.

A further use of rhythmicity-based identification is to investigate interactions between speech rhythm and segmental articulation, epenthesis, or deletion. Previous work based on envelope spectra has shown that the likelihoods of phonetic deletions of vowels and consonants are influenced by the rhythmicity of the context in which they occur (Tilsen, 2008). Furthermore, the method can be used to obtain highly naturalistic conversational stimuli for use in rhythm perception experiments in which correlations between perceived rhythmicity and envelope-based metrics could be investigated.

In Sec. III C, envelope metrics were applied to chunks extracted from the SLab corpus to characterize cross-linguistic differences in speech rhythm. Regarding cross-linguistic differences, the power distribution metrics showed that English exhibited significantly more low-frequency periodicity than all of the other languages. This is partly consistent with previous studies that have used interval-based metrics and classify English as a prototypical stress-timed language (e.g., Ramus *et al.*, 1999; Grabe and Low, 2002; White and Mattys, 2007). It is worth noting that German, typically seen as a stress-timed language too (e.g., Grabe and Low, 2002), did not pattern with English and was not differentiated from canonical "syllable-timed" languages. Overall these cross-linguistic comparisons indicate that languages can differ with regard to the strength of supra-syllabic periodicities. However, envelope metrics provide no strong evidence for rhythm classes based on a stress-timing vs syllable-timing distinction; rather, there appears to be a cline along which languages differ: Supra-syllabic rhythms can be more or less prevalent but they are clearly always present. In turn this would suggest that even languages that lack word-level stress, such as Korean (Jun, 2005), do utilize a form of salience that occurs less often than every syllable.

Rate metrics (instantaneous frequencies of IMF1 and IMF2) also exhibited significant cross-linguistic differences. On the one hand, English showed relatively low values of syllable-time-scale frequencies (ω_1) when compared to Greek and Italian, a result that concords with the observation that English has a comparatively low speaking rate (Clopper and Smiljanic, 2011). Spanish and German also exhibited low ω_1 compared to Italian. It is worth noting, however, that both these effects were very small (on the order of 10 ms in mean period). More differences were observed in supra-syllabic rate, where German exhibited significantly higher ω_2 compared to Greek, Korean, and Spanish, suggesting that stress-level prominences are more frequent in the former than the latter; in this case, the differences in mean periods are on the order of 40–50 ms, which is about 10% of the average period of this time-scale, hence of some perceptual relevance. Together, these findings lead to the conclusion that instantaneous frequencies may capture robust cross-linguistic similarities in syllable- and stress-level periodicities. This conclusion also accords with the observation of Dauer (1983) that stresses appear cross-linguistically every 500 ms or so (a rate very similar to that obtained here for ω_2).

It is apparent from these data that the rate metrics did not pattern similarly to the power distribution metrics; this suggests that rate metrics do not capture the same sort of information as power distribution metrics and that rate and power distribution are independent aspects of rhythm. For example, although Spanish and Korean ω_2 pattern with English, their power distribution metrics are quite dissimilar.

This implies that although the rate of prominent syllables in these languages are similar, relatively more envelope power is associated with the stress time-scale in English. This could arise from a difference in the phonetic manifestations of prominence between the languages, for example, if English employs intensity and duration to a greater extent than Spanish and Korean. One of the desirable qualities of the EMD algorithm is that it allows for rate and power distribution to be dissociated in this way.

Like rate metrics, rhythmic stability metrics exhibit significant cross-linguistic differences. Rhythmic stability metrics reflect the degree to which the instantaneous frequency of an IMF remains constant throughout a stretch of speech. One consistent feature of the observed patterns is that Korean and Spanish exhibit the least amount of variance in both syllabic and supra-syllabic time-scales; however, the languages that exhibit the highest variances differ between time-scales. For the syllabic time-scale associated with IMF_1, it is English and German that have a relatively higher variance than Korean and Spanish, whereas for the supra-syllabic time-scale, it is Greek and Italian. Because English and German exhibit greater variation in syllable structure than other languages in the SLab corpus, as well as more complex syllable types (Dankovičová and Dellwo, 2007), this makes sense: English and German have a substantial degree of variation in syllable duration and hence higher variance of ω_1 (cf. Crystal and House, 1990). In contrast, Greek and Italian may be more similar to Korean and Spanish by virtue of lacking variation in syllable duration, so high values of var. ω_1 would not be anticipated. As for the observation that Greek and Italian—but not English and German—exhibited greater var. ω_2 than Korean and Spanish, we suggest that these differences may reflect the fact that the occurrence of stresses is less regular in Greek and Italian than English and German. Both Italian and Greek allow lapses of several unstressed syllables and neither has a rhythm rule equivalent like German and English (cf. Farnetani and Kori, 1990, on Italian; Arvaniti, 1994, 2007, on Greek).

