Loudness predicts prominence: fundamental frequency lends little.

G. Kochanski and E. Grabe and J. Coleman and B. Rosner

(2005/06/07 13:50:33 UTC)

Running title: Fundamental Frequency Lends Little Prominence

The University of Oxford Phonetics Laboratory, 41 Wellington Square, Oxford OX2 2JF, United Kingdom

Submitted to J. Acoustical Society of America, November 2004. Until publication, it may be referred to as `http://kochanski.org/gpk/papers/2005/04pnp.pdf` . G. Kochanski may be contacted at `gpk@kochanski.org` .

ABSTRACT

We explored a database covering seven dialects of British and Irish English and three different styles of speech to find acoustic correlates of prominence. We built classifiers, trained the classifiers on human prominence/non-prominence judgements, and then evaluated how well they behaved. The classifiers operate on 452 ms windows centered on syllables, using different acoustic measures. By comparing the performance of classifiers based on different measures, we can learn how prominence is expressed in speech. Contrary to textbooks and common assumption, fundamental frequency (f_0) played a minor role in distinguishing prominent syllables from the rest of the utterance. Instead, speakers primarily marked prominence with patterns of loudness and duration. Two other acoustic measures that we examined also played a minor role, comparable to f_0. All dialects and speaking styles studied here share a common definition of prominence. The result is robust to differences in labeling practice and the dialect of the labeler.

PACS numbers: 43.70.Fq, 43.66.-x, 43.71.-k

I. INTRODUCTION AND BACKGROUND

In English, some syllables are special and more important, and others less so. The important ones are described, variously, as bearing "stress accents" (Beckman, 1986), as "prominent", or by other terms, but a definition strictly in terms of their acoustic properties has been lacking.

Our central question is the following: Using acoustic data, what property allows the best machine replication of the prominence judgements of human listeners? The experiments here focus on acoustic cues in a window that includes the syllable under consideration and the neighboring syllables. We explore seven dialects and three different styles of speech: lists of sentences, story paragraphs, and a retelling of a story.

Many people have looked at cues to prominence, and they have reached a variety of answers. Passy (1891, pp. 41–42), Passy (1906, p. 27), Sweet (1906, p. 47), Trager and Smith (1951), and others impressionistically described English prosody in terms of "force" or "accent" (equated to loudness), and "intonation" (equated to pitch). Fry (1955, 1958) did early perceptual studies on minimal pairs of synthesized English words that are distinguished by a difference of stress placement (e.g. súbject *vs.* subjéct). He found that the more prominent syllable was marked, in decreasing order of importance, by duration, f_0, and amplitude. His results have achieved wide currency in the linguistic community, despite the study's limitation to single, isolated words.

Other experiments with careful, "laboratory" speech have yielded a variety of results. Lieberman (1960), for instance, described a very early system for deducing lexical stress from acoustics. His work indicates that f_0, amplitude, and duration, are similarly important and that each individually is a good predictor of prominence. However his exceptionally good classification probabilities are due to the explicit selection of clearly enunciated and unambiguous speech: utterances were used as stimuli only when four human judgements all agreed on the stress placement.

Likewise, synthesis studies (discussed in §III.E) showed that f_0 bumps can induce the perception of prominence (Gussenhoven *et al.*, 1997; Rietveld and Gussenhoven, 1985; Terken, 1991).

Other laboratory work is often taken to support the importance of f_0. For instance, Cooper *et al.* (1985) and Eady and Cooper (1986) found significantly different f_0 patterns in a sentence as a function of the focus position (roughly, the pattern of prominences). However, this result needs to be interpreted carefully. These papers reported statistically significant changes to the average f_0 of a group of utterances. While this is useful from a descriptive point of view, the usual listener only hears only one utterance at a time and does not have the luxury of averaging several repetitions before responding. Consequently, while averages of two classes may be significantly different, the distributions of individual measurements may overlap enough so that a listener could not usefully decide what has been said, based on a single utterance.

On the other hand, Beckman (1986) saw substantial correlations of prominence with a combination of amplitude and duration. Turk and Sawusch (1996) have conducted synthesis experiments, comparing isolated instances of (e.g. máma *vs.* mamá). to tease apart the relative importance of loudness and duration to perception judgements. They come to two main conclusions. The first is that these two acoustic measures were perceived together as a single percept; the second is that loudness made a negligible contribution to the results of their rating scale experiment.

Tamburini (2003) has had success with a prominence detection system for more natural speech that assumes an important role for amplitude contrasts between neighboring syllables. However, he did not measure what the differences were between prominent and non-prominent syllables; he simply reported that a particular algorithm achieved 80% correct classification on a corpus of American English.

Another system for automated prominence transcription, built by Silipo and Greenberg (1999, 2000) was tested on an American English corpus with several plausible acoustic correlates of prominence. This study was a first attempt to understand prominence of natural speech, as opposed to careful laboratory speech, although there is not a complete published description of the experiment. In their work, f_0 was shown to have relatively little importance. Comparisons to this work are difficult in that their system had strong assumptions wired in (which we test instead of assuming). For instance, they assumed that f_0 induced prominence only through a single f_0 contour: a symmetrical bump. However, they achieved good performance (≈80% correct classification) by operating their system on the product of syllable-averaged amplitude and vowel duration, which suggests that amplitude and duration are good indicators of prominence. The strong assumptions built into the classifier mean that little can be said about their other, less successful combinations of acoustic features.

In summary, the literature is not completely clear on what acoustic properties of speech communicate prominence, but f_0 is not the complete story. Nevertheless, much work on intonation and prosody, especially in the field of intonational phonology, implicitly assumes that prominence is primarily a function of f_0 (see Terken and Hermes (2000) and Beckman (1986) for reviews).

Prominence of a syllable is sometimes explicitly equated with special f_0 motions in its vicinity. For instance, Ladd (1996) states:

> A pitch accent may be defined as a local feature of a pitch contour – usually, but not invariably a *pitch change,* and often involving a local minimum or maximum – which signals that the syllable with which it is associated is *prominent* in the utterance. … If a word is prominent in a sentence, this prominence is realized as a pitch accent on the 'stressed' syllable of the word.

Similarly equating pitch motions with prominence, Welby (2003) writes: "The two versions [of an utterance] differ in that (1) has a pitch accent, a prominence-lending pitch movement…". A standard textbook by Roca and Johnson (1999, p. 390) claims that f_0 patterns can be used to prove the reality of abstract lexical stress: they state that one can test syllables for stress by looking at f_0 in their vicinity. Another textbook, Clark and Yallop (1995, p. 349), gives a less extreme view but still espouses the primary importance of f_0 when discussing the acoustic implementation of lexical stress: "Our perception is in fact likely to be more responsive to the pitch pattern than other factors." Similar views were put forth by Bolinger (1958); 't Hart *et al.* (1990) and others. Since the assumption that pitch implies prominence underlies much work, it needs to be thoroughly tested.

To do this, we studied seven dialects of British English. We looked for patterns in f_0 and other acoustic properties that could separate prominent from non-prominent syllables.

II. DATA AND METHODS

II.A. Overview

This experiment is conducted on a large corpus of natural speech. Listeners judge the prominence of syllables, and the speech is analyzed to find the acoustic basis of their judgements.

We measure a selection of acoustic properties in a window that centers on a syllable. Listeners mark the syllables as either prominent or not. Five time-series are then computed from the speech signal: measures of loudness, aperiodicity, spectral slope, f_0, and a running measure of duration. These measures are transformed into coefficients for Legendre polynomials and fed into a classifier that is trained to reproduce the human prominent/non-prominent decision. Finally, the classifier performance is measured on a test set, and the result reveals how consistently the speakers used each of the measured properties to mark prominent syllables.

The first step in the analysis is the extraction of prominence marks (§II.C) from a labeled corpus (§II.B). Second, the five time series ("acoustic measures") are computed from the speech; details are in §II.D. Third, each property is normalized (§II.E), then, fourth, the data are represented as a best-fit sum of Legendre polynomials (§II.G.1). The coefficients of the polynomials that result from the fit are a compact representation of the shape of the time-series in the window. (Some of these coefficients are easy to interpret: the first coefficient is the average over the window; the second coefficient captures the overall rate of change.)

These coefficients form a feature vector, which is the input for the fifth stage of the analysis. The feature vector specifies a point in a space; hence the Legendre polynomial analysis maps an acoustic time-series into a single point into a (e.g.) 6-dimensional feature space. Each point in that space (each syllable) is labeled as prominent or non-prominent by a human. Fifth, we build a classifier (§II.H) on those vectors to reproduce the human prominence marks as well as possible. We use a Quadratic Discriminant Forest classifier, which should be reasonably efficient for our features, which are roughly multivariate Gaussian and have no obvious complex structure.

We chose this classifier partially because it is a variant of a quadratic classifier, and can capture classes that are linguistically interesting. For instance, if f_0 indicated prominence by being either high or low at the syllable center (and non-prominent by being intermediate), we could capture that behavior. Likewise, if f_0 indicated prominence by slopes or extra variance, a quadratic classifier could capture such classes.

Since each class is defined by a full covariance matrix among all the orthogonal polynomial (OP) coefficients, it can represent complex patterns of low and high pitch combined with large and small standard deviations. Specifically, using this design of classifier will let us test models of prominence where $f_0(t)$ on a syllable is measured relative to any linear combination of the surrounding f_0 measurements. This includes many plausible normalizations of $f_0(t)$ relative to preceding and/or following syllables, such as a consistent declination slope. We put quantitative limits on the classifier performance in §III.D.

Sixth, after the classifiers are built and tested, we compare the error rates for classifiers based on different acoustic measurements (§III) to deduce how much information is carried by each acoustic property.

II.B. Corpus

We use the IViE (**I**ntonational **V**ariation **i**n **E**nglish) corpus (Grabe *et al.* (2001)), which is freely available on the Web at `http://www.phon.ox.ac.uk/ivyweb` . The IViE corpus contains equivalent sets of recordings from seven British English urban dialects: Belfast, Bradford (speakers of Punjabi heritage), Cambridge, Dublin, Leeds, London (speakers of Jamaican heritage), and Newcastle. Speech of six of the twelve speakers per dialect have been intonationally labeled. The speakers were students in secondary schools, with a mean age of 16 years ($\sigma < 1$ year). We use data from three styles of speech: the "sentences", "read story" and "retold story" sections of IViE. (Abbreviated below as "sentences", "read" and "retold".)

In "sentences", speakers read lists of sentences like "We were in yellow." or "May I lean on the railings?" The "read" section involved reading a version of the Cinderella story, containing narration and dialog. In the "retold" section, the subjects re-told the story in their own words, from memory. The IViE corpus has about 240 minutes of annotated data, which includes about 7200 intonational phrases and 14400 accents. For this analysis, we use all the annotated IViE single-speaker data.

II.C. Prominence marks

The IViE corpus contains files marking prominent syllables. We adopted these as the primary data source. In IViE, all accented syllables are prominent and *vice versa*. The marks were made by two phoneticians (one of whom, EG, is an author), who are experienced in the analysis of English intonation. The phoneticians were native speakers of Dutch and German who acquired RP English before adolescence. They consulted with a third phonetician who was a native speaker of British English. Accented syllables were marked according to the British tradition defined by O'Connor and Arnold (1973) and Cruttenden (1997), using the prosodic prominence hierarchy of Beckman and Edwards (1994). During labeling, the speech was heard and the speech waveform and f_0 trace were displayed on a screen.

Non-prominent syllables were not marked in IViE, but word boundaries were. Using the boundaries, one can deduce the locations of most non-prominent syllables and automatically mark them. We built a dictionary containing the number of syllables in a typical conversational version of each word or word fragment. An analysis program then scanned through the labeled part of the corpus. As each word was encountered, the program placed the correct number of syllable marks, evenly spaced throughout the word. Any syllables that IViE shows as prominent then replaced the nearest automatically generated mark. The remaining non-prominent syllables are needed as a comparison to the prominent syllables because the classifiers are trained and tested on their ability to separate two classes.

