What marks the beat of speech?

Greg Kochanski[a] and Christina Orphanidou
Phonetics Laboratory, University of Oxford, 41 Wellington Square, Oxford OX1 2JF, UK

(Dated: June 20, 2008)

Which acoustic properties of the speech signal differ between rhythmically prominent syllables and non-prominent ones? A production experiment was conducted to identify these acoustic properties. Subjects read out repetitive text to a metronome, trying to match stressed syllables to its beat. The analysis searched for the function of the speech signal that best predicts the timing of the metronome ticks. The most important factor in this function is found to be the contrast in loudness between a syllable and its neighbors. The prominence of a syllable can be deduced from the specific loudness in an (approximately) 360 millisecond window centered on the syllable in question relative to an (approximately) 800 millisecond-wide symmetric window.

PACS numbers: 43.72.Ar, 43.71.Sy, 43.70.Fq

This paper is published in the http://asa.aip.org, (ISSN 0001-4966) as http://dx.doi.org/10.1121/1.2890742, May 2008, Volume 123, Issue 5, pages 2780–2791. A preprint can be found at http://kochanski.org/gpk/papers/2006/2006tapping.pdf .

I. INTRODUCTION

Patterns of prominence define the rhythmicity of speech, which is an important characteristic of stress-timed languages, English among them. The purpose of this study is to find which acoustic properties of the speech signal mark rhythmic prominence. We investigate this with a new production-based experiment; it involves auditory perception only to the extent that subjects extract a beat from a metronome tick.

While a number of papers have studied acoustical correlates of prominence, human perception of speech has played a central part in most of them, with listeners being asked to identify prominent units in speech. The prototypical experiments are Fry (1955, 1958), who synthesized isolated disyllabic words and asked listeners to choose which syllable was stressed ($\approx$ prominent). These papers are mostly of historic interest because the synthesis may not have been very realistic and the stimuli were very simple compared to natural speech.

Other workers (Beckman, 1986; Streefkerk *et al.*, 1999; Brenier *et al.*, 2005; Silipo and Greenberg, 2000; Kochanski *et al.*, 2005) have studied more realistic speech, some using actual conversations. These studies investigated the relative importance of a variety of acoustic factors to prominence. Streefkerk *et al.* (1999) tested several acoustic features such as duration, loudness, spectral slope of vowels, as well as median f_0 over a syllable, and the range of f_0 over a syllable. They concluded that all but the spectral slope had promise as predictors of prominence.

Brenier *et al.* (2005) tested 12 acoustic and lexical features and found that the maximum intensity was the most effective for emphasis detection, followed by duration and f_0. Silipo and Greenberg (2000) found that a combination of intensity and syllable duration was the best predictor of their "prosodic stress" (prominence by our definition). Kochanski *et al.* (2005) tested syllable prominence in seven dialects of British and Irish English and concluded that loudness and duration primarily marked prominence in speech, with f_0 relatively unimportant. The experimental consensus is that louder and longer syllables tend to be heard as prominent.

Most of these papers suffer from a common problem: the task used to define prominence is highly artificial and unnatural. Prominence was evaluated by a variant of the following experimental procedure:

1. Display the speech waveform and f_0 trace on a computer monitor.
2. Allow the subject (typically a linguist) to listen to the utterance or parts thereof many times.
3. Have the subject find and mark any prominent syllables on the computer screen.

Participants in a conversation normally don't see a graphical representation of the speech, and they hear each utterance only once. Further, they do not consciously classify each syllable as prominent or not,[1] nor is a mouse click the normal behavioral response to speech. Consequently, it is important to see if one would get similar results with a different task.

In our experiment, we employ a more natural approach by giving subjects a production rather than a perception task. The classic production experiment is Lieberman (1960), who built a machine classifier to study acoustic properties of prominence. He made perceptual judgments unimportant by selecting sentences that had unambiguous prominence patterns, at the cost of studying utterances that were presumably much more carefully articulated than normal speech. Fear *et al.* (1995) is a noteworthy modern representative of such experiments. Rhythmic speech, often spoken to a metronome, has been extensively studied in relation to stress in speech production and perception (Fowler, 1979; Cummins and Port, 1998; Lehiste, 1973). Boutsen *et al.* (2000) is a recent production experiment; they concluded that speakers use intensity to mark stress patterns, in agreement with the majority of perception-based experiments. Our experiment is most closely related to Cooper and Allen (1977) who collect very similar data however we use a different analysis to look at the loudness contrasts in more detail. They focus on the differences between normal subjects and stutterers; we look at what aspects of the speech of normal speakers most accurately reflects the metronome rhythm.

[a]Electronic address: greg.kochanski@phon.ox.ac.uk

We ask subjects to read text to a metronome, matching their reading to the beat. Before the experimental part of the task, there is a training part which is intended to accustom the subjects to treating metronome ticks as proxies for prominence. We then analyze the speech to find stable timing relationships between acoustic properties and the metronome ticks.

One important feature of our work is that we use naive native speakers, rather than linguists. This reduces the possibility that the prominence marks may be influenced by theoretical expectations. No conscious judgments of prominence are required and no visual cues are involved in this experiment. This experiment provides a different experimental view of prominence, with different biases and limitations from the classical technique.

One could argue that the task is not ideal in that the phrases must be spoken repetitively, which brings in the risk that repetitive speech is different from more natural speech. However, we have checked this possible problem in other, more recent work (Kochanski and Orphanidou, 2007). We found that spectral differences between repetitive speech and speech from a list of randomized phrases are not large, so one would expect other aspects of speech, like prominence, to be similarly unaffected.

Work has been published in the related field of music analysis, where algorithms have been designed to extract the beat and/or metrical pattern of music using metronomic stimulation. See Scheirer (1998), Klapuri *et al.* (2006), Todd and Brown (1996), Large and Palmer (2002) and references therein. The techniques differ from ours for various reasons, for instance, unlike speech, many musical sounds have sharp onsets (e.g. drums, pianos). Further, algorithms can exploit the fact that the beat of music is very nearly periodic; they can keep long-term correlations and use them to help predict the next beat. Conversely, our algorithm is designed to be applicable to normal speech, which is not very periodic. It therefore operates on a small window of time, with no explicit memory of previous prominences.

II. EXPERIMENTAL METHODS

The experiment involved several tasks and lasted for approximately 1 hour; only some of the later tasks are analyzed. The earlier tasks were intended to train the subjects to put the metronome beat on metrically prominent syllables.[2]

A set of 53 short phrases (4–6 syllables) were central to the experiment (see Appendix B for a list). The phrases had 4 different metrical patterns: 12 of *SuSu*, 12 of *uSuS*, 13 of *SuuSuu*, and 16 of *SuuS* where *S* denotes a stressed syllable and *u* unstressed.[3] The phrases were selected from Project Gutenberg (Hart and volunteers, 2006), based on patterns of stress predicted by Unisyn (Fitt and Isard, 1999; Fitt, 2002).[4] Phrases were selected for broad coverage of phonemes, minimal repetition of words, and a lack of obsolete and unusual vocabulary. They were reviewed to confirm that the Unisyn stress assignments led to a reasonable reading. All phrases have at least one polysyllabic word and they have an average of 1.4 monosyllabic words.

As a warm-up, subjects read out a poem (Nesbitt, 2001, "My Excuse") and then read it out again while tapping their finger to what they considered to be stressed syllables.

Next, a metronome was connected to an earphone, and subjects were asked to choose the two most comfortable rates at which they could read text with a strong metrical pattern. To pick a comfortable metronome rate, the metronome was started and the subject was asked to read the poem again. After a few lines, the subject was asked "faster, slower, or is that OK?" If necessary, the metronome was adjusted by one click (typically 4 beats per minute) and the process continued until the subject said "OK." They were then asked to pick a second rate, either 4 beats faster or slower than the first one. They then read the poem at both rates, matching their reading to the beats of the metronome.

In the next tasks, subjects read 48 short paragraphs from which the phrases had been extracted, then read a randomized list of 264 phrases which included 5 repetitions each of the above set of phrases.[5]

In the final set of tasks, subjects were presented with the above set of phrases, from which 4 groups of 12 were randomly selected, with the groups balanced by metrical pattern. Subjects were asked to read out 10 consecutive repetitions of each phrase. The repetitive task was intended to allow the subject to settle upon the easiest (perhaps the most natural) metrical pattern. The number of repetitions was intended to allow subjects to conveniently say them all in one breath. The first group of 12 was simply read out. For the second group, the subject was asked to read and simultaneously "Tap your finger to what you consider a stressed syllable."