Effects of elicitation method were observed for all metrics. In the power distribution and rate metrics, the effects distinguish between spontaneous and read speech. Spontaneous speech utterances exhibited relatively more low-frequency power than read speech (passages and sentences), relatively slower syllabic rates (compared to sentences), and supra-syllabic rates (compared to passages). One likely explanation for why spontaneous speech differs from read speech in these ways involves greater irregularity in speech rate arising from variation in semantic/conceptual planning processes (cf. Levelt, 1989). In read speech, formulation of a conceptual message is externally driven, and so this may facilitate planning and hence increase rate. Exactly why this has an effect on power distribution metrics is less clear, but in any event, these differences have implications for studies addressing cross-linguistic differences on rhythm as they show that conclusions based on read speech cannot be generalized to conversational speech.

Finally, we note for all metrics that there were no effects of syllable structure in the read sentences. The absence of such an effect suggests that syllable structure does not have a strong influence on the oscillatory variation in the speech envelope. This is quite interesting in comparison with interval-based metrics for which effects of syllable structure have been demonstrated by previous studies (Arvaniti, 2012; Prieto *et al.*, 2012). This difference in results affirms that envelope metrics are to some extent tapping into an alternative source of information than interval metrics—continuous variation in the vocalic energy amplitude envelope.

Analysis of amplitude envelopes provides a different approach to characterizing utterance rhythmicity than does analysis of interval durations. The difference resides in whether phonological units or fluctuations in signal energy are the objects of investigation. We make no claim that one approach is preferable to the other as this likely relies upon the analytic goals of an investigation. It is noteworthy that the direction of causality between rhythm and interval durations has never been established: Do the phonological properties that can lead to variation in interval durations cause the perception of rhythmic differences, are they a consequence of them, or are they both consequences of an underlying phenomenon? Envelope metrics, which were found to exhibit only weak dependence on syllable structure, may be good candidates for addressing these questions precisely because they are less dependent on syllable phonotactics than interval metrics.

V. CONCLUSION

This paper has presented a new method for characterizing speech rhythm based on empirical mode decomposition of the vocalic energy amplitude envelope. Accordingly, several types of envelope-based metrics (EMs) for analyzing utterance rhythm were presented. A distinction was drawn among three types of metrics: Power distribution metrics, which capture the relative power in syllabic vs supra-syllabic oscillations in the envelope; rate metrics, which capture the frequencies of those oscillations; and rhythmic stability metrics, which capture the stability of the oscillations. The application of these metrics to the Buckeye and SLab corpora showed that EMs are sufficiently flexible to capture information about periodicities that likely correspond to different linguistic constructs (such as the syllable, foot, and phrase), while they can also be used to examine rhythmicity in speech and investigate cross-linguistic differences in rhythmicity.

Envelope metrics are derived from a physical and dynamic representation of speech, the amplitude envelope. However, it has yet to be established how well envelope metrics correlate with rhythm perception. Future studies should assess the degree to which envelop-based metrics correlate with perceptual judgments of rhythmicity. Moreover, it is becoming clear that additional dimensions of speech are involved in rhythm perception: For example, pitch appears to play an important role (e.g., Kohler, 2008). Nevertheless, we have shown that envelope metrics can be useful for identifying rhythmic speech and studying cross-linguistic differences in rhythm. Further uses may be found in automated

measurements made in clinical or pedagogical contexts, and in speech recognition technology.