The primary set of speech data includes 2173 prominence marks out of an estimated 5962 syllables in the “sentence” style; 1919/5134 in the “read” style; and 805/2341 in the “retold” style. Most are on syllables that have primary or secondary lexical stress. In the “read” style in the primary set, the Belfast and Cambridge dialects had considerably more data labeled than other dialects: 34 and 32 audio files, respectively, *vs.* a total of 18 for the other five dialects. Otherwise, the data was almost evenly balanced between dialects.

Two other sets of prominence marks were produced independently, to ensure that the primary data source reflected widely perceived properties of the language, rather than something specific to the primary labelers. These two secondary sets were smaller, but (unlike the primary data set) they also contained marks for the centers of non-prominent syllables. Data files were chosen randomly from “read” data obtained in Cambridge, Belfast, and Newcastle, from audio files that had transcriptions. The secondary sets were created by two people with significantly different training and dialects from the primary labelers.

In the labeling for the secondary sets, the labelers attempted to mark syllables that perceptually “stand out”, giving minimal attention to meaning or syntax. No attempt was made to discriminate between lexical stress, focus, and other causes of prominence. No attempt was made to decide what type of accents were present or to define intonational phrases. One secondary labeler (GK, an author) is a native speaker of American English (suburban Connecticut), trained as a physicist. The GK set has 454 prominence marks out of 1385 syllables. The other secondary labeler (EL) is a native speaker of Scottish English (Glasgow), trained as a Medieval English dialectologist. The EL set has 775 prominence marks among 2336 syllables.

During the secondary labeling, only the speech waveform and word boundaries were displayed; IViE labels were not displayed; and the primary labelers were not consulted. Marks were placed without regard to a detailed phonetic segmentation; syllables were marked somewhere between the center of the voiced region and the temporal center of the syllable. The secondary labelers had the option of not labeling a word if the number of syllables was unclear or if it was a fragment. Otherwise, they marked each syllable as prominent or non-prominent.

The secondary sets include some data that are not in the primary set: 3/12 audio files in the GK set are not in the primary set, 8/24 in the EL set are not in the primary set, and only two audio files are common between the GK and EL sets. The secondary sets thus bring in new data and are almost independent of the primary set, but they have enough overlap to allow some limited comparison of the consistency of label placement.

Overall, the median spacing between neighboring syllable centers in the secondary sets is 180 ms (which is also the median syllable duration). The median distance between prominence marks is 440 ms in the primary set and about 600 ms in the secondary sets.

II.D. Acoustic measures

We based the paper on five acoustic measures that are plausibly important in describing prosody. All are time-series, and they describe the local acoustic properties. We used approximations to perceptual loudness, and phone duration, a measure of the voicing (aperiodicity), the spectral slope, and the fundamental frequency. In addition to the three classic contenders, we added a spectral slope measure because of the success of Sluijter's spectral slope measurement (Sluijter and van Heuven, 1996). Aperiodicity was added simply as a relatively unexplored candidate: it is sensitive to some prosodic changes (e.g. pressed vs. modal vs. breathy speech) and so might plausibly be correlated with prominence. Additionally, it is sensitive to the relative durations of vowels and consonants in syllables, and therefore might capture some duration changes associated with prominence. Loudness and duration, together, capture at least some of the acoustic features of vowel quality; reduced vowels tend to be quiet and short; more open vowels tend to be louder.

II.D.1. Loudness

The loudness measure is an approximation to steady-state perceptual loudness (Fletcher and Munson, 1933). The analysis implements a slightly modified version of Stevens' Mark VII computation (Stevens, 1971), which is an improved version of the ISO-R532 Method A standard noise measurement. We modified it to use 0.7 octave frequency bins rather than the full- or third-octave bands for which it was originally defined. It operates on the spectral power density derived from an $L = 50$ ms wide, $1 + cos(2\pi(t - t_c)/L)$ window, and supposes that the RMS speech level in an utterance is 68 dB relative to 20 $\mu\mathrm{N/m}^2$ sound pressure.

The IViE recordings were obtained in whatever spaces were available, so background noise is sometimes audible. The noise could affect our analysis because the weight we assign to acoustic measures depends on the loudness (§II.F), and changes in the weight will affect the orthogonal polynomial coefficients (§II.G.1). To minimize this problem, we subtracted an estimate of the background noise from the loudness.

The correction was

$$L^3(t) = \max(0,\ L_r^3(t) - \hat{L}_r^3), \tag{1}$$

where $L_r(t)$ is the raw (Stevens) loudness measure, $L(t)$ is a corrected loudness, excluding the background noise, and $\hat{L_r}$ is an estimate of the background noise loudness. Equation 1 is approximate and assumes that the speech and noise spectrum have the same shape; The ratio of peak speech power to the background noise is typically about 30 dB, however, so the correction only affects the quietest parts of most utterances. $\hat{L_r}$ was conservatively set equal to the fifth percentile of L_r, as all the utterances contained at least 5% silence. In other words, the analysis assumed that the quietest 5% of the data contained no speech and could be used to estimate the background noise level.

II.D.2. Running Duration Measure

The running duration measure, $D(t)$ is a time series whose value at each moment approximately equals the duration of the current phone. It is derived by finding regions with relatively stable acoustic properties and measuring their length. Longer phones, especially sonorants, will tend to have long regions that have nearly constant spectra and will give large values for $D(t)$. Shorter phones will give small values for $D(t)$. Stops are treated as the edge of a sonorant plus a silence, and bursts are effectively treated as separate entities. Short silences have the expected duration, but $D(t)$ is ill-defined for long silences.

To compute the running duration measure do:

- For every 10 ms interval in the utterance, compute a perceptual spectrum, $\psi(t_c, j)$, where t_c is the time of the window center and j is the frequency index, in Bark. The frequency interval from 300 to 5500 Hz is used. The Fourier transform of the signal is taken over an $L = 30$ ms wide, $1 + \cos(2\pi(t - t_c)/L)$ window[1]. Then the power spectrum is normalized by the total power within the window. The spectral power density is then collected into 1 Bark wide bins on 0.5 Bark centers, and a cube-root is taken of the power in each bin. (The summed power across all bins is of the order of unity.)

Then, in a second pass, compute the $D(t)$ at each 10 ms interval as follows:

- Starting at $t = t_c$ with $\eta = 0$, and moving t forward from t_c in 10 ms steps, accumulate $\eta = \eta + \sum_j (\psi(t, j) - \psi(t_c, j))^2$. This is a measure of how much the spectrum has changed over the interval between t_c and t.
- In the same sweep, accumulate $\Delta_{\text{fwd}} = \Delta_{\text{fwd}} + e^{-\eta/C}$, with $C = 600$. As long as the accumulated difference is smaller than C, Δ_{fwd} will approximately equal the time difference, $t - t_c$, but when the spectrum changes and η becomes bigger than C, the accumulation will slow down and stop. The final value of Δ_{fwd} will be approximately equal to how far one can go in the forward-time direction before the spectrum changes substantially.
- Do the same in the reverse direction, to compute Δ_{rev}.
- The $D(t_c)$ is then (10 ms)$\cdot(\Delta_{\text{rev}} + \Delta_{\text{fwd}} - 1)$, where the final "-1" corrects for double counting of the sample at t_c.

Figure 1 shows a section of acoustic data and the resulting time series of $D(t)$ for a phrase "...go to the ball...", along with the input waveform. The values of $D(t)$ near each sonorant center approximately match the phone duration.

II.D.3. Aperiodicity

The aperiodicity measure, $A(t)$, ranges from 0 to the vicinity of 1. It assigns zero to regions of locally perfect periodicity, and numbers near one where the waveform of the signal cannot be predicted. (For stationary signals, the maximum value is unity, but amplitude changes, especially on 20 ms or shorter time scales, will locally change the maximum.) It is related to Boersma's Harmonics-to-Noise ratio (Boersma, 1993) (HNR) and can be approximated by $A(t) \approx (1 + 10^{\text{HNR}/10})^{-1/2}$. $A(t)$ can also be considered a measure of voicing, as voiced speech is often nearly periodic and unvoiced speech is typically aperiodic.

To compute $A(t)$, the audio signal first had low-frequency noise and DC offsets removed with a 50 Hz fourth-order time-symmetric Butterworth high-pass filter, and then was passed through a 500 Hz single-pole high-pass filter for pre-emphasis. The aperiodicity measure was derived by taking a section of the filtered signal defined by a Gaussian window with a 20 ms standard deviation and comparing it to other sections

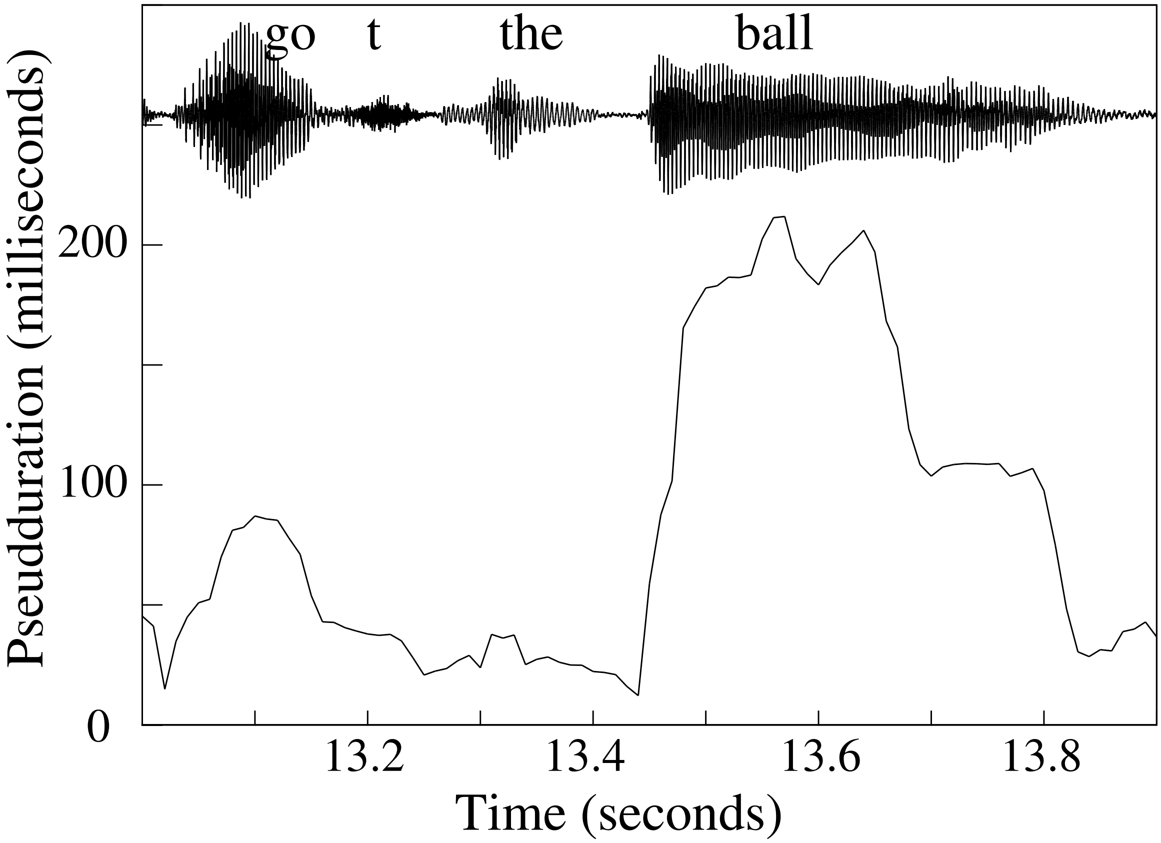

FIG. 1. Running duration measure, $D(t)$ (below, smooth curve) and acoustic waveform (top) for "…go to the ball…" The sharp downwards step in $D(t)$ near 13.65 s corresponds to the transition between the vowel and liquid in "ball"; the two adjacent sounds have different durations.

shifted by 2 to 20 milliseconds. If the acoustic signal were exactly periodic with f_0 between 50 and 500 Hz, then one of the shifted windows would exactly match the starting window, and the difference would be zero. The value of $A(t)$ is proportional to the minimum RMS mismatch between the windows.