We analyzed the data from the third and fourth groups of 12 phrase, which were read out to metronome ticks. One group was read at each of the two rates that the subjects chose earlier. Subjects were instructed to "Read, trying to follow the beat of the metronome." The metronome rates were 86 ± 7 beats per minute (0.71 ± 0.06 s intervals between beats), and the mean length of 10 repetitions of the phrase was 13.4 ± 2.4 s. (In this paper, means and standard deviations are given in the form $\mu \pm \sigma$ where σ is the standard deviation of the distribution, not the standard error of the mean.) There were 0.91 ± 0.15 metronome ticks per stressed syllable (assuming stress as predicted by Unisyn).

Participants were linguistically naive speakers of Standard Southern British English. All were either undergraduate or graduate students at Oxford University. Five females and four males were recruited by mailing list advertisements.

Each subject was recorded with an electret microphone positioned approximately 10 cm from his/her mouth, to the side of the breath stream. Recordings were taken in a acoustically insulated recording booth. The audio signals were sampled at 32 kHz with 16 bit resolution.

Metronome ticks were fed to the subjects through an "earbud" style earphone on the opposite side from the microphone that recorded their speech. It was adjusted to a comfortable loudness level for each subject. The metronome ticks were recorded on one channel of a stereo recording and cannot be heard in the other channel which carries the microphone signal.

The start and end points of the speech were automatically determined by an algorithm that finds the borders of a loud interval, surrounded by quiet regions on both edges. It is a modified version of techniques used in (Kochanski *et al.*, 2005). All endpoints were manually checked at the same time we checked the tick marks. Fewer than 10% were adjusted.

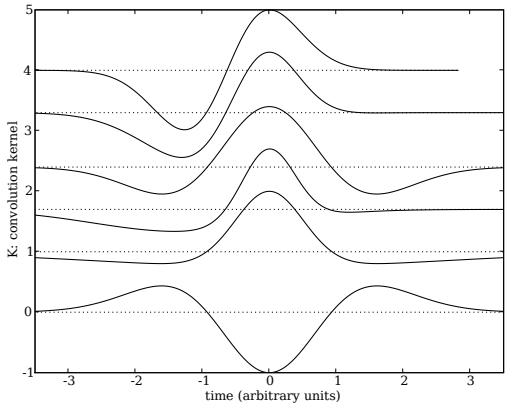

FIG. 1. Possible shapes for the convolution kernel, $K(\tau)$. More possible shapes can be obtained by reversing the time (delay) axis or the vertical axis.

III. ANALYSIS TECHNIQUES

In the analysis, we search for the function of the speech data that is best at predicting the metronome ticks. The function is expressed as a simple algorithm that has several adjustable parameters. It takes a waveform and yields a set of predictions; we compute the timing accuracy of the predictions and adjust the parameters to maximize that accuracy. We can then interpret the parameters to see how the subjects express rhythmical prominence.

A. The Algorithm

The algorithm computes a signal from the speech waveform, convolves the signal with a kernel, and picks maxima of the result. Large maxima become predictions of metronome ticks. The signal we use is related to the perceived loudness of the speech, but we also try modifications that include information from the pitch of the speech, voicing, and the overall slope of the local speech spectrum.

It is motivated by perception experiments (Kochanski *et al.*, 2005) showing that prominent syllables are typically louder. The algorithm begins by computing a time series of the specific loudness, $L(t)$, from Kochanski *et al.* (2005)[6], derived originally from Stevens (1971). This specific loudness is then convolved with a kernel K to yield $x(t)$ as an intermediate result. We use[7]

$$x = K * (L \cdot g). \tag{1}$$

Other acoustic properties of the speech will be included via g, and will be described in more detail in §III.B. We will test different values for g, beginning with $g = 1$, then functions of f_0, voicing periodicity, and spectral slope.

We chose a product form in Equation 1 to express the fact that the other acoustic parameters become irrelevant as L approaches zero. For instance, if the signal is quiet enough relative to near-by regions, the pitch of the quiet region will be perceptually unimportant.

We choose a convolution kernel, K, which has a zero mean and is the difference between two Gaussians:

$$K(\tau) = e^{-(\tau - a\sigma)^2/2(b\sigma)^2} - be^{-\tau^2/2\sigma^2}, \tag{2}$$

where τ is the time delay, σ controls the overall width of the kernel, b (which is typically less than 1) is the relative width of the positive Gaussian, and a controls where the positive Gaussian sits. If $a = 0$, K is symmetric in time and corresponds to the difference between a syllable and both of its neighbors; if $a \neq 0$, then the contrast with either the left or the right neighbor is more important. One can interpret σ as the width of the region that is used to normalize the local loudness information.

Figure 1 shows the range of shapes that $K(\tau)$ can take. The top two would respond best to step-wise increases in $L \cdot g$, the next down responds to the contrast of a syllable with the nearest syllable on each side. The next two show contrasts against broader regions than the nearest neighbor(s), and the lowest curve shows a kernel that responds to a local minimum in $L \cdot g$.

In the next step, we consider regions where $x(t)$ is positive and take the time of the largest value in each region, $t_i^{\max}$. The final set of syllable times, T, is then computed by dropping any values of $t_i^{\max}$ where either $x(t_i^{\max}) < r \cdot x(t_{i-1}^{\max})$ or $x(t_i^{\max}) < r \cdot x(t_{i+1}^{\max})$. In other words, one drops t-values whose x is substantially smaller than their neighbors. The process is controlled by the adjustable parameter r: if $r \ll 1$ all the maxima are preserved, while if r is close to 1 only the largest few maxima in each utterance survive.

B. Acoustic Properties Beyond Loudness

As there is some evidence that acoustic properties other than L contribute to prominence judgments, we investigated a set of alternatives for g in Equation 1. They were:

1. $g_1 = 1$. The resulting timing estimates are based only on L.

2. .

$$g_2(t) = 1 + \eta \cdot V(t) \cdot \left(\frac{f_0(t)}{\langle f_0 \rangle} - 1 \right), \tag{3}$$

where $f_0(t)$ is the speech fundamental frequency, as determined by the `get_f0` program from the ESPS package (Entropic Corp.). $V(t)$ is 1 or 0, indicating whether or not the speech is voiced (it also is produced by `get_f0`), and $\langle f_0 \rangle$ is the average f_0 over the voiced parts of the utterance. Thus, if $\eta > 0$, voiced sounds with relatively high f_0 are emphasized, while if $\eta < 0$, voiced sounds with relatively low f_0 would be treated as louder.

3. .

$$g_3(t) = 1 + \zeta \cdot V(t) \cdot \left| \frac{f_0(t)}{\langle f_0 \rangle} - 1 \right|, \tag{4}$$

with variables as above. Thus, if $\zeta > 0$, voiced sounds with either high or low f_0 are emphasized relative to

those with f_0 near average. Conversely, if $\zeta < 0$, voiced sounds with near-average f_0 or unvoiced sounds would be treated as more important.

4. .
$$g_4(t) = 1 - \alpha A(t), \tag{5}$$

where $A(t)$ is the aperiodicity measure from Kochanski *et al.* (2005). (Related measures have been developed by de Krom, 1993 and Boersma, 1993.) Vowels have small aperiodicities (typically less than 0.5), while fricatives have aperiodicities near 1, so α controls how important fricatives (and other consonants) are in the expression of prominence.

5. .
$$g_5(t) = 1 - \gamma S(t), \tag{6}$$

where S is a measure of the average slope of the speech spectrum. It is related to the ratio of power below 1 kHz to the power above 1 kHz, and is described in detail in Appendix A. Related measures have been developed by Heldner (2001); Sluijter and van Heuven (1996); and references therein.

We define S so that its histogram is approximately centered around zero, and positive S corresponds to excess high-frequency power. Consequently, a positive γ would cause sounds with more high frequency power (like fricatives and harshly spoken vowels) to be treated as louder; on the other hand, sounds like /m/ and gently spoken vowels would be treated as emphasized if $\gamma < 0$.

6. .
$$g_6(t) = g_2(t) \cdot g_4(t) \cdot g_5(t) \tag{7}$$

We conduct an analysis where g is the product of Equations 3, 5 and 6. This allowed for prominence to be determined by an arbitrary combination of voicing irregularity, spectral slope and f_0.

C. Optimizing the Parameters

We based our analysis on the metaphor of coupled oscillators (Large and Kolen, 1994; Saltzman and Byrd, 2000; Port, 2003). Following that metaphor, we constructed a time series of phase for both the tick sequence and for the set of syllable times T produced by the algorithm. The phase is 0, 2π, 4π,... at successive ticks, linearly increasing in between. The metronome phase $\phi(t)$ is then a linear function of time, with a slope equal to 2π divided by the interval. A similar phase function, called $\psi(t)$, is defined from T; it increases by 2π for each element of T. It is defined between the first and last elements of T that fall within the speech.