ACKNOWLEDGMENTS

We would like to thank the audience at the special session on "Speech Rhythm in Production, Perception and Acquisition," 162nd Meeting of the Acoustical Society of America, San Diego, October 31–November 4, 2011, for their helpful comments. Thanks are also due to SLab members Younah Chung, Page Piccinini, Tristie Ross, and Nadav Sofer and to Noah Girgis for their help with the collection and analysis of the SLab corpus data. The financial support of the University of California San Diego Committee on Research through Grant No. LIN201G to Amalia Arvaniti with Tristie Ross as GSR is hereby gratefully acknowledged.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Edinburgh), Chap. 6, pp. 89–110.
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.
Arvaniti, A. (**1994**). "Acoustic features of Greek rhythmic structure," J. Phonetics **22**, 239–268.
Arvaniti, A. (**2007**). "Greek phonetics: The state of the art," J. Greek Linguist. **8**, 97–208.
Arvaniti, A. (**2012**). "The usefulness of metrics in the quantification of speech rhythm," J. Phonetics **40**, 351–373.
Bertinetto, P. M. (**1989**). "Reflections on the dichotomy 'stress' vs. 'syllable timing,' " Rev. Phonét. Appl. **91-93**, 99–129.
Chatfield, C. (**1975**). *The Analysis of Time Series* (Chapman and Hall, London), Chap. 7, pp. 127–132.
Clopper, C. G., and Smiljanic, R. (**2011**). "Effects of gender and regional dialect on prosodic patterns in American English," J. Phonetics **39**, 237–245.
Crystal, T. H., and House, A. S. (**1990**). "Articulation rate and the duration of syllables and stress groups in connected speech," J. Acoust. Soc. Am. **88**, 101–112.
Cummins, F. (**2009**). "Rhythm as affordance for the entrainment of movement," Phonetica **66**, 15–28.
Cummins, F., and Port, R. (**1998**). "Rhythmic constraints on stress timing in English," J. Phonetics **26**, 145–171.
Dankovičová, J., and Dellwo, V. (**2007**). "Czech speech rhythm and the rhythm class hypothesis," in *Proceedings of 16th International Congress of Phonetic Sciences*, Saarbrücken, pp. 1241–1244.
Dauer, R. M. (**1983**). "Stress-timing and syllable-timing reanalyzed," J. Phonetics **11**, 51–62.
Dauer, R. M. (**1987**). "Phonetic and phonological components of language rhythm," in *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallinn, pp. 447–449.
Dellwo, V. (**2006**). "Rhythm and speech rate: A variation coefficient for deltaC," in *Language and Language-Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*, edited by P. Karnowski and I. Szigeti (Peter Lang, Frankfurt am Main), pp. 231–241.
Farnetani, E., and Kori, S. (**1990**). "Rhythmic structure in Italian noun phrases: A study on vowel duration," Phonetica **47**, 50–65.
Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., and Scott, S. K. (**2002**). "Amplitude envelope onsets and developmental dyslexia: A new hypothesis," Proc. Natl. Acad. Sci. U.S.A. **99**(16), 10911–10916.
Grabe, E., and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), pp. 515–546.
Howell, P. (**1988**). "Prediction of P-center location from the distribution of energy in the amplitude envelope," Percept. Psychophys. **43**(1), 90–93.
Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (**1998**). "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proc. R. Soc. London, Ser. A **454**, 903–995.
Jun, S.-A. (**2005**). "Korean intonational phonology and prosodic transcription," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford), pp. 201–229.
Kiesling, S., Dilley, L., and Raymond, W. (**2006**). "The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of conversational speech," Department of Psychology, Ohio State University, Columbus, OH, available at www.buckeyecorpus.osu.edu (Last viewed 11/07/2012).
Kohler, K. J. (**2008**). "The perception of prominence patterns," Phonetica **65**, 257–269.
Kohler, K. J. (**2009a**). "Whither speech rhythm research?" Phonetica **66**, 5–14.
Kohler, K. J. (**2009b**). "Rhythm in speech and language. A new research paradigm," Phonetica **66**, 29–45.
Lee, C. S., and Todd, N. P. M. (**2004**). "Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora," Cognition **93**, 225–254.
Levelt, W. J. M. (**1989**). *Speaking: From Intention to Articulation* (The MIT Press, Cambridge, MA), Chap. 4, pp. 107–160.
Lloyd James, A. (**1940**). *Speech Signals in Telephony* (Pitman and Sons, London), Chap. III, pp. 16–27.
Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual centers (P-centers)," Psychol. Rev. **83**(5), 405–408.
Nolan, F., and Asu, E. L. (**2009**) "The Pairwise Variability Index and coexisting rhythms in language," Phonetica **66**, 64–77.
O'Dell, M. L., and Nieminen, T. (**1999**). "Coupled oscillator model of speech rhythm," in *Proceedings of the XIVth International Congress of Phonetic Sciences*, edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey (AIP, New York), Chap. 2, pp. 1075–1078.
Pike, K. (**1945**). *The Intonation of American English* (University of Michigan Press, Ann Arbor), pp. 34–35.
Pitt, M., 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**, 89–95.
Pompino-Marschall, B. (**1989**). "On the psychoacoustic nature of the P-center phenomenon," J. Phonetics **17**, 175–192.
Prieto, P., Vanrell, M., Astruc, L., Payne, E., and Post, B. (**2012**). "Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish," Speech Commun. **54**, 681–702.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **75**, 265–292.
Tilsen, S. (**2008**). "Relations between speech rhythm and segmental deletion," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Vol. 44, 211–223.
Tilsen, S. (**2009**). "Multitimescale dynamical interactions between speech rhythm and gesture," Cogn. Sci. **33**, 839–879.
Tilsen, S., and Johnson, K. (**2008**). "Low-frequency Fourier analysis of speech rhythm," J. Acoust. Soc. Am. **124**, EL34–39.
White, L., and Mattys, S. L. (**2007**). "Calibrating rhythm: First language and second language studies," J. Phonetics **35**, 501–522.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (**2010**). "How stable are acoustic metrics of contrastive speech rhythm?," J. Acoust. Soc. Am. **127**, 1559–1569.
Yu, A. (**2010**). "Tonal effects on perceived vowel duration," in *Laboratory Phonology 10*, edited by C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée (Mouton de Gruyter, Berlin), pp. 151–168.