To compute the aperiodicity measure:

- For each possible shift, δ, between 2 and 20 milliseconds, compute $p_\delta(t) = (\tilde{s}(t + \delta/2) - \tilde{s}(t - \delta/2))^2$, where $\tilde{s}(t)$ is the filtered acoustic waveform at time t.
- Compute $P(t) = \tilde{s}^2(t)$
- Convolve $p_\delta(t)$ and $P(t)$ with 20 ms standard deviation Gaussians to yield $\bar{p}_\delta(t)$ and $\bar{P}(t)$, respectively.
- Compute $\hat{p}(t) = \min_\delta\{p_\delta(t)\}$, i.e. find the minimum error at each time, minimizing over all the shifts, δ.
- The aperiodicity measure is then $A(t) = \hat{p}^{1/2}(t)/(2\bar{P}(t))^{1/2}$.

Figure 2 shows a small section of acoustic data and the resulting time series of $A(t)$ near the end of the word “railings”, along with the input waveform and enlarged sections of the pre-processed (high-pass filtered) waveform.

II.D.4. Spectral Slope

The spectral slope estimator is intended to approximate the average slope of the power spectrum near the glottis, i.e. , to be relatively insensitive to the formant structure of the speech. It takes a local spectrum of the speech waveform, computed in a 30 ms window, and collects the power in 1 Bark bins (the bins are overlapping, on 0.5 Bark centers). Next, a cube-root is taken to yield an approximation to the perceptual response in each frequency band. Finally, the spectral slope estimate, $S(t)$ is the slope of the best-fit to the Bark-binned spectrum between 500 Hz and 3000 Hz.

Related measures are described in Heldner (2001); Sluijter and van Heuven (1996) and references therein. Our measure is not identical to prior measures; but should have a substantial correlation with them. We chose it because it could be computed easily and reliably on a large corpus, in a strictly automated manner.

II.D.5. Fundamental Frequency

We compute an estimate of the fundamental frequency, $f_0(t)$, with the *get_f0* program from the ESPS package (Entropic Corp.). The program also produced a

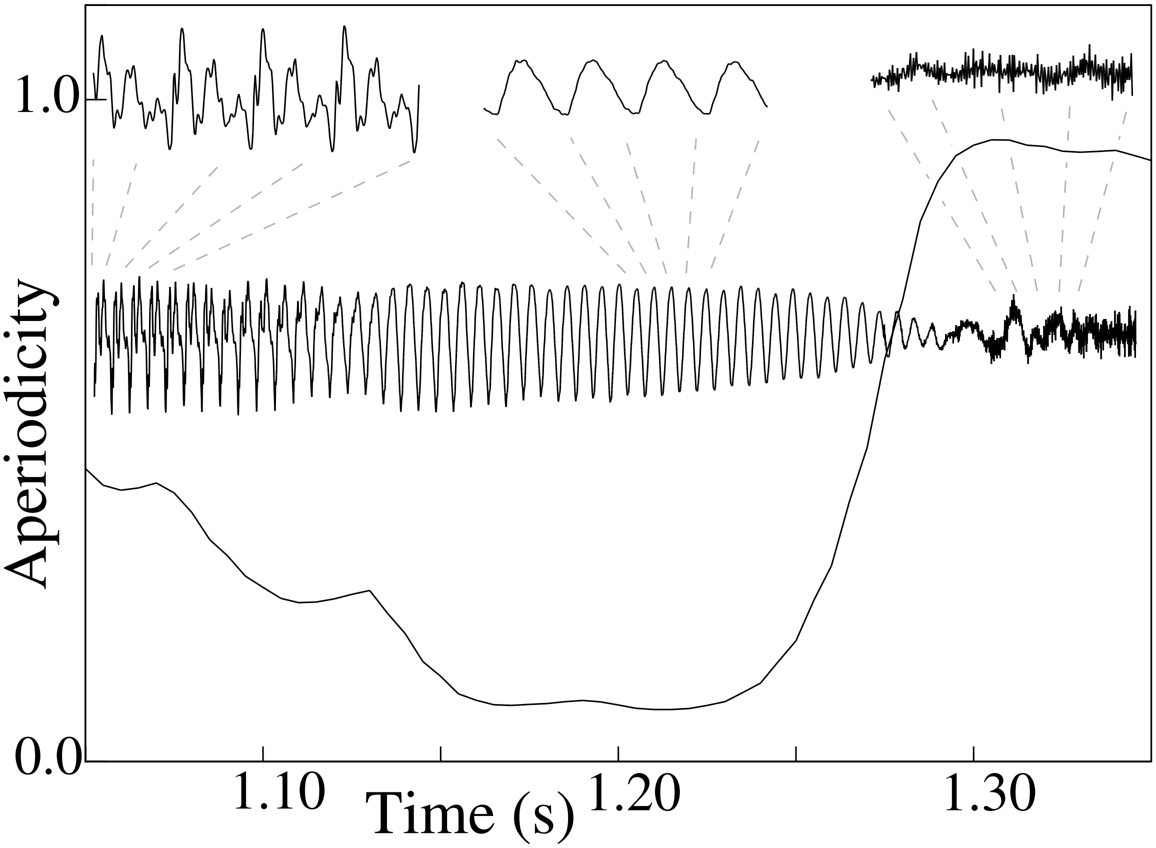

FIG. 2. Aperiodicity measure (below, smooth curve) and acoustic waveform (middle) for the end of "railings." Enlarged sections of the high-pass filtered waveform are shown above. The data show a vowel, nasal, and the final unvoiced fricative. Modest changes from period to period can be seen in the leftmost section of the waveform, leading to an intermediate value of $A(t)$; the middle section is more periodic, so $A(t)$ is close to zero; and the fricative produces a large value of $A(t)$.

voicing estimate, $V(t)$, which was zero or one at each 10 ms interval. Before further analysis, the f_0 tracks were inspected for gross errors. An automated procedure that (a) searched for substantial jumps, and (b) looked for f_0 values close to the subject's minimum and maximum f_0 was used to identify likely problem areas. A roughly equal number of problems were identified during manual inspections driven by various checks not directly associated with f_0. About half of the utterances were manually inspected, and we checked f_0 on every utterance that we inspected. Finally, another set of utterances was inspected because the mean-squared error of the Fourier fit was unusually large.

Once an utterance was identified as having possible problems with its pitch tracking, a labeler inspected each area and marked a change to $f_0(t)$ or $V(t)$ if *get_f0* results did not match the perceived sound. The labeler had the option of shifting f_0 up or down by a factor of 1.5, 2, or 3, and/or marking a region as irregularly voiced or unvoiced. In all, 498 regions in 254 utterances were marked, of which 75 regions included upwards octave shifts of f_0, while 15 were f_0 shifts by other factors. The remaining majority were either marked as irregular phonation or no phonation. The median length of the marked regions is 56 ms.

II.E. Normalization

To compute the orthogonal polynomial coefficients, we take data from a window of width w, centered on the relevant syllable. We then normalized the time axis so that the data ranged from -1 to 1, in preparation for fitting OPs to the data. This converted the t-axis to an x-axis via $x = 2(t - t_c)/w$, where t_c is the time of the center of the syllable.

Additionally, we normalized each acoustic property relative to a weighted average of the corresponding speaker's data of that property over the corpus. For $f_0(t)$, we divided by the 10%-trimmed weighted average[2] of $f_0(t)$. For $A(t)$, we divided by the 35%-trimmed weighted average[3]. For $S(t)$, we subtracted the 10%-trimmed weighted average. Finally, because the microphone placement was not controlled in the recordings in the IViE corpus, we normalized $L(t)$ locally, so motions of the speaker would not have much effect on the normalized amplitude. We normalized $D(t)$ and $L(t)$ by dividing by the 5%-trimmed weighted average over the window. This local normalization reduces the sensitivity of the analysis to changes in the speaking rate or microphone position between one utterance and another.

II.F. Weighting the data

Not every part of the acoustic measures are equally valuable. For instance, f_0 information is meaningless in unvoiced regions, as is S, D, and A. It is necessary, then, to give a weight to each point in the data when we later compute the orthogonal polynomial fits in Equation 4. The weight function is written $W_\alpha(t)$, where α indicates one or another of the acoustic measures.

The detailed form of the weight functions are somewhat arbitrary, but we made plausible choices, then tested that they are close to optimal (see Appendix A). All the weight functions are computed from the acoustic measures before normalization.

The weights are different for each acoustic measure, but they share some common features. Specifically, using weights that increase with loudness will emphasize regions that may be more perceptually important. Under real-world conditions, speech more than 15 dB below the peaks is often buried in ambient noise, and thus has less importance[4].

For f_0, in addition to perceptual importance, we were motivated by considerations of the accuracy and reliability of the pitch tracker. We took $W_{f_0}(t) = L^2(t) \cdot \max(1 - A^2(t), 0)^2 \cdot V(t) \cdot I^2(t)$, where $V(t)$ is the voicing estimate from §II.D.5. The component $I(t)$ is a semi-automatic indicator of irregular voicing. It is a product of factors:

- A factor that de-weights the edges of a voiced region to reduce the impact of segmental effects. It is unity everywhere except in the first and last 10 ms of each voiced region where it is 0.5.

- A factor that de-emphasizes unstable f_0 readings: $(1 + (\delta/10\text{ Hz})^2)^{-1}$, where δ is the pitch change over the 10 ms interval between samples.

- A factor that is 1, except 0.5 in regions hand-marked as irregularly voiced (see §II.D.5).

This weight function forces the orthogonal polynomial fit to $f_0(t)$ to be most precise in loud regions that are periodic, such as syllable centers.

For the spectral slope, we suppressed the unvoiced regions to avoid the large jumps in $S(t)$ that occur across voiced-unvoiced transitions. Thus, we used $W_S(t) = L^2(t) \cdot V(t)$. For aperiodicity and the running duration measure we used $W_A(t) = L^2(t)$ and $W_D(t) = L^2(t)$. Finally, for $L(t)$, we used a uniform weight: $W_L(t) = 1$. The net result of our weighting choices is to focus on the peak of the syllable, paying less attention to the margins, especially consonant clusters.

Weighting the data with a power of the loudness gives us some sensitivity to the relative timing of f_0 excursions with respect to syllable centers. For instance, f_0 peaks that appear earlier than syllable centers will have the largest weight applied to their falling edge. The resulting OP coefficients will be biassed towards those those of a falling accent. A delayed f_0 peak will have more weight placed on its rising edge and will push the coefficients towards those of a rising accent.

II.G. Orthogonal polynomials

We use orthogonal polynomials because the intentionally controlled aspects of intonation are, by and large, smooth and continuous. This is especially true for $f_0(t)$ (Kochanski *et al.* (2003, section 1.2), Kochanski and Shih (2000)), because f_0 is controlled by muscle tensions that are smooth functions of time. We chose Legendre polynomials (Hochstrasser, 1972) which have the property of orthogonality:

$$\sum_x P_i(x) \cdot P_j(x) \cdot \omega(x) \cdot dx = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Here, $P_i(x)$ is the i^{th} Legendre polynomial, $\omega(x)$ is the weight function that specifies the family of orthogonal polynomials ($\omega(x) = 1$ for Legendre polynomials). The sum is computed on the 10 ms grid where the acoustic measures are computed. Note that $\omega(x)$ and $W(x)$ aren't the same: $\omega(x)$ is a global property of the entire analysis; $W(x)$ is the weight function used to fit the sum of polynomials to a particular utterance.

This orthogonal polynomial analysis is similar to a Fourier transform in that the low-ranking polynomials pick out slowly-varying properties and the higher-ranking polynomials pick out successively more rapidly varying properties. The n^{th} Legendre polynomial has $(n-1)/2$ peaks and the same number of troughs, if we count a high (low) point at an edge of the utterance as half a peak (trough).