If the algorithm produced a periodic series of predictions in T, then $\psi(t)$ would also be a straight line. For a regular pattern with ticks and predictions coming at the same average rate, the slopes of ϕ and ψ will be equal, and the difference,

$$\Delta(t) = \phi(t) - \psi(t), \tag{8}$$

will be a constant. At each moment, Δ can be interpreted as a phase difference between the stream of ticks and the stream of predictions in T.

One could use the variance of $\Delta(t)$ as a measure of how well the two sequences are related, but it would be extremely sensitive to missed events. One missing or extra element in T would change Δ by 2π for the remainder of the utterance. Thus a single missed prediction could dramatically increase the variance, as can be seen from the dashed line of Figure 2.

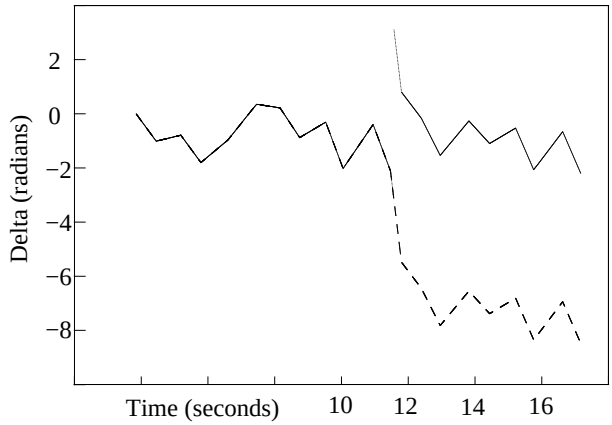

FIG. 2. The phase $\Delta(t)$ for a typical utterance plotted against time. The algorithm missed a prediction near $t = 10.8$ s, then behaved well for the rest of the utterance (dashed line). The solid line shows Δ, wrapped into the range from $-\pi$ to π to simulate the behavior of Equation 9.

Instead, we compute[8]

$$I = \frac{1}{\mathbb{L}} \int \exp(i \cdot \Delta(t)) \cdot dt, \tag{9}$$

where the integral is taken over the part of the utterance where ψ and ϕ are both defined; $\mathbb{L}$ is the integral's length (in seconds).

If the the times in T match the metronome ticks so that $\Delta(t)$ is constant, then the exponential will be a complex number of constant phase and unit magnitude, and the magnitude of I will be unity. This gives the largest possible magnitude for I; any mismatches will reduce it. So, $|I|$ tells how well the two sequences match, and the phase of I describes when the ticks occur relative to T.

Putting $\Delta(t)$ into an exponential makes the analysis relatively insensitive to dropping or adding one prediction. In either case, the argument to the exponential changes by 2π radians during the gap, but the value of the exponential swings around and returns to the value it had before the gap.[9] In such a case, $|I|$ would be reduced from 1 to about 0.9 for our conditions. The solid line in Figure 2 shows the effect of the exponential, wrapping together values of Δ that differ by 2π.

We can interpret I in two related ways. Starting with $I = 1$ as implying perfect correlation, I is reduced whenever there is a missing or extra prediction. Alternatively, we note that when $\mathrm{var}(\Delta)$ is much less than 1, $I \approx 1 - \mathrm{var}(\Delta)$, so that a decrease in I can be interpreted as an increase in the variance of Δ and thus as increased timing errors between the ticks and T.

Now we can compute I for an utterance, given the parameters that control the algorithm. Then, for a corpus, we can compute the average of the magnitude of I, i.e. $\langle |I| \rangle$. $\langle |I| \rangle$ is thus a function of the algorithm's parameters. It is an over-all figure of merit for how well the algorithm's predictions match the ticks.

We then find the optimal parameters for the algorithm by evaluating $\langle |I| \rangle$ for 90,000 randomly chosen combinations of parameters and taking the one that produces the largest absolute value. This simple technique for finding the best parameters was chosen because I, and thus $\langle |I| \rangle$, is a discontinuous function of the algorithm's parameters. (More efficient optimization algorithms are not applicable, as they typically assume that the function to be optimized is continuous, and often that it has continuous first derivatives.)

By optimizing in this way, we are explicitly searching for acoustic properties that repeat with the same periodicity as the metronome ticks. The resulting parameters will be most representative of those utterances with one nominally stressed syllable per metronome tick.

D. Bootstrap Resampling and Confidence Intervals

The above technique is good for finding the parameters that give the best match to the data but it needs to be extended to find error bars and confidence intervals. For this purpose, we use a Bootstrap Resampling scheme (Davison and Hinkley, 1997). We compute 3400 artificial corpora, constructed by choosing utterances from the real corpus, sampling randomly with replacement. In each artificial corpus, most utterances appear once, but some do not appear at all and some appear several times. The same analysis procedure can then repeated for each artificial corpus, leading to optimal parameters for each. In practice, we use a mathematically equivalent but faster technique. We compute the values of I for each combination of utterance and parameters, then the Bootstrap resampling is implemented as a weighted average in the computation of $\langle |I| \rangle$. (Each weight is just equal to the number of times that datum was chosen.)

The resulting distribution of optimal parameters then approximates what one would get by replicating the entire experiment, with new subjects drawn from the same pool. So, if one wishes to estimate the probability that some parameter p exceeds a threshold, X, one can simply count the fraction of artificial corpora where the optimal $p > X$. Thus, we use Bootstrap resampling to generate confidence intervals for $|I|$ and the algorithm's parameters.

E. Timing Data

We processed both the metronome and speech channels with the same algorithm (§III.A). Its use on the speech channel is a major focus of this paper; we used it to find the metronome ticks merely out of convenience. The metronome data is short bursts of oscillation amid silence, and almost any algorithm will be equally successful at finding the ticks.

For the metronome channel, rather than computing L (see §III.A), we used the RMS power above 100 Hz in the metronome output, averaged over a 15 ms window. This signal was then used in Equation 1 in place of L to yield an initial set of tick times. The algorithm's parameters were set by informal experiment to $g = 1$, $a = 0$, $\sigma = 0.110$ s, $b\sigma = 0.015$ s, and $r = 0.5$. The result was inspected and no errors were found.

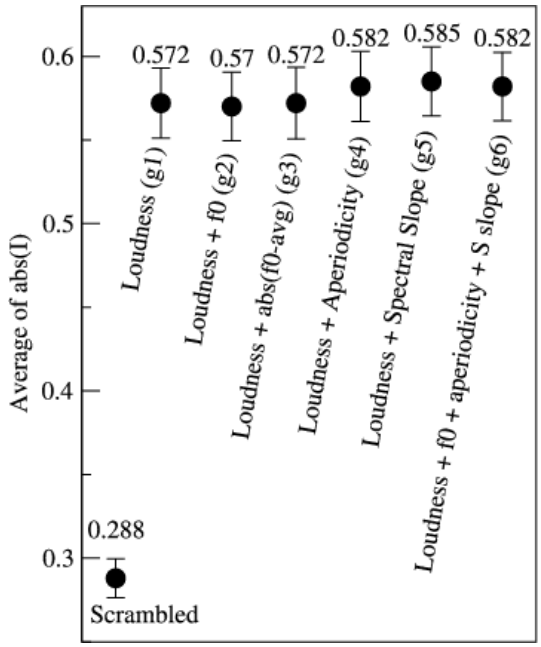

FIG. 3. Values if $|I|$, averaged over the corpus. The lower left measurement is the baseline, where acoustic data are shuffled with respect to ticks. Other conditions correspond to the six cases of §III.B, in order: running the analysis only on L ($g = g_1$), and L enhanced by other acoustical properties ($g = g_{2...6}$).

The start and end points of the speech were automatically determined by an algorithm that finds the borders of a loud interval, surrounded by quiet regions on both edges. It is a modified version of techniques used in (Kochanski *et al.*, 2005). All endpoints were manually checked at the same time we checked the tick marks. Fewer than 10% were adjusted. In total, approximately 3000 syllables (and tick positions) were examined.

IV. RESULTS AND DISCUSSION

A. Relevant acoustic measures

We first confirm that the algorithm performs better when given acoustic data than it would without any data. This is the basic check that it can usefully predict anything. To do this, we first perform the analysis with $g = 1$ and find $\langle |I| \rangle = 0.572 \pm 0.02$. Then, we repeat the analysis, except we shuffle the data so that we use ticks from one utterance and acoustic data from another. Comparing, the shuffled results are substantially poorer than the actual analysis, with $\langle |I| \rangle$ only 0.290 ± 0.013. The difference between normal and shuffled analyses is statistically significant at $z > 9$ or $P \ll 10^{-5}$. Thus, based on L, metronome ticks can be predicted at much better than chance levels. All choices of g show similarly large and significant changes.