II.G.1. Deriving coefficients – fitting the acoustic data

One can derive coefficients that represent an acoustic measurement by fitting it with the sum of Legendre polynomials,

$$y(x; c) = \sum_x c_i \cdot P_i(x), \tag{3}$$

using a regularized, weighted linear regression. In Equation 3, c_i are the coefficients that multiply each Legendre polynomial, and $y(x; c)$ is a model for the data. The model (y) is $x-$ (e.g. time) dependent, and also depends on the coefficients, c. To compute the coefficients that best represent some data $\alpha(x)$, we minimize

$$\mathbb{E}_\alpha = \sum_x W(x) \cdot (y(x; c) - \alpha(x))^2 + \gamma \cdot c_i^2. \tag{4}$$

The first term is the normal sum-squared-error term; the second term is a regularization term. In Equation 4, $\alpha(x)$ stands for each of the five acoustic time series, and γ is the strength of the regularization. The regularization causes $c_i \to 0$ when $\gamma \to \infty$, and is equivalent to assuming a Gaussian prior probability distribution with a width proportional to γ^{-1} in a maximum a posteriori probability (MAP) estimator. Descriptions of the method can be found in Press *et al.* (1992, pp. 808–813) and Gelman *et al.* (1995).

We use linear regularization because some of the syllables have $W(x) \approx 0$ over 50% or more of the window; an example might be a syllable with a long fricative when one is fitting $f_0(t)$. In such a case, Equation 4 becomes nearly degenerate when $\gamma = 0$ and yields large, cancelling values of the coefficients c_i. The resulting c_i are far outside the distribution obtained for most syllables and degrade the classifier performance by violating its assumption of Gaussian classes.

Regularization can limit these spurious values of c_i. We chose $\gamma = 10^{-4} \sum_x W(x)$, which has the effect of reducing most c_i by only about 1%, but yields fairly good behavior for the hard cases that have large regions in which $W(x) \approx 0$.

By experimentation, we found that good fits to the time-series of acoustic data can be obtained by using $1 + w/2\tau_\alpha$ orthogonal polynomials, where w is the length of the analysis window and $\tau_L = 60$ ms, $\tau_D = 70$ ms, $\tau_{f_0} = 90$ ms, $\tau_A = 80$ ms, and $\tau_S = 90$ ms. We make τ_L small because the loudness contours have sharp features which require a higher density of orthogonal polynomials in order to get a good fit; τ_{f_0} is adequate to represent the relatively slow f_0 variations. Others are in between.

The fits are generally quite accurate. The weighted RMS error between the normalized time series and the fit is 0.008 (about 1 Hz) for f_0, 0.14 (i.e. 15%) for loudness, 0.09 (i.e. about 8 ms) for $D(t)$, 0.13 (i.e. about 13% of the median) for aperiodicity, and 0.003 (i.e. less than a 1 dB shift in the spectral power density at 3000 Hz relative to the power at 500 Hz). This is probably good enough to be indistinguishable by human perception, so we presumably capture most of the relevant information. In Appendix B shows that our results are relatively insensitive to the values of τ_α or (equivalently) to the accuracy of the fit.

The weighted orthogonal polynomial fit to $f_0(t)$ is not strongly affected by small changes in which regions are voiced. Indeed, if $f_0(t)$ were fit exactly, de-voicing small regions would have no effect at all. Much of the $f_0(t)$ time series is indeed smooth and well-fitted by the polynomials, so changes to voicing are primarily captured by $L(t)$ and $A(t)$.

II.G.2. Transforming coefficients to make them more Gaussian

Next, we transform the coefficients to remove any obvious nonlinear correlations. We saw that for f_0 and especially loudness, the scatter plot of c_0 vs. c_1 was crescent-shaped. However, it could be made much closer to a Gaussian by the following adjustments: $c_0 \Leftarrow c_0 - \kappa c_1^2$. Since the histograms for c_2 and other coefficients also had visible curvature, all the coefficients except c_1 were adjusted via

$$c_i \Leftarrow c_i - \kappa_{\alpha,i} c_1^2. \tag{5}$$

A linear least squares procedure was used to determine $\kappa_{\alpha,i}$, from the union of the prominent and non-prominent data. For each coefficient (except c_1) and for each acoustic measure, α, the scatter-plot of $c_i\ vs\dot{}\ c_1$ was fitted to $\hat{c_i} = \eta_{\alpha,i} + \nu_{\alpha,i}c_1 + \kappa_{\alpha,i}c_1^2$, and c_i was then corrected via Equation 5. We did not need to consider η further, as it is picked up by μ in the classifier, and ν becomes part of the classifier's covariance matrix. The transformed c_i is the feature vector that will be used by the classifier.

II.H. Classifier

We developed a Bayesian Quadratic Forest Classifier, inspired by the forest approach of Ho (1998). The classifier is a straightforward application of Bayes' Theorem. To build the classifier, assume that there are M classes, each defined by a multivariate Gaussian probability distribution

$$P(\vec{z}\,|\text{class } i) = P(\vec{z}\,|\mu_i, H_i) = (2\pi)^{-N/2} \cdot \det(H_i) \cdot \exp(-(\vec{z} - \vec{\mu_i})^T \cdot H_i \cdot (\vec{z} - \vec{\mu_i})) \quad (6)$$

on the input coordinates, $\vec{z}$, where N is the dimension of $\vec{z}$, $\vec{\mu_i}$ is a vector that defines the center of the i^{th} class, H_i is the inverse of the i^{th} class's covariance matrix, and $\det(H_i)$ is its determinant. There are then M hypotheses: the input coordinates belong to one or another of the M classes ($M = 2$ here, i.e. prominent or non-prominent).

One can then use Bayes' Theorem to compute $P(\text{class } i|\vec{z})$ from the set of $P(\vec{z}\,|\text{class } i)$ and the relative frequency with which one observes the various classes. The classifier output is then the class that has the largest probability, given $\vec{z}$. If the classes are observed equally often, this boils down to picking the class with the largest $P(\vec{z}\,|\text{class } i)$. The classifier is defined by a choice of M triplets of (μ_i, H_i, ϕ_i), where ϕ_i is the prior probability of observing of each class. The algorithm operates like a linear discriminant analysis in that it chooses ϕ_i, μ_i and H_i so as to maximize the product, of the probabilities that the feature vectors are classified correctly.

Figure 3 shows a sample set of feature vectors that are sent to the classifier. The figure shows loudness data for Cambridge (all styles). Only two of the eight components of the feature vector are shown, so the separation in this two-dimensional projection is not as good as is possible in the full eight-dimensional space. The dashed line is an approximate class boundary derived from the machine-classifications of the data[5].

One limitation of a standard discriminant classifier is that when the number of feature vectors becomes small, the border between classes becomes poorly-defined; many algorithms fail entirely when the number of training points is smaller than the number of parameters necessary to define the classifier.

To avoid this, we computed an ensemble of good estimates by way of a Markov Chain Monte-Carlo process, rather than limiting ourselves to a single "best-estimate" of the classifier parameters. The Markov Chain Monte-Carlo process generates samples of μ, H, and ϕ from the distribution $P(\mu_i, H_i, \phi_i|\vec{z})$. This distribution is sharp as long as the number of feature vectors is much larger than the number of parameters

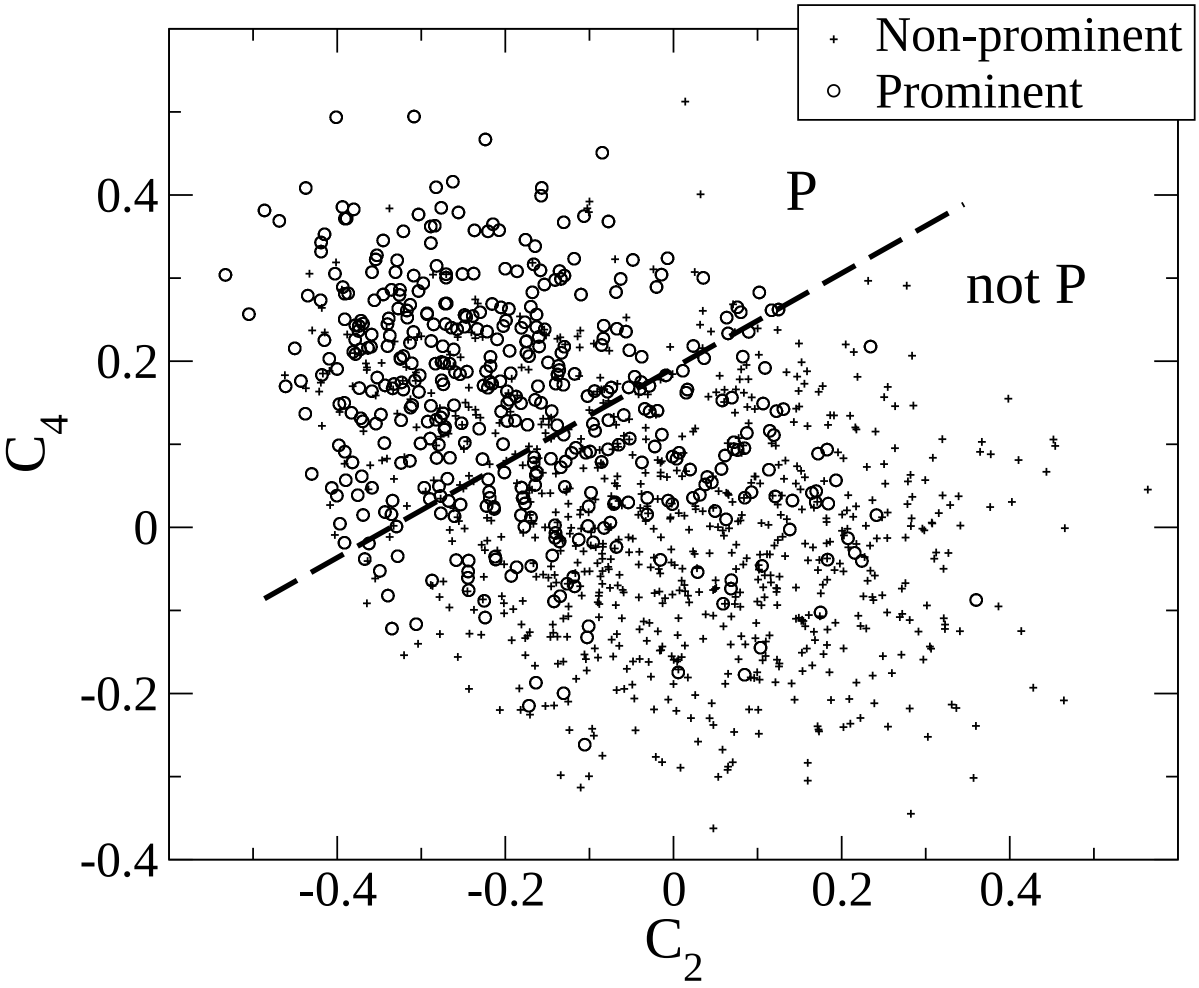

FIG. 3. Scatter plot showing two of the eight components of the feature vector for a loudness classifier for Cambridge data. Each point corresponds to a syllable. Prominent syllables are marked with circles, non-prominent by "+". The dashed line is an approximation to the classifier boundary, derived from the machine classifications of the syllables.

that defines the classifier (which is $(M-1)\cdot(N+N\cdot(N+1)/2+1)$). For small numbers of feature vectors, however, the probability distribution of the covariance matrix will become broad and heavy-tailed, and the prior distribution of $P(\mu_i, H_i, \phi_i)$ becomes important. We chose a prior that is constant, independent of μ, H, and ϕ.

In practice, we found it useful to select the best few from among the classifiers generated by the Markov-Chain Monte-Carlo algorithm. This makes the algorithm far less sensitive to the termination conditions of the Monte-Carlo process and also makes the definition of classification probabilities more comparable with those reported for other classifiers. For this study, we test $Q = 10N$ (about 50) candidate-classifiers against the training set. Of the Q candidates, we keep the best $Q^{1/2}$, where "best" is defined by the fraction of the training set that is correctly classified.