Next, we investigate which acoustic properties carry the most information, as shown by an improvement in the tick prediction. Can adding other information to L improve the tick prediction? To do this, we insert different choices of g from §III.B into Equation 1 and see if $\langle |I| \rangle$ increases. The

results are displayed in Figure 3. Changing g to bring in other acoustic data gives little or no improvement over $g = 1$.

Because the analysis with $g = g_6$ involves more parameters, it might be argued that we do not sample densely enough in the higher-dimensional parameter space to find the maximum. While an exhaustive test is not practical, we did check for this possibility by recomputing the $g = g_6$ case with 10 times as many samples ($9 \cdot 10^5$ sets of parameters). This makes only a modest change: $\langle |I| \rangle$ increases by just 0.01. Since the change is small, it suggests that we have indeed sampled parameters close to the maximum even in the higher dimensional $g = g_6$ parameter space, and that little further increase would be expected even with more samples.

Figure 3 shows that L is an important correlate of the ticks, but it would not exclude the possibility that g is of comparable importance to L, if g and L marked the same locations. We can check that by looking at the distribution of α, γ, η and ζ in suitable optimizations. If an acoustic measurement (e.g. f_0) were important, we would expect that the distributions of (e.g.) η (which indicates changes in f_0 - see section IIIB) would be narrow. Assume, for instance, that ticks are marked by high f_0. In that case, we would expect a positive η so that the contributions of f_0 and L would reinforce each other. Running the algorithm with negative η would cause the contributions to cancel. L would then cancel the main effect of f_0 and vice versa, leaving only the fluctuations; one would not then expect the peaks of x to be correlated with either metronome ticks, L or f_0.

Similarly, if f_0 and L were anti-correlated, the same logic would apply and η would be negative. (If η were near zero, it would imply that the best tick predictions are done without use of f_0, which contradicts the hypothesis that f_0 is important.) Either way, if f_0 were important, we would expect that η would have a reasonably narrow distribution, on one side or the other of zero. A broad distribution of η, roughly centered on zero, is thus evidence that the algorithm gets no useful information from f_0 via Equation 3. The same logic applies to α, γ, and ζ.

For α, η, and ζ the distribution is broad and overlaps zero. The optimal values of η for the $g = g_2$ analysis is 0 ± 0.5. This provides additional evidence that f_0 does not usefully contribute to the prediction of metronome ticks (i.e. rhythmic prominence),[10] since the distribution includes $\eta = 0$. The deviation of f_0 from the average ($g = g_3$) is also unimportant, with $\zeta = 0.2 \pm 0.9$: subjects do not reliably mark prominent syllables by pushing f_0 away from the utterance mean. Likewise, the aperiodicity is not important; optimal $\alpha = -0.3 \pm 0.7$.

On the other hand, the spectral slope has some relationship to the ticks. The distribution of optimal values is $\gamma = 0.6 \pm 0.3$, so that distribution overlaps zero only slightly: just 1% of the sets of optimal parameters are negative. So, syllables on the beat have some excess high frequency power. However, it is not a large effect; we estimate that $g_5(t)$ gives the same effect as a 20% change in L (roughly equivalent to a 5 dB change in acoustic power). (Perceptually, a twenty percent changes in loudness is not large.) This is consistent with our result that $\langle |I| \rangle$ does not substantially increase when spectral slope information is added to L (Fig. 3). The small size of the effect is further confirmed by computing the correlation coefficient between $L(t)$ and $S(t)$, the spectral slope measure defined in APPENDIX A: it is just 0.03.

TABLE I. Parameters that yield the largest $\langle |I| \rangle$. The analysis operates on L only ($g = 1$). The right column shows the distribution of values that were tested in the optimization procedure (90,000 samples), and the center column shows the distribution of optimal values that were found (3400 bootstrap corpora).

Parameter	Optimal Value	Distribution of Test Evaluations: mean±stdev
σ	0.176 ± 0.028	Gaussian 0.19 ± 0.07
a	-0.01 ± 0.03	Gaussian 0 ± 0.4
b	0.83 ± 0.07	Uniform on [0.5,1]; thus 0.75 ± 0.16
r	0.04 ± 0.04	Uniform on [0,0.4]; thus 0.20 ± 0.12

The straightforward interpretation of this is just that the specific loudnesses at frequencies over 1 kHz go up more, on the beat, than the specific loudnesses at lower frequencies. Such a shift in the spectral balance for vowels was measured by Glave and Rietveld (1975) and Gauffin and Sundberg (1989).

This weak correlation of a spectral slope measure with metronome ticks is in general agreement with the results of Kochanski *et al.* (2005), though the spectral slope computations differ. A comparison with Sluijter and van Heuven (1996) is not simple; They analyzed dependences on focus (i.e. accent) and stress separately, while we effectively compare +focus,+stress with the neighboring syllables which are either −focus or −stress. Heldner (2001) showed that his measure, called "spectral emphasis" is a good predictor of accent. Our results do not support this, but are not inconsistent, as his spectral emphasis measure is substantially different.

B. How localized is prominence?

Since the primary acoustic marker of rhythmical prominence is L, we now focus on that case ($g = 1$) and describe the remaining parameters. Three parameters (a, b, σ) define the shape of the convolution kernel K, and one (r) controls the rejection of small peaks. Table I shows optimal parameters. The rightmost column shows the distribution of parameters we sampled; in each case, the distribution of optimal parameters can be seen to be narrow and not too far off center, so our choice of sampling distribution is not seriously constraining the distribution of optimal parameters.

We also check that our analysis is not simply detecting the gap between repetitions. This is confirmed by noting that the median spacings between predictions in T (0.67 s for metronome for $g = 1$) is substantially smaller than the median length of a repetition (1.30 s). Thus, we are thus not locked onto a single prediction per repetition; in fact, we are seeing close to one prediction per stress, or two predictions per repetition. Also, we note that the average integral over *uSuS* patterns is essentially the same as that for *SuSu* patterns (0.549 vs. 0.558, not significantly different) while we might expect a substantial difference if predictions were tied to the beginning of each repetition but the metronome ticks followed

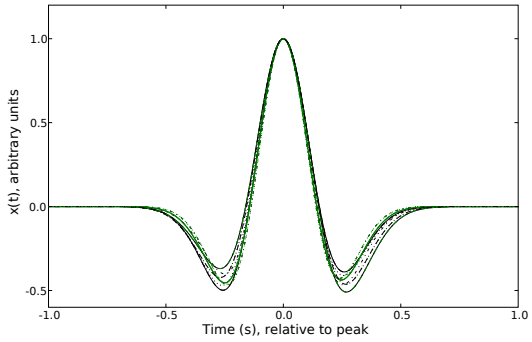

FIG. 4. Convolution kernels, $K(\tau)$, that are optimal for bootstrap samples of the data. The maxima of the curves are aligned at $t = 0$. (These kernels maximize $\langle |I| \rangle$, and have α, γ, and η zero, corresponding to $g = 1$.)

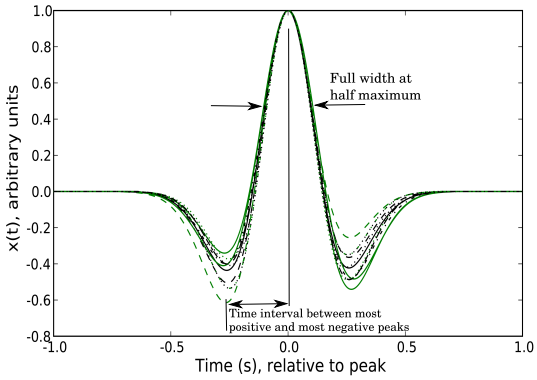

FIG. 5. Optimal bootstrap samples of the convolution kernel, $K(\tau)$ for the $g = g_6$ analysis. The maxima of the curves are aligned at $t = 0$. (These kernels maximize $\langle |I| \rangle$, and allow α, γ, and η to be nonzero.)

the stresses.

The optimal values of r are small, so that the last step of the algorithm is almost moot; it rejects very few candidates. The other three parameters are closely coupled, and can most easily be interpreted visually, by plotting $K(\tau)$. Figure 4 shows a set of convolution kernels that are computed from sets of (a, b, σ) which are optimal parameters for bootstrap corpora. The curves show examples of $K(\tau)$ that are consistent with the data, analyzed assuming $g = 1$. Since the analysis procedure is invariant with respect to shifts of the peaks of x, we can align the peaks of K for clarity without changing $\langle |I| \rangle$. The shape of K is important because it tells us which contrasts are important for predicting a tick. In this case, the contrast is approximately symmetrical: the syllables preceding and following the metronome tick are equally quiet.