We split the data, randomly assigning 75% to the training set and 25% to the test set. We built classifiers for 8 different splits into training and test set, to allow us to estimate errors for the classification accuracy. Consequently, there were 8 selected ensembles, $80N$ candidate classifiers, and a total of $8 \cdot (10N)^{1/2} \approx 55$ selected classifiers. This approach is a variant on a cross-validation procedure (Webb, 1999, p. 323).

Given this selected ensemble of classifiers, we computed the overall classification accuracy on a test set, averaging the accuracy across the selected classifiers. The accuracy we report is the averaged percent of correct classification, or 100% minus the sum of false-negative and false-positive errors.

One advantage of this Monte-Carlo procedure is that it correctly reproduces the longer tails of Student's t-statistic in the one-dimensional case, whereas any quadratic classifier with a single best value of μ and H cannot.

III. RESULTS AND DISCUSSION

We express results in terms of the classifier effectiveness,

$$K = \frac{F[\text{correct}] - P[\text{chance}]}{1 - P[\text{chance}]}, \tag{7}$$

where $F[\text{correct}]$ is the fraction of the test set that is correctly classified, and $P[\text{chance}]$ is the accuracy of the classification in the absence of acoustic information. Consequently, $K = 0$ implies the acoustic data was useless and yielded chance performance, while $K = 1$ implies perfect classification.

$P[\text{chance}]$ is determined by randomly shuffling the labels to break any association with the acoustic parameters (Table I). We classified such de-correlated data for five different window widths between 446 ms and 458 ms, with the average w chosen to match the $w = 452$ ms that the bulk of the paper discusses (§III.A and thereafter). $P[\text{chance}]$ depends on the acoustic measure, ranging from 59.0% for the loudness classifier to 61.4% for the spectral slope classifier. The differences are significant

Loudness	$D(t)$	f_0	Irregularity	Spectral Slope
59.1%	59.7%	61.1%	60.0%	61.4%

TABLE I. P[chance] for classifiers based on the different acoustic features. These are the probabilities of correctly classifying the acoustic data, after shuffling so that there is no correlation between prominent/non-prominent labels and acoustic properties.

($F(4, 524) = 7.1$; $P < 0.01$), but not large. Presumably they are due to differences in the shape of the distributions. The performance on de-correlated data is slightly worse than the theoretical limit of 63.6% derived by predicting all syllables to be non-prominent.

Unless noted, all results will be for classifiers that are trained for a particular dialect and style of speech (e.g. "read passages in Leeds"). This corresponds to communication within a dialect group. When we present a single value for K, it will refer to the average of all classifiers over the entire corpus. We chose this approach of building many dialect-specific classifiers because of the strong dialect-to-dialect variation that was seen in the IViE corpus in f_0 contours Grabe *et al.* (to appear); Fletcher *et al.* (2004). and well-known cross-dialect differences in the question-statement distinction Cruttenden (1997).

Figure 4 shows the performance as a function of window size for classifiers built from each of the five acoustic measures. The plotted performance is the average over 21 dialect/style combinations. Each classifier separates prominent from non-prominent syllables based on acoustic time series in a window centered on the syllable.

Three important results appear in Figure 4. First, the classifiers based on loudness consistently outperform other classifiers by a substantial margin: they are about 50% better than classifiers based on running duration and more than twice as good as classifiers based on f_0 for most window sizes.

Second, the absolute performance of the f_0 classifiers is unimpressive. With a 452 ms window, the average f_0 classifier predicts only 66.3% of the syllable prominences correctly, which is little better than the 61.1% that can be achieved without the data (P[chance]). We found this surprising, as the prominence marks in the primary data set were made by labelers who expected that pitch motions often induced prominence. Further, they worked under labeling rules that encouraged the association of prominences with pitch events. This result contradicts the widespread view that a set of commonly employed f_0 patterns underlie the perception of prominence or accent.

The classifier can separate a wide variety of f_0 patterns. It can separate prominent from non-prominent patterns if they consistently differ over any 100 ms region within the analysis window, either in f_0 or slope of f_0. It can also discriminate if there are large differences in variance between prominent and non-prominent syllables.

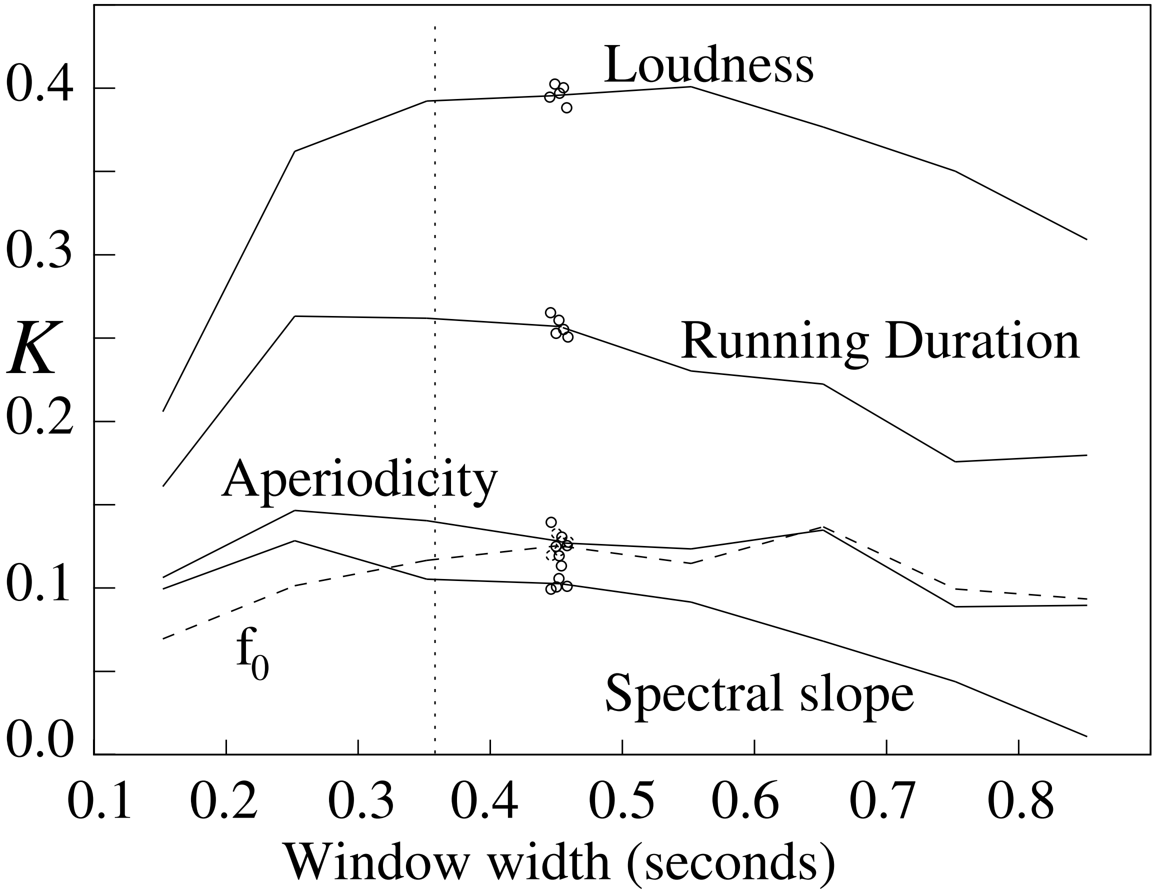

FIG. 4. Classifier performance *vs.* the size of the analysis window, w. Each curve shows performance of classifiers based on a different acoustic feature (f_0 is shown dashed to separate it from its neighbors). The vertical axis is the K-value, which shows how well each classifier performs relative to chance (shown as zero) and exact duplication of the human labels (shown as one). Plotted K-values are averages over seven dialects and three styles of speech. The vertical dotted line marks where the window includes neighboring syllable centers. The small clusters of points near $w = 0.45$ s show the reproducibility of the classifiers, derived from five classifier runs with slightly different window sizes.

Additionally, because we have enough feature vectors to compute the full covariance matrix of all the orthogonal polynomial coefficients for each class, the classifier can also separate classes based on a combination of f_0, its slope and variance. These capabilities are sufficient to yield good classification of syllables based on loudness or duration; the poor results for f_0 then suggest that f_0 simply is not strongly correlated with prominence. Quantitative examples are in section §III.D.

Figure 5 supports this observation that f_0 is not usefully correlated with prominence. It shows that histograms of f_0 at the center of prominent and non-prominent syllables overlap strongly. (Values of f_0 are computed at the window center from the orthogonal polynomial fits to the entire window; this provides interpolation into unvoiced regions.) While the mean f_0 for the entire set of prominent syllables is significantly (in the statistical sense) larger than for non-prominent syllables, that fact is nearly useless to a listener who is attempting to classify a single syllable as prominent or not. For any given f_0, there are roughly equal numbers of prominent and non-prominent syllables, so no measurement of central f_0 for a single syllable provides much evidence as to whether the syllable is prominent or not.

The third result shown in figure 4 is that the loudness and running duration classifiers improve dramatically as the window encompasses the neighboring syllables. This means that prominence depends not just on the loudness or duration of a syllable, but (as one might expect) on a contrast between a syllable and its neighbors. The decline in classifier performance beyond $w \approx 600$ ms is not understood in detail, but some of the decline is certainly caused by longer windows running off the ends of the utterances. Part of it may also be due to the increasing complexity of the classifiers relative to the constant amount of data.

Given that the loudness classifiers continue improving up to $w \approx 500$ ms, and given that the f_0 classifiers are simpler for the same window size since $\tau_{f_0} > \tau_L$, the f_0 classifiers should make efficient use of the available information up to about 500 ms or beyond. In other words, since the loudness classifiers substantially improved by including neighboring syllables, the classifier complexity is probably not limiting the performance for f_0. Thus, the small change in f_0 classifier performance as the neighboring syllables are included in the analysis window suggests that the f_0 of neighboring syllables carries little information.

In the remainder of the paper, we focus on classifiers with $w = 452$ ms, unless noted. We chose this size because it gives nearly peak performance for each acoustic measure.

III.A. Dependence of *K*-values on acoustic measure, dialect, and style of Speech

Figure 6 shows the classifier K-values separated by acoustic measure and speech style, for classifiers trained on a single style/dialect combination. The results

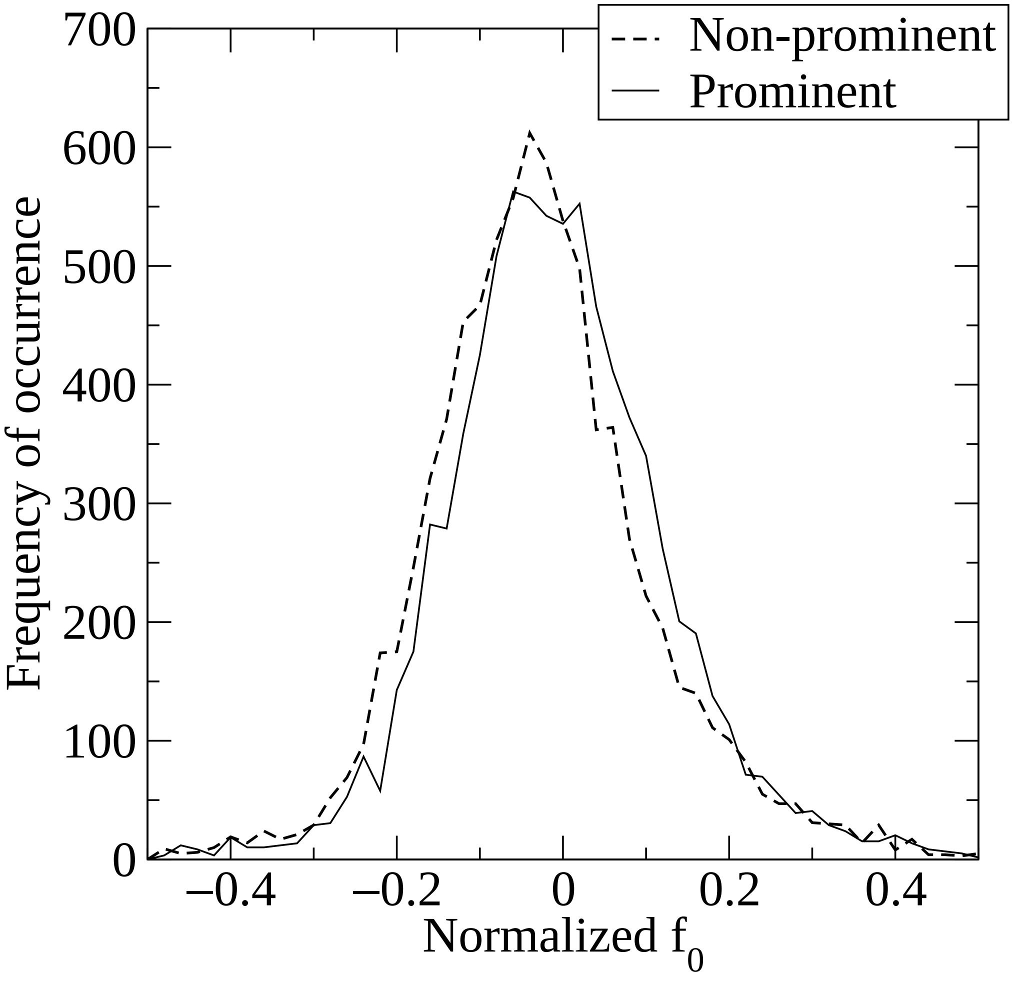

FIG. 5. Histograms of f_0 at the center of a $w = 0.152$ s window for prominent (solid) and non-prominent (dashed) syllables. The distributions are nearly identical, showing that neither the central f_0 nor variance can effectively separate the two classes of syllables.