Figure 5 shows convolution kernels that are optimal for the $\langle |I| \rangle$, $g = g_6$ analysis. The analysis is entirely consistent with Figure 4, but with somewhat larger scatter as we are fitting a more complex model to the data.

The widths and relative spacings of the peaks are remarkably consistent. Values for σ which is related to the overall width of the kernel are in Table I. The full-width at half-maximum for the highest peak in K is 0.184 ± 0.008 s, and the magnitude of the time interval between the most positive and most negative peaks is 0.31 ± 0.04 s for the $g = 1$ analysis, so the entire window of relevant speech data spans approximately 800 milliseconds. The numbers are similar for the $g = g_6$ analysis, 0.183 ± 0.01 for the FWHM and 0.26 ± 0.01 for the peak spacing.

The positive peaks of the kernel are as wide or wider than a typical vowel, and about two-thirds of the mean syllable spacing, 0.27 s. The full extent of the kernel, given by $2\sigma \approx 0.34$ s, is wider than the mean syllable spacing, and the time interval between the positive and negative peaks is just about equal to the mean syllable spacing. Ticks are thus correlated with the properties of a region larger than a single syllable. Our analysis is consistent with the hypothesis that ticks are related to a large loudness contrast between a syllable and its nearest neighbor(s).

This conclusion is consistent with findings by Kochanski *et al.* (2005), who show an (approximately symmetric) loudness pattern around syllables that were judged to be prominent. The loudness pattern observed there was somewhat narrower, as might be expected, given the faster speech in that study. There are some differences in detail, however. Kochanski *et al.* (2005) reported small but significant correlations of f_0 and A with prominence; but we see none here. The difference may be due to the different tasks (e.g. production vs. perception).

An interesting feature of the these results is that the algorithm unifies loudness and duration changes into a single time series, $x(t)$. For vowel durations shorter than κ's positive peak, an increase in duration has much the same effect as an increase in loudness; the convolution can be approximated as an integral of loudness over about 0.2 s. It can then be further approximated as the vowel's loudness times the vowel duration. Thus, peaks of $x(t)$ and consequently predictions in T will tend to occur on longer syllables, rather than shorter ones.

An analogous effect, but occuring in speech perception rather than production, can be found in the psychophysics literature. Munson (1947) showed that perceived loudness is a generally increasing function of duration, and Plomp and Bouman (1959) modeled the effect as convolution of the specific loudness with a kernel. They obtained equivalent widths of their kernel near 0.25 s. This is sufficiently close to our results to support the idea that speech production should be matched to speech perception; the peak value of our $x(t)$ may simply be related to the perceived loudness of the syllable.

Our results are similar to those of Beckman (1986), who found strong correlations of prominence with a similar combination of amplitude and duration, and Silipo and Greenberg (1999, 2000), who had the best success at predicting prominence with a product of syllable-averaged amplitude and vowel duration. It also parallels (on the production side) at least one claim of Turk and Sawusch (1996) – viz that duration and loudness are perceived together as a single percept.

This agreement with other work is perhaps remarkable, given the differences in experimental technique. Not least because the experiments mentioned above involve perceptual judgments (e.g. which syllables are prominent), while our

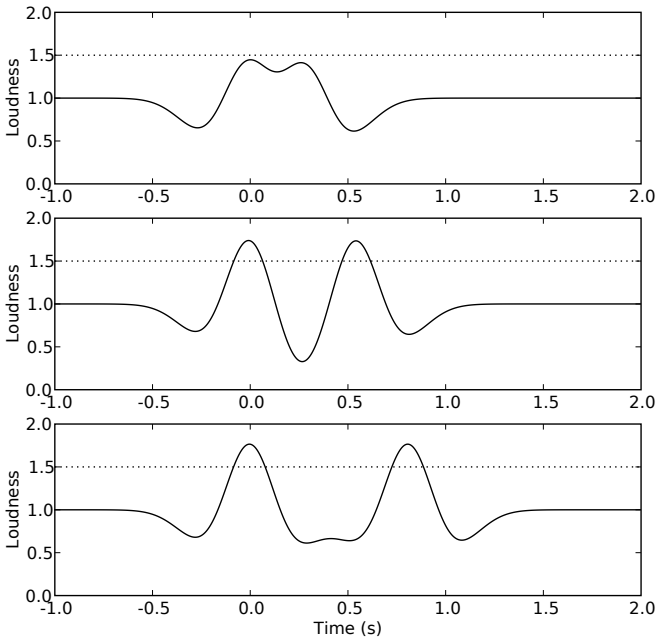

FIG. 6. Theoretical loudness contours based on this work for prominences that are (top) on adjacent syllables, (middle) separated by one and (bottom) by two syllables. The dashed line is a loudness reference.

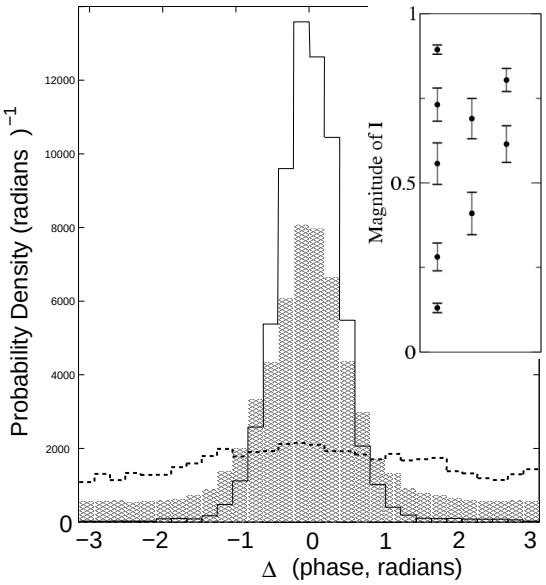

FIG. 7. Phase histograms (Δ, relative to the phase of I) for three different subjects. The subjects shown have (reading from top to bottom, in the center) the largest, median, and smallest value of $\langle|I|\rangle$. The subfigure shows values of $\langle|I|\rangle$ for each subject, with 1-standard deviation error bars on the average. (The horizontal axis in the subfigure has no meaning – it just separates subjects.)

experiment is strictly a production experiment.

C. Alternating Metrical Patterns and Implications for Phonology

This result that prominence is expressed by a contrast between a syllable and its neighbors is interesting because it may provide a reductionist explanation of the alternating metrical patterns that are common in many languages.

Consider a loudness pattern like Figure 4, and place it on a uniform background corresponding to the average loudness of speech. For the sake of argument, suppose that loudness patterns of different syllables add. Then, Figure 6 shows the resulting loudness patterns for prominences that are separated by 1, 2, and 3 syllables.

One can see that the loudness patterns interfere with one another when the prominent syllables are adjacent, and the resulting loudness peaks are then not as dramatic as when the syllables are farther apart. If, hypothetically, the listener had a loudness threshold for the perception of prominence, the case of adjacent syllables would not be perceived as prominent.

To make the adjacent case appear prominent, the speaker would have to make those syllables unusually loud, and/or to push the syllables farther apart so that the loudness patterns would not interfere so strongly. We suggest that speakers may avoid this case because of the extra effort and complexity needed to ensure that a listener will perceive the correct prominences. Over time, this avoidance may become enshrined in phonological rules that reduce the number of adjacent stresses. If speakers avoid 1-syllable feet, the next simplest foot is 2-syllable and this potentially might help explain the cross-linguistically common preference for an alternating metrical pattern.

D. Performance variation between subjects

Subjects differ substantially in terms of how consistently they follow the metronome. This performance variability is seen in Figure 7, which shows the phase histograms for three speakers. Histograms for the other speakers are similar, displaying the same properties: large $\langle|I|\rangle$ is correlated with a tall, narrow peak and a low background level, while small $\langle|I|\rangle$ implies a short broad peak sitting on a higher background.

From subject to subject, $\langle|I|\rangle$ varies from 0.13 to 0.90, while the uncertainty in each subject's mean due to intra-subject variation is just 0.05. This is a large range, as the possible range of $\langle|I|\rangle$ is just from 0 to 1.

ANOVA rejects the hypothesis that the subject means are equal with $P < 10^{-6}$ ($F(9,214) = 7.5$). Some subjects are therefore speaking reliably along with the metronome, while the speech of some others has little consistent relationship to the ticks. By observation, the subjects with the lower values of $\langle|I|\rangle$ show a mixture of two problems: irregular pauses that lead to jumps in Δ, and speech that is simply not synchronized with the metronome rate, leading to a gradual drift in Δ.