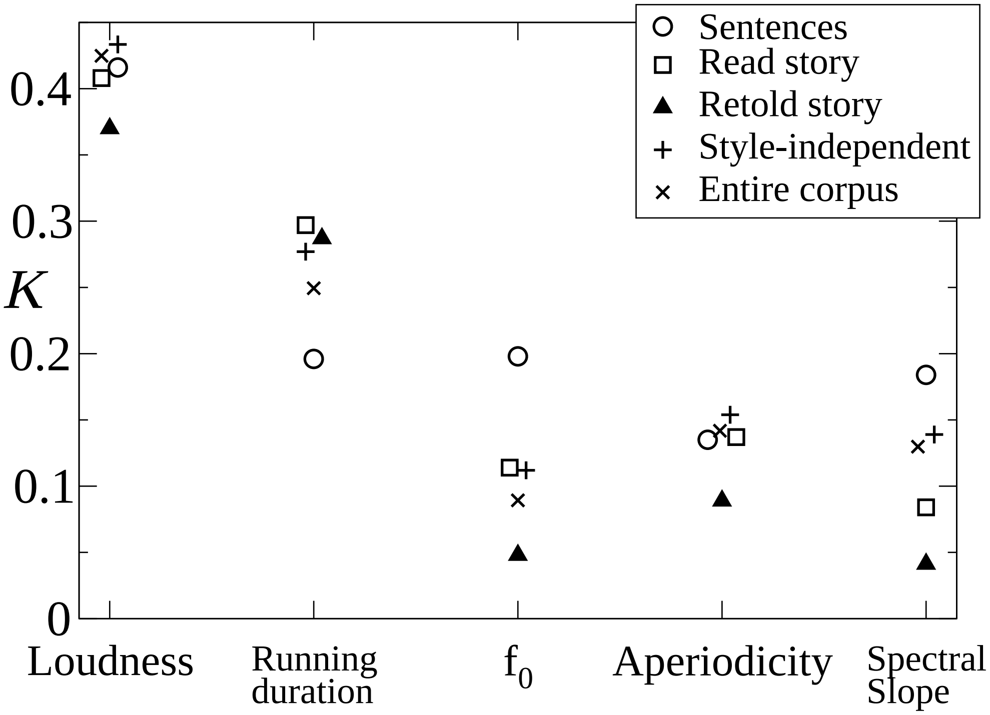

FIG. 6. Classifier performance *vs.* acoustic measure and style of speech. The vertical axis shows performance on a scale where $K = 0$ corresponds to chance and $K = 1$ corresponds to perfect prediction of prominence marks. Classifiers based on loudness perform substantially better than the others for all three styles. The "+" shows style-independent classifiers, and "×" marks classifiers that are both dialect- and style-independent.

Number of dialect/style combinations	Scope of the Classifier	K
1	One style of speech in one dialect.	0.201
3	All styles of speech in one dialect,	0.208
7	One style of speech, covering all dialects.	0.223
21	All styles of speech in all dialects.	0.207

TABLE II. Performance of classifiers trained on increasingly broad portions of the corpus. The top line shows the performance of classifiers trained on a single dialect/style combination; the bottom line is for classifiers of the same complexity, trained on the entire corpus. In the rightmost column, K is averaged over all five acoustic measures.

from classifiers built from the loudness measure are substantially and significantly ($P < 0.001$) better than classifiers based on f_0, spectral slope, or aperiodicity. This conclusion holds true across all styles of speech. However, there are statistically significant differences between different styles of speech (e.g. “retold” *vs.* “sentences”), so one cannot always rank one acoustic measure as better or worse than another. For instance, classifiers built on running duration outperform classifiers built on f_0 for the “read” and “retold” styles, but are effectively equal for the “sentence” style[6]. The statistical errors on these points were derived from the classifier’s cross-validation estimates. They are not uniform, but average to $\sigma_K = 0.02$. Most differences larger than 0.06 are significant at the 0.05 level.

Figure 6 also shows the results of classifiers that are trained on all the styles of speech together (e.g. a classifier is built for all of Belfast speech, rather than just Belfast “read” speech). The average performance of these style-independent classifiers is then plotted as “+”. These classifiers embody the assumption that prominence is marked the same way for all styles of speech. Finally, a style- and dialect- independent classifier (cross“×”, trained on the entire corpus) was built for each acoustic measure. This classifier embodies the assumption that prominence is marked the same way for all dialects and all styles of speech that we studied. These more broadly defined classifiers perform about as well as the average of the style-specific classifiers; this suggests that all of the corpus indeed shares the same definition of prominence.

Table II shows the performance of classifiers that make different assumptions about the breadth of application of the definition of prominence. All the classifiers in the table have the same size feature vector, so they are equally capable of representing the classes. If each dialect had a unique definition of prominence, dialect-independent

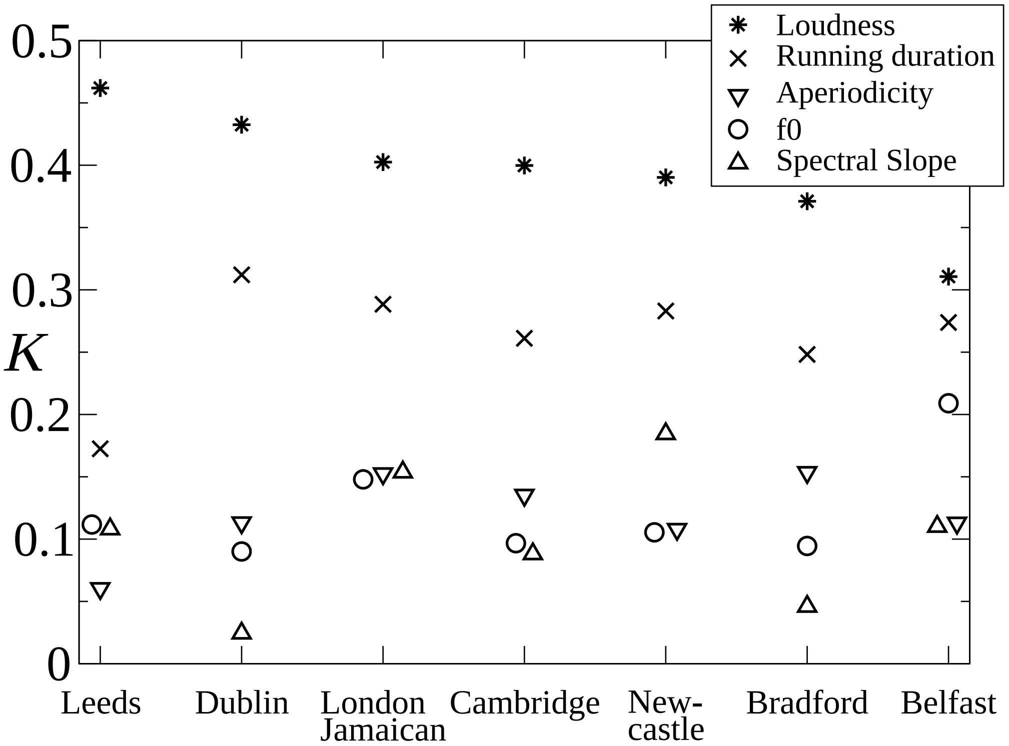

FIG. 7. Classifier performance for the five acoustic measures as a function of dialect. Each classifier is trained on a single dialect/style combination; symbols show the average over the three styles of speech.

classifiers that attempt to represent seven dialects with one set of classes should give poor performance. Likewise, if prominence were encoded differently in the three styles of speech, the style-independent classifiers that use a common definition of prominence for all three styles should give a low K. Instead, different scopes yield nearly the same performance, differing by only 0.03 in K. The near-equality of K-values in Table II implies that there is a useful common definition of prominence across all these dialects and styles of English.

Figure 7 shows the dependence of K-values on acoustic measure and dialect. Again, classifiers based on loudness consistently out-perform all others, with running duration in second place. Some dialect-to-dialect variations exist: most notably, $D(t)$ is relatively unimportant for Leeds, and f_0 is relatively important in Belfast. On average, the classifier's cross-validation error estimates for these points are $\sigma_K = 0.03$, so most differences larger than 0.09 are significant.

III.B. Reconstructing the acoustic properties

The dependence of K_L on window size in Figure 4 implies that prominence depends on a loudness pattern. Reconstructing a loudness profile within the window reveals the details of this pattern. The reconstruction starts with the style- and dialect-independent classifier that represents the entire corpus. We then take all the syllables in a class (e.g. prominent syllables) that are correctly classified. The correctly classified points are represented by OP coefficients, which one can think of as points in a multidimensional space. (Each appears multiple times, once for each classifier in the forest that classified it correctly.) We then compute the centroid of this cloud of points to get the OP coefficients corresponding to a typical, correctly-classified prominent syllable.

Next, these OP coefficients for each class of syllables are converted back into a loudness contour via Equation 3. The resulting curves are averages but are quite representative of individual contours. As we include only contours where the human and machine classifications agree, these resulting contours emphasize the ones where loudness consistently induces a prominence judgement in the listener.

Figure 8 shows loudness reconstructions, as described above. Prominent syllables typically have a loudness peak near the labeled position ($t = 0$), which follows an unusually quiet preceding syllable ($t \approx -180$ ms). The prominent syllable is nearly three times as loud as its predecessor. The following syllable is also quieter than average, but the difference is less dramatic. In contrast, non-prominent syllables typically lie in the midst of a fairly flat loudness profile, with the preceding syllable being slightly louder on average. The secondary data sets are discussed further in §IV.A.

Figure 9 shows a similar reconstruction of $D(t)$. Prominent syllables have a longer region of stable acoustic properties (presumably a longer vowel), following a relatively short preceding syllable. The prominent syllable is nearly three times as long as the preceding syllable and twice as long as the following syllable.

Figure 10 shows the equivalent $f_0(t)$ reconstruction. As expected, prominent syllables typically show a peak in fundamental frequency, but the peak is not large (about 20% in f_0, or about 30 Hz). This plot represents only those utterances which are correctly machine-classified on the basis of f_0. A similar plot based on all utterances would be diluted by the large number of utterances that cannot be correctly classified on the basis of f_0, and would show much less contrast between prominent and non-prominent syllables.