For instance, if we select utterances that were spoken with approximately 1 tick per stressed syllable (i.e. 19–21 metronome ticks within the 10 repetitions or 0.95–1.05 ticks per stressed syllable), we compute $\langle|I|\rangle = 0.69 \pm 0.28$ for those utterances ($n = 119$). This contrasts sharply with $\langle|I|\rangle =$

0.25 ± 0.2 for utterances which are apparently unsynchronized with the metronome, having 0.65–0.85 ticks per stress ($n = 44$). The three subjects who have the largest number of these unsynchronized utterances are also the three subjects with the smallest values of $\langle |I| \rangle$.

Because of the large range of subject performance, and because it can be automated, this task may be useful as a measure of the ability to process metrical patterns. One possible application is a evaluation tool for stuttering (e.g. Boutsen *et al.*, 2000 and Cooper and Allen, 1977, showed that stutterers had much larger timing variance in repetitive speech than normal subjects). The wide range of inter-subject performances that we have observed (c.f. Figure 7) parallels the results of Cooper and Allen (1977) for normal subjects, who found that some subjects had timing variances roughly ten times larger than others.

E. Analysis of high-performing subjects

Our analysis is designed on the assumption that metronome ticks are proxies for prominence. But, ticks are clearly not perfect proxies, especially for some subjects. This raises the possibility that our results are affected by utterances that are unsynchronized with the metronome or that have other synchronizations (e.g. two ticks per prominence).

The majority of the utterances were spoken with approximately a 1:1 ratio between metronome ticks and nominal stresses: 56% of the utterances contained 19–21 metronome ticks for their 20 nominally stressed syllables. A total of 18% were near other small-integer ratios that might suggest different patterns of synchronization: 5% contained 9–11 ticks (near a 1:2 ratio), 11% contained 14–16 (near a 2:3 ratio, but informal investigation shows that many of these utterances have no obvious synchronization), 1% contained 29–31 (3:2) and 1% contained 39–41 ticks (2:1). The ratios of the remaining 26% do not suggest synchronization between the speech and the metronome.

To check that possibility, we repeated our analysis on the subjects that gave us the five largest values of $\langle |I| \rangle$. These subjects generally produced utterances that are well synchronized, with one prominence per metronome tick, and thus are particularly well adapted to our analysis. By comparing the analysis on this subset to our main results (§IV.A), we can check that our results are robust.

As expected, the value of $\langle |I| \rangle$ for the shuffled analysis is statistically indistinguishable from the shuffled analysis for the full data set. Also, as expected, the values for the $g_{1\ldots6}$ analyses have increased: for the g_1 analysis, $|I|$ has increased from 0.572 ± 0.02 to 0.739 ± 0.02. Even so, there are no differences among the various $|I|$ values for the $g_{1\ldots6}$ analyses that are larger than the corresponding standard deviations. This supports our contention that the other acoustic properties are not very important.

Likewise, α, η, and ζ still have distributions that strongly overlap zero, providing further confirmation that f_0 does not play a role in the senses of Equations 3 and 4, and that $A(t)$ is likewise unimportant. The spectral slope remains weakly important, with just 4% of the optimal parameter sets having $\gamma < 0$.

The shapes of κ do not change substantially from those shown in Figure 5; neither σ, a, b, or r show changes larger than the error bars shown in Table I. This supports our use of the metronome tick as a proxy for rhythmic prominences.

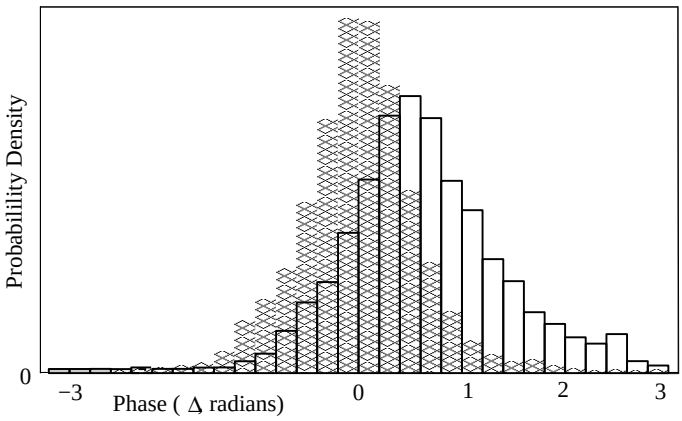

FIG. 8. Phase histogram (Δ) for a typical speaker (outline) and the histogram of Δ relative to the average phase of each utterance (filled). The peak of the dashed histogram shows the typical phase relationship between metronome ticks and the algorithm's predictions for that subject. The width of the histograms show timing inconsistencies between the subject's speech and the metronome.

F. Phase variation between utterances

The analysis so far only considers how stable the predictions (T) are, in relation to the ticks. Any uniform phase shift between the ticks and T will just change the phase of I but not its magnitude, so $|I|$ and $\langle |I| \rangle$ are insensitive to the average phase relation within an utterance. Our analysis essentially minimizes the variance of the difference between the predicted and actual tick positions. Thus, our analysis differs from Allen (1972), who assumes that the moment when the tick happens is the critical part of the syllable.

However, because the optimal κ happens to be symmetrical and fairly compact, we can use it to identify the point in each phrase where $x(t)$ is maximal. The loudness contour in a region around this point gives the most consistent prediction of the metronome ticks. We computed this for $g = 1$ and optimized parameters; peaks in $x(t)$ thus correspond fairly accurately to peaks in $L(t)$.

Figure 8 shows the histogram of Δ for all utterances produced by one subject. It also displays the histogram of Δ relative to the phase of I for the corresponding utterance. We display data from the subject who had the median value of $\langle |I| \rangle$. The histogram of relative phase is noticeably narrower and taller, as might be expected, indicating that different utterances have somewhat different alignments between the speech and the metronome.

For this subject, the peak of the Δ histogram (outline) is not at zero, indicating that the peak of $x(t)$ is not aligned with the metronome ticks, but that the metronome ticks occur somewhat before the peaks of $x(t)$. In other words, this subject speaks with the ticks early in the syllable, before the vowel's loudness peak. However, subjects differ in their average alignment.

Figure 9 shows a vector plot of the phases per utterance,

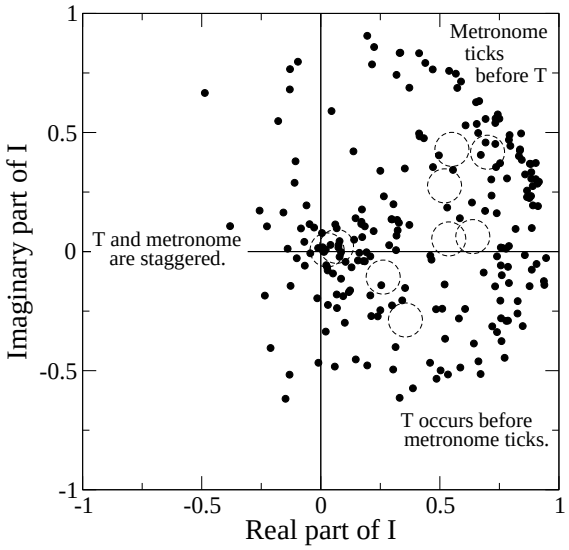

FIG. 9. The figure shows the phase relationship between the metronome ticks and peaks of $x(t)$ (loudness, loosely speaking). Each utterance (i.e. 10 repetitions of one text by one speaker) is represented by a dot at the (complex) value of I, with the real part on the horizontal axis and the imaginary part on the vertical axis. The distance from the origin is thus proportional to $|I|$ for that utterance. Dots near the origin represent utterances that did not have a consistent phase relationship between loudness and the metronome; dots on the unit circle would have a perfectly consistent phase relationship. The angle of the point, when viewed from the origin matches the phase of I, and dots just to the right of the origin come from utterances where the peak in $x(t)$ is aligned with the metronome ticks. Dashed circles represent the average of each subject's utterances.

for all subjects. The phases are distributed around zero, indicating that on average, subjects align the peaks of $x(t)$ (and thus $L(t)$) with the metronome ticks. If we wished to discuss a "pr(oduction)-center" in analogy to Allen (1972)'s p(erceptual)-centers, the pr-center appears to be near the syllable's peak in L: the mean phase of the peak is 13 ± 44 degrees before the tick. The concept of "pr-centers" has been discussed by Morton J. Morton (1976) who has argued that P-centers are ambiguous and could possibly be exerting an influence on the production as well as the perception of words. However, there is substantial inter-subject variation.[11] In the figure, the dashed circles show averages of I for each subject. Apart from the two that are near zero and thus have have no consistent phase relationship to the ticks, some subjects place the loudness peak before the tick, some almost on the tick, and some place it after the tick.

V. CONCLUSIONS

In English, the beat of repetitive speech is marked by an increase in the loudness relative to the immediate neighborhood. This neighborhood is approximately one syllable on either side of the beat, at the speech rates we studied.