Similar plots would show that prominent syllables have a lower aperiodicity and a more positive spectral slope than their neighbors (i.e. they have more regular voicing

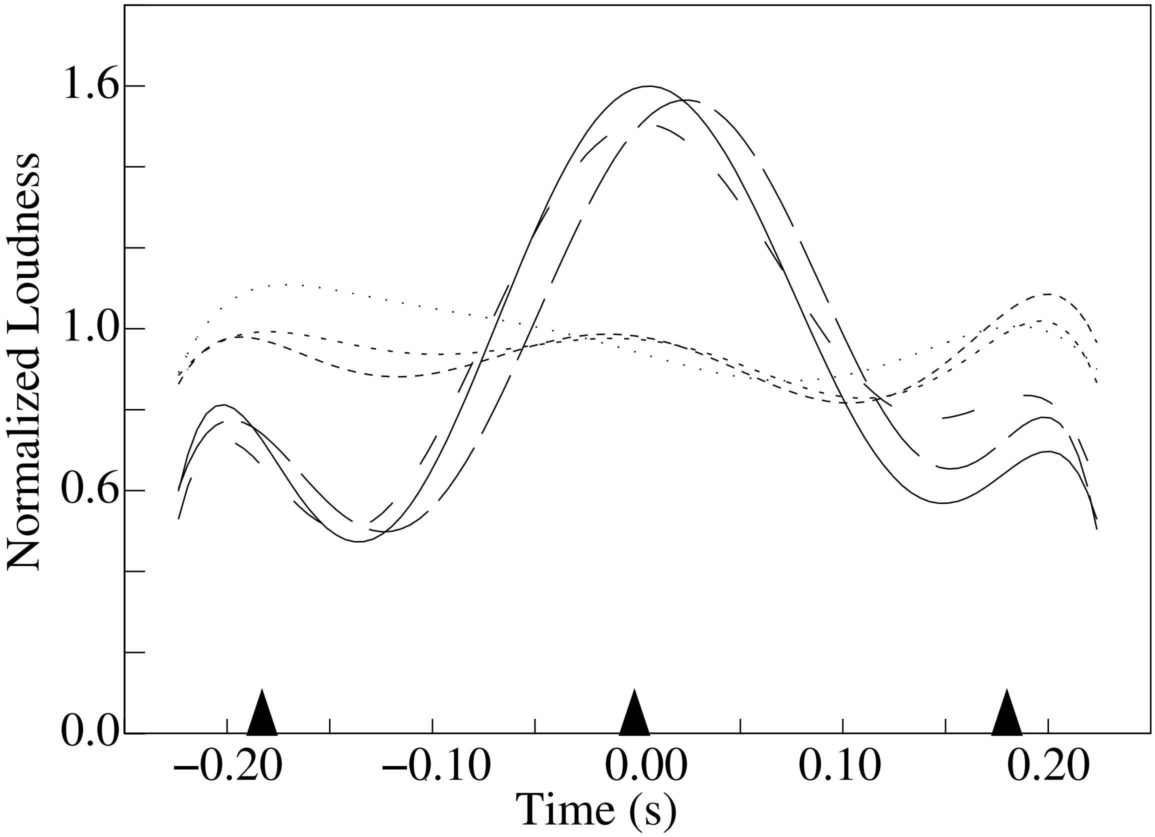

FIG. 8. Reconstructed loudness profiles for prominent (long dashes) and non-prominent (short dashes) syllables, for the primary and two secondary data sets. In each group, the primary data set is plotted with the most ink, followed by secondary sets GK then EL. The black triangles mark the median position of syllable centers. Zero on the time axis corresponds to the prominence mark.

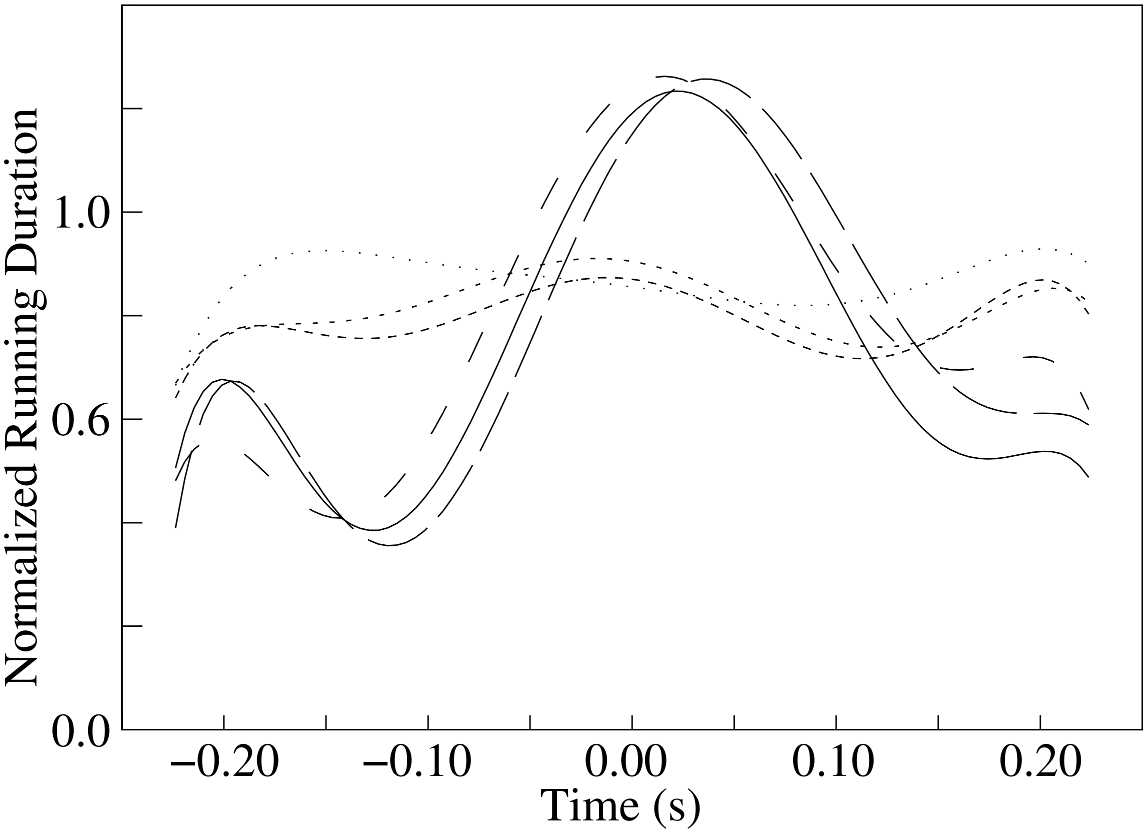

FIG. 9. Reconstructed running duration contours for prominent (long dash) and non-prominent (short dash) syllables. See Figure 8.

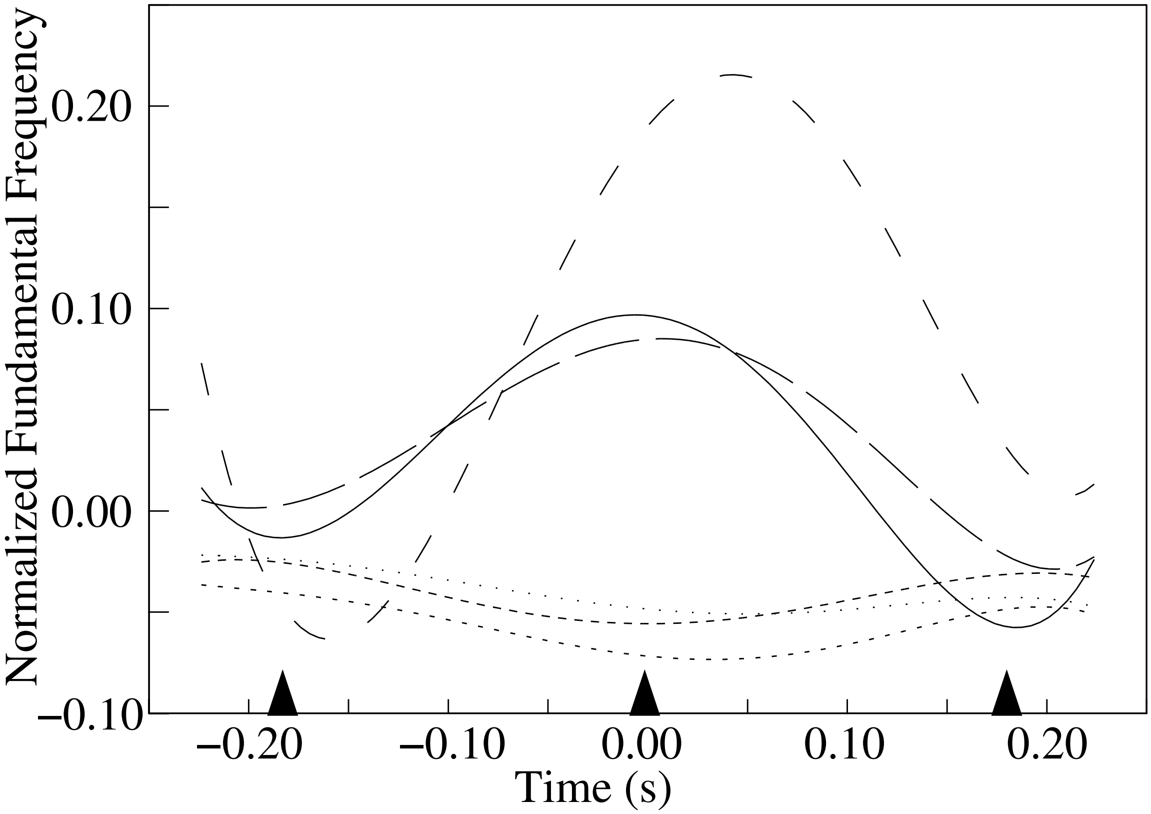

FIG. 10. Reconstructions of the time-dependence of $f_0(t)$. See Figure 8.

and have more high frequency power in voiced regions). As with the other measures, the contrasts are strongest with the preceding neighbor.

Overall, a variety of differences appear between prominent and non-prominent syllables, perhaps extending beyond the vowel into consonantal regions. Furthermore, the acoustic markers for prominence are not restricted to the prominent syllable; contrasts between a syllable and neighboring ones are important. These reconstructions are an acoustical representation of the alternating metrical pattern of English.

III.C. Qualitative limits of the analysis

- We search for patterns of f_0 and other acoustic measures defined in terms of absolute time offsets from the syllable center. If the patterns stretched as syllable durations changed, so that the the positions of peaks and valleys would move, the features will be blurred in this analysis. The classifier would then be unable to make full use of the information in such patterns. This is another possible explanation for the fall of classifier performance for $w >$ 600 ms: duration changes accumulate across the window, so the position of the second- or third- nearest neighbor syllable is correspondingly less certain than the nearest neighbor. However, this effect will strike classifiers with small τ_α first, so the logic (in §III) that implies good efficiency of f_0 classifiers still holds. We are confident that f_0 and its contrasts with adjacent syllables carry relatively little information about prominence.

- The analysis does not take account of position in the utterance or intonational phrase. For instance, final lengthening doubtless dilutes the results based on running duration by introducing a population of long syllables that are only occasionally prominent. Likewise, initial syllables tend to be loud, but are not especially likely to be prominent. This will reduce the K of the loudness-based classifiers.

- We analyze f_0, not pitch. Although the correlation between f_0 and pitch is quite tight for pure tones, there has been less work on the psychophysics of speech-like sounds; perhaps the correlation is weaker. Or, perhaps the linguistic usage of the term “pitch” doesn’t agree with the psychophysical definition of the term.

- We ignore the dependences of the acoustic measures on segmental structure. For instance, /m/ and /s/ have intrinsically different values of aperiodicity. This acts as an extra source of noise in our classification, increasing the class variances relative to the difference between the means, thus reducing K.

- Loudness and duration are correlated in our corpus, so a decision of which of the two is more important may not be completely reliable.

III.D. Quantitative limits of the analysis

As the analysis does not detect strong correlations of f_0 with prominence, we should confirm that the weak result for f_0 is not an artifact of our analysis procedure.

We explored the limits of the analysis procedure by adding in an artificial f_0 component to the prominent syllables between normalization and OP fitting. We repeated the analysis, then adjusted the size of the artificial component until $K_{f_0} \approx 0.5$. This reveals how large the motion of f_0 would have to be for detection by the classifier. Since $K_{f_0} \ll 0.5$ with the unmodified data, this allows us to set an upper limit to the size of f_0 motions that might be associated with prominent syllables.

We first explored the possibility that a locally raised f_0 marked prominence. To check this, we added bumps in the shape of a $\sigma = 100$ ms Gaussian, centered on the prominence mark. The classifiers detected these bumps, reaching $K = 0.5$ when the bump size was 2.4 semitones (about 25 Hz for a speaker with mean f_0 of 170 Hz). Since K is much smaller than that for our unmodified data, we can exclude the possibility that prominence is commonly associated with such an f_0 bump or larger, because the analysis would have detected it. This is a conservative upper limit, as we base the limit on $K = 0.5$, whereas the unmodified f_0 data yielded only $K = 0.12$.

However, a standard assumption in the intonation literature is that many different pitch patterns can lend prominence to a syllable (e.g. Ladd (1996); Cruttenden (1997)). A bump centered on a syllable is only one of many options. Background on this topic can be found in Wichmann *et al.* (1997).

We tested three more patterns to map out more limits of the analysis:

- A region of sloping f_0. A bump in the form $\frac{t-t_c}{\sigma}e^{-(t-t_c)^2/2\sigma^2}$ was added, with $\sigma = 100$ ms. This function has a broad peak 100 ms after the prominence mark, a valley 100 ms before the mark, and a smooth slope in between. It was detected with $K = 0.5$ when the peak-to-valley difference was 2.8 semitones (about 27 Hz), and the slope was 14 semitones/s.

- A region of increased variance of f_0. We used a random mixture of the Gaussian bumps and the sloping contours, above. Instead of using a single amplitude, the amplitudes were chosen from a zero-mean Gaussian distribution. This corresponds to the possibility that prominence is marked by *either* a bump, a dip, a peak-valley pattern a valley-peak pattern, or some mixture thereof. Non-prominence would presumably be indicated by relatively flat contours.

  This choice can generate a very broad range of intonational patterns, covering many of the suggested possibilities. Even with this wide variety of possible f_0 patterns, the classifier reached $K = 0.5$ when the standard deviation of the bump amplitude was 3.1 semitones, along with a 3.8 semitone standard deviation for the peak-to-valley difference for the slope component.

- We added a Gaussian bump with $\sigma = 100$ ms, but we let the amplitude and position vary from prominence to prominence. The bump center was chosen from a $\sigma = 100$ ms Gaussian probability distribution, to simulate random choices of peak alignment, and the amplitude was chosen from a zero-mean Gaussian.

  This corresponds to the possibility that prominence is marked by either an f_0 bump or dip, whose timing is not precisely tied to the syllable center. The analysis was not as effective at this test, detecting it at $K = 0.5$ only when the standard deviation of the normalized amplitude was 0.5, corresponding to 10 st (about 85 Hz).

The limits that this analysis can set on f_0 excursions depend on the complexity of the pattern and the accuracy with which it is anchored to the prominence mark. However, most 1/2 octave motions would be easily detectable, if they existed in the data. The analysis can exclude most f_0 features that have a fixed time-alignment with the syllable center and are larger than 3 semitones.

While we do not categorically rule out f_0 as a indicator of prominence, we do rule out many simple associations of f_0 with prominence. Most of the possibilities that we do not exclude would involve fairly complex patterns and/or rather loose associations between the position of the pattern and the syllable center. It is currently uncertain whether such a tenuous association of f_0 with a syllable is sufficient to communicate the prominence to a human listener.

III.E. Comparison to synthesis experiments

When comparing these results to other studies in the literature, it is important to maintain the distinction between acoustic properties that can induce the perception of prominence and acoustic measures that are actually used to mark prominent syllables. They need not be identical. Speech is only one of several inputs that the human auditory mechanism processes, and other uses, such as monitoring environmental sounds, might define the way the auditory system functions. Additionally, articulatory constraints may make certain ways of inducing prominence easier than others.

This distinction is crucial for understanding the synthesis-based experiments that show that f_0 can induce the perception of prominence (Gussenhoven *et al.*, 1997; Rietveld and Gussenhoven, 1985; Terken, 1991). Despite appearances, their results are consistent with this study, because they use larger f_0 excursions than are normally found in our corpus. For instance, a typical stimulus in these papers above contains a sharp triangular f_0 excursion of 1/2 octave amplitude and a full-width at half-maximum of 200 ms or less.

We looked for such peaks in our database by computing a peak-height statistic h matched to the shapes used in the above synthesis-based studies. We take f_c as

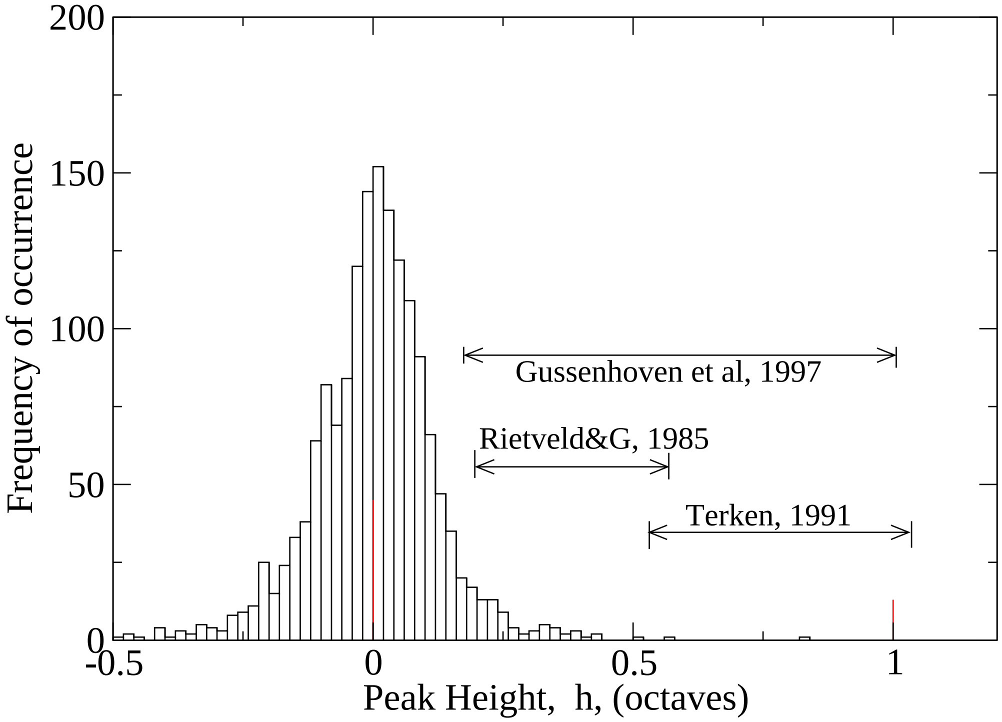

FIG. 11. Peak height statistic, h for prominent syllables in the IViE corpus (histogram). For comparison, the ranges of the f_0 swings used as experimental stimuli in Gussenhoven *et al.* (1997), Rietveld and Gussenhoven (1985), and Terken (1991) are shown.

an average of f_0 over a 50 ms wide region centered on a syllable. The average is weighted with $W_{f_0}(t)$ from §II.F. Similarly, f_e is an average of f_0 over a pair of regions between 100 and 150 ms to the left and to the right of the prominence mark. We then compute $h = \log_2(f_c/f_e)$; this statistic is close to zero for linear f_0 contours and nearly equal to the bump height (in octaves) for contours used in the papers cited above.

For prominent syllables in the IViE corpus, h has an approximately Gaussian distribution (Figure 11) with a standard deviation of h=0.11 and a mean of zero. Only 2% had bump heights exceeding a quarter octave. Most of the stimuli studied in these papers have bumps that are larger than that. Consequently, they studied bumps that are larger than those commonly found in British English. Their results are thus completely consistent with our conclusion that f_0 is relatively unimportant for prominence, as English is normally spoken. Their experiments, like ours, indicate that a 10% pitch change induces little prominence

III.F. Importance of the spectral slope

Our result that K_S is small is somewhat unexpected, given work by Heldner (2001) and Sluijter and van Heuven (1996). Heldner found his spectral emphasis measure to be a good predictor of prominence. However, Heldner's measure is different, and is applied to a different language (Swedish). His measure is the difference between the power in the first harmonic and the rest of the spectrum in voiced regions, and is zero in unvoiced regions. So, his measure obtains almost all its information from the low-frequency parts of the spectrum, mostly below $3f_0 \approx 600$ Hz, unlike ours, which extends up to 3000 Hz. A further difference is that his measure responds differently to voiced/unvoiced distinctions than ours.

Sluijter and van Heuven, consistent with this work, found that syllables with contrastive focus have a flatter spectrum. Their experiment yields a strong effect of spectral slope, but that is expected, as their classification task is far easier. Their sentences were read carefully by speakers instructed to produce contrastive focus on certain words. The authors then selected sentences for a clear contrast between the +FOCUS and −FOCUS versions. Thus, they allowed no ambiguous utterances, Their paired comparison between ±FOCUS renditions of the same word in the same position in an utterance also allows for a more sensitive comparison than is normally available to a human listener to natural speech. They proved that speakers *can* produce contrasts in spectral slope, not that speakers normally *do* produce such contrasts.

IV. FURTHER EXPLORATIONS

IV.A. Comparison with secondary data sets

It might be argued that the similarity of our results between dialects is due to the fact that the same pair of labelers marked each dialect rather than because of an intrinsic similarity. To check this, we conducted the same analysis on the two secondary data sets. Our secondary labelers speak different dialects and are trained differently from the primary labelers. If the process of labeling says more about the labeler than about the speech, the secondary data sets should give substantially different results from the primary data set.

Table III compares the primary and two secondary sets. Inspection of a sample of the marks reveals some disagreements about prominence, some disagreements about the number of syllables (primarily non-prominent syllables), a few unlabeled words in the secondary sets, and a few long syllables where the labelers agree but placed

Comparison	Agreement on Alignment	Agreement of syllables that align
Primary *vs.* GK	84%	73%
Primary *vs.* EL	79%	72%
GK *vs.* EL	75%	84%

TABLE III. Alignment and agreement of syllable and prominence marks between the various data sets. The "alignment" column counts marks that match within 60 ms. Of the aligned marks, the right-hand column counts what fraction agree in terms of prominence/non-prominence judgements.

marks more than 60 ms apart. It is hard to compare these alignment and agreement numbers with the literature (e.g. Yoon *et al.* (2004) and references therein), because published studies of inter-transcriber reliability typically have trained the transcribers to a specific standard in an attempt to minimize the disagreement. In contrast, we wished to find the natural limits of the idea of "prominence," so we did not train labelers.

As can be seen in Figures 8, 9 and 10, the reconstructions of the primary and secondary sets are quite similar. The most obvious discrepancy is that the EL set is shifted about 20 ms later, relative to the marks. This is unimportant, as a review of the marks indicates that EL placed labels slightly earlier in the syllable than the other labelers. Reconstructions for the irregularity and spectral slope measures (not shown) are also similar to reconstructions based on the primary data set. The classifier performance on the primary and secondary sets also match well (Figure 12). These figures suggest that the dialect and academic background of the labeler makes little difference. This supports our main conclusion and suggests that a perceptual, theory-independent definition of prominence may be possible.

The secondary sets of labels also gave an opportunity to check that the algorithm we used to assign the location of non-prominent syllables in the primary was adequate. The agreement of the secondary and primary sets confirms that the automatic generation of non-prominent syllable positions is good enough.

IV.B. Loudness *vs.* RMS *vs.* peak-to-peak

We use a loudness measure rather than the more common RMS amplitude or peak amplitude measurements because the former is a better match to the listener's perception. However, to allow comparison with prior work, we also built classifiers based