The critical factor appears to be the average loudness over an approximately 360 millisecond interval, so that as the vowel is shortened below 200-300 milliseconds, the duration reduction will play the same role as a reduction in loudness. Other acoustical properties such as f_0 are not strongly correlated with the beat; one exception is that on average, speakers produce somewhat more high frequency power on the beat than off.

We suggest that this preference for loudness contrasts as a marker of the beat may provide a partial explanation for the relative rarity of adjacent, prominent syllables. We note also that the width of the region over which prominence is expressed is well matched to human auditory perception, as might be expected.

We conducted a production experiment that had minimal involvement of speech perception, in contrast to much prior work on prominence. Despite that our results are in general agreement with prior experiments. Prominence seems a broadly useful idea, one that is shared both by linguists and naive speakers of English, and one that has a straightforward relationship to measureable acoustic quantities.

Acknowledgments

_We thank Burton Rosner for discussions and comments and 'Oiwi Parker Jones for comments. This research was funded by the UK's Economic and Social Research Council via grant RES-000-23-1094.

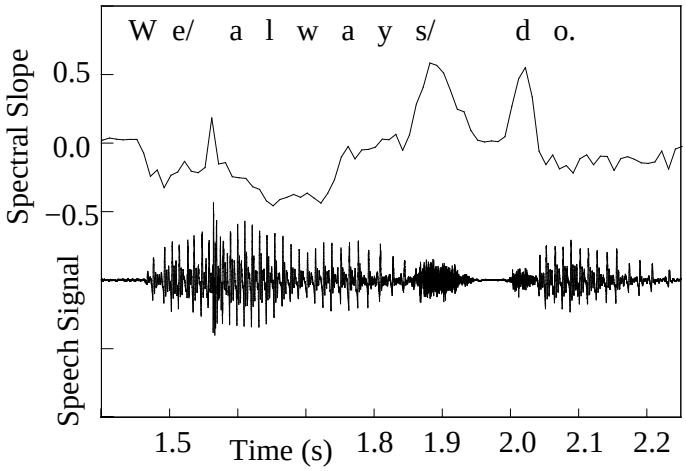

FIG. 10. Sample audio data along with the spectral slope, $S(t)$. This shows one repetition of "We always do."

APPENDIX A: DEFINITION OF SPECTRAL SLOPE MEASURE

The spectral slope measure we use is based on a comparison of specific loudnesses at high and low frequencies. The computation begins by filtering the speech into bands that are 1 Erb (see Moore and Glasberg, 1983) wide, on $1/2$-Erb centers, between 50 and 6000 Hz. (We use the filterbank recommended by Baumgarte (2002).) Then, we half-wave rectify each filter output and take the cube root of the local intensity to approximate the nonlinear response of the ear. This yields a set of specific loudness measures, $\ell_k(t)$, where k indexes the frequency band. The center frequency of the k^{th} band is $f(k)$ Hz.

Next, we low-pass filter these with a frequency-dependent time constant derived from Plomp and Bouman (1959) to yield $\tilde{\ell}_k(t)$. We approximate their time constants as

$$\tau(f) = 0.4 - 0.27\frac{\log(f/200)}{\log(10^4/200)}, \tag{A1}$$

where $\tau(f)$ is in seconds.

We sum the result into two broad-band specific loudnesses,

$$L^H(t) = \sum_k \tilde{\ell}_k(t) \text{ where } f(k) > 1000 \text{ Hz}, \tag{A2}$$

and

$$L^L(t) = \sum_k \tilde{\ell}_k(t) \text{ where } f(k) < 1000 \text{ Hz}. \tag{A3}$$

Finally, we combine L^L and L^H into a spectral slope measure by computing

$$S(t) = \frac{cL^H(t) - L^L(t)}{\hat{s}}, \tag{A4}$$

where $c = 2.8$ is a constant chosen so that the average of $S(t)$ is approximately zero, and $\hat{s}$ (defined below) is a factor to normalize out the average loudness. This form for $S(t)$ has the advantage that it is well-behaved in and near pauses, unlike measures based on intensity ratios of two frequency bands (Boersma, 1993; de Krom, 1993; Sluijter and van Heuven, 1996; Streefkerk *et al.*, 1999; Kochanski *et al.*, 2005). Also, it is well-behaved in regions of uncertain or no voicing, unlike the measure used by Heldner (2001).

We use

$$\hat{s} = c\frac{\sum_t s^2(t)}{\sum_t s(t)}, \tag{A5}$$

where $s(t) = cL^H(t) + L^L(t)$ and the sum over time is computed over the entire data file containing the utterance. (Typically, the data file contains about 20% silence, counting pauses between repetitions and a few hundred milliseconds at each end. We chose this form for $\hat{s}$ because it relatively insensitive to the amount of silence in the data file.) An example of $S(t)$ is shown in Figure 10.

APPENDIX B: LIST OF PHRASES

Exactly so.	Another time.
We always do.	It surely is.
You never will.	I understand.
Upon my word.	She goes away.
They lie apart.	The business here.
Indeed it had.	I noticed that.
Nothing matters.	Checker players.
William Roper.	Scarlet Letter.
Goodness gracious!	Lightning presses.
Worldly wisdom.	Banking systems.
Phosphorescence.	None whatever.
Good beginning.	This is funny.
Testing the instriments	Philip was conquered
Love and integrity.	Spare me your history.
Let us experiment.	Notes on the editing.
Like you suggested.	Here are the banners.
Not to my knowledge.	Run for the cellar
Not that it mattered.	Talking of wandering.
Wendy was scandalised.	Going away.
Always ahead.	Then I refuse.
Nothing at all.	Open the door.
Maybe she had.	Probably not.
Only a mouse.	Say it again.
Billy was there.	Under the desk.
Carrots and greens.	Shirley declared.
That ruined me.	Freddy Observed.
Tracy announced.	

ENDNOTES

1. As discussed in (Kochanski, 2006), it is not obvious to what extent people have accurate conscious access to the process of speech perception and understanding. For instance, several papers have raised the possibility (Dyde and Milner, 2002; Haffenden and Goodale, 1998; Aglioti *et al.*, 1995) that visual processing for perception is substantially different than processing for action. If so, reported perceptions could disagree on the actions we take in response to the same stimuli. If a similar effect occurs in speech, our reports of what we hear might not reflect our conversational behavior.

Hickok and Poeppel (2000) argue for a similar distinction between tasks that require sub-lexical awareness and those that do not.

2. Part of this data was intended for a similar finger-tapping experiment, but we dropped its analysis when we found that the finger-taps were too loud in the speech channel. However, the tapping parts of the experiment nicely serve the purpose of training the subjects to associate the beat with prominence.

3. One male was accidentally given *uuSuuS* phrases instead of *SuuS*; that data was included in the analysis as we did not find any strong differences between metrical patterns. We don't include those phrases in the appendix because they were only used once.

4. Unisyn is a software package for predicting pronunciation in a variety of English dialects. It is based on a lexicon and includes transformation rules.

5. This part of the data was designed to check that the repetitive speech as not too different from a more typical laboratory experiment. See (Kochanski and Orphanidou, 2007) and the discussion in §I.

6. Note that calling $L(t)$ loudness as was done by Kochanski *et al.* (2005) is somewhat of a misnomer. $L(t)$ is closer to the sum of all the specific loudnesses at each moment, to use Zwicker (1977)'s terminology. Munson (1947) and Plomp and Bouman (1959) showed that perceived loudness is an average of the specific loudness over approximately a 250 ms interval, while the Kochanski *et al.* (2005) measure averages only over a 25 ms (FWHM) window. Averaging $L(t)$ over a 250 ms window should approximate perceived loudness.

7. The convolution $a * b$ is the integral $\int a(\tau)b(t-\tau)\,d\tau$. It can be thought of as taking a time-series b, and filtering it in a way specified by function a. If this were a perception experiment, this equation would correspond to Munson (1947)'s sensation integral.

8. Recall that $\exp(i\cdot\Delta) = \cos(\Delta) + i\cdot\sin(\Delta)$, so that it is periodic, not rapidly increasing as one would get by taking $\exp()$ of a real argument. If one plots the real and imaginary parts of $\exp(i\cdot\Delta)$, one finds that they trace out a unit circle, centered at zero.

9. We rely on the fact that $\exp(i\Delta) = \exp(i(\Delta + 2\pi))$ for any Δ.

10. We recognize that f_0 *can* induce the perception of prominence, as shown by Gussenhoven *et al.* (1997); Rietveld and Gussenhoven (1985); Terken (1991). However, as shown in Kochanski *et al.* (2005), f_0 swings in speech are frequently not large enough to induce a prominence judgment. Consequently, f_0 is potentially important to prominence, but apparently not important in practice.

11. Unfortunately a direct comparison to the p-center location is not practical in this experiment.

Aglioti, S., DeSouza, J. F., and Goodale, M. A. (**1995**). "Size-contrast illusions decieve the eye but not the hand", Current Biology **5**, 579–685.

Allen, G. (**1972**). "The location of rhythmic stress beats in English: An experimental strudy I.", Language and Speech **15**, 72–100.

Baumgarte, F. (**2002**). "Improved audio coding using a psychoacoustic model based on a cochlear filter bank", IEEE Transactions on Speech and Audio Processing **10**, 495–503.

Beckman, M. E. (**1986**). *Stress and Non-Stress Accent*, volume 7 of *Netherlands Phonetic Archive* (Dordrecht : Foris).

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Institute of Phonetic Sciences, University of Amsterdam, Proceedings **17**, 97–110, URL `http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf`, uRL viewed 11/2006.

Boutsen, F. R., Brutten, G. J., and Watts, C. R. (**2000**). "Timing and intensity variability in the metronomic speech of stuttering and nonstuttering speakers", J. Speech Language Hearing Research **43**, 513–520.

Brenier, J. M., Cer, D., and Jurafsky, D. (**2005**). "The detection of emphatic words using acoustic and lexical words", in *Proceedings of the 9th European Conference on Speech Communication (EUROSPEECH-05)* (International Speech Communications Association), lisbon, Portugal, 4-8 September 2005.

Cooper, M. H. and Allen, G. D. (**1977**). "Timing control accuracy in normal speakers and stutterers", J. Speech and Hearing Research **20**, 55–71.

Cummins, F. and Port, R. (**1998**). "Rhythmic constraints on stress timing in english", J. Phonetics **26**, 145–171.

Davison, A. C. and Hinkley, D. V. (**1997**). *Bootstrap Methods and their Application* (Cambridge University Press).

de Krom, G. (**1993**). "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", J. Speech and Hearing Research **36**, 254–266.

Dyde, R. T. and Milner, D. A. (**2002**). "Two illusions of perceived orientation: one fools all of the people some of the time; the other fools all of the people all of the time", Experimental Brain Research **144**, 518–527.

Fear, B. D., Cutler, A., and Butterfield, S. (**1995**). "The strong/weak syllable distinction in English", J. Acoustical Society of America **97**, 1893–1904.

Fitt, S. (**2002**). "Unisyn lexicon release", URL `http://www.cstr.ed.ac.uk/projects/unisyn/`, UNISYN_1_1, Downloaded 4/2006.

Fitt, S. and Isard, S. (**1999**). "Synthesis of regional English using a keyword lexicon", in *Proceedings: Eurospeech 99*, volume 2, 823–826 (International Speech Communications Association), URL `http://homepages.inf.ed.ac.uk/sue/Publications/Fitt_1999_a.ps`, URL viewed 11/2006.

Fowler, C. A. (**1979**). ""perceptual centers" in speech production and perception", Perception and Psychophysics **25**, 375–388.

Fry, D. B. (**1955**). "Duration and intensity as physical correlates of linguistic stress", J. Acoustical Society of America **27**, 765–768.

Fry, D. B. (**1958**). "Experiments in the perception of stress", Language and Speech **1**, 126–152.

Gauffin, J. and Sundberg, J. (**1989**). "Spectral correlates of glottal voice source waveform characteristics", J. Speech Hearing Research **32**, 556–565.

Glave, R. D. and Rietveld, A. C. M. (**1975**). "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?", J. Acoustical Society of America **58**, 875–879.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., and Terken, J. (**1997**). "The perceptual prominence of fundamental frequency peaks", J. Acoustical Society of America **102**, 3009–3022.

Haffenden, A. M. and Goodale, M. A. (**1998**). "The effect of pictorial

illusion on prehension and perception", J. Cognitive Neuroscience **10**, 122–136.
Hart, M. S. and volunteers (**2006**). "Project gutenbert", URL www.gutenberg.org, project Gutenberg; URL viewed 4/2006.
Heldner, M. (**2001**). "Spectral emphasis as an additional source of information in accent detection", in *Prosody in Speech Recognition and Understanding*, paper #10, October 22-24, Molly Pitcher Inn, Red Bank, NJ, USA.
Hickok, G. and Poeppel, D. (**2000**). "Toward a functional neuroanatomy of speech perception", Trends in Cognitive Sciences **4**.
J. Morton, S. Marcus, C. F. (**1976**). "Perceptual centers (p-centers)", Psychological Review **83**, 405–408.
Klapuri, A. P., Eronen, A. J., and Astola, J. T. (**2006**). "Analysis of the meter of acoustical musical signals", IEEE Trans. Audio, Speech, and Language Processing **14**, 342–355.
Kochanski, G. (**2006**). "Prosody beyond fundamental frequency", in *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer, Language, Context and Cognition, 89–122 (Walter de Gruyter, Berlin, New York).
Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (**2005**). "Loudness predicts prominence: Fundamental frequency lends little", J. Acoustical Society of America **118**, 1038–1054.
Kochanski, G. and Orphanidou, C. (**2007**). "Testing the ecological validity of speech", in *Proceedings of the 16th International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry, URL http://kochanski.org/gpk/papers/2007/icphs.pdf, conference website at http://www.icphs.de viewed 9/2007.
Large, E. and Kolen, J. F. (**1994**). "Resonance and the perception of musical meter", Connection Science 177–208.
Large, E. W. and Palmer, C. (**2002**). "Perceiving temporal regularity in music", Cognitive Science **26**, 1–37.
Lehiste, I. (**1973**). "Rhythmic units and syntactic units in production and perception", J. Acoustical Society of America **54**, 1228–1234.
Lieberman, P. (**1960**). "Some acoustic correlates of word stress in American English", J. Acoustical Society of America **32**, 451–454.
Moore, B. C. J. and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acoustical Society of America **74**, 750–753.
Munson, W. A. (**1947**). "The growth of auditory sensation", J. Acoustical Society of America **19**, 734–735.
Nesbitt, K. (**2001**). *The Aliens Have Landed at Our School!* (Meadowbrook Press, Minnetonka, MN).
Plomp, R. and Bouman, M. A. (**1959**). "Relation between hearing threshold and duration for tone pulses", J. Acoustical Society of America **31**, 749–758.
Port, R. F. (**2003**). "Meter and speech", J. Phonetics **31**, 599–611.
Rietveld, A. C. M. and Gussenhoven, C. (**1985**). "On the relation between pitch excursions and prominence", J. Phonetics **13**, 299–308.
Saltzman, E. and Byrd, D. (**2000**). "Task-dynamics of gestural timing: Phase windows and multifrequency rhythms.", Human Movement Science **19**, 499–526.
Scheirer, E. D. (**1998**). "Tempo and beat analysis of acoustic musical signals", J. Acoustical Society America **103**, 588–601.
Silipo, R. and Greenberg, S. (**1999**). "Automatic transcription of prosodic stress for spontaneous English discourse", in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS99)*, edited by J. J. Olds, Y. Hasegawa, M. Ohala, and A. C. Bailey, 2351–2354 (The Regents of the University of California).
Silipo, R. and Greenberg, S. (**2000**). "Prosodic stress revisited: Reassessing the role of fundamental frequency", in *Proceedings of the NIST Speech Transcription Workshop*.
Sluijter, A. M. C. and van Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress", J. Acoustical Society of America **100**, 2471–2485.
Stevens, S. S. (**1971**). "Perceived level of noise by Mark VII and decibels", J. Acoustical Society of America **51**, 575–602.
Streefkerk, B. M., Pols, L. C. W., and ten Bosch, L. F. M. (**1999**). "Acoustical features used as predictors for prominence in read aloud Dutch sentences used in ANNs", in *Proceedings of the 6th European Conference on Speech Communication (EUROSPEECH-99)*, volume 1, 551–554 (International Speech Communication Association), URL http://citeseer.ist.psu.edu/streefkerk99acoustical.html, uRL viewed 5/2007.
Terken, J. (**1991**). "Fundamental frequency and percieved prominence of accented syllables", J. Acoustical Society of America **89**, 1768–1776.
Todd, N. P. M. and Brown, G. J. (**1996**). "Visualization of rhythm, time and meter", Artificial Intelligence Review **10**, 253–273.
Turk, A. E. and Sawusch, J. R. (**1996**). "The processing of duration and intensity cues to prominence", J. Acoustical Society of America **99**, 3782–3790, URL doi:10.1121/1.414995.
Zwicker, E. (**1977**). "Procedure for calculating loudness of temporally variable sounds", J. Acoustical Society of America **62**, 675–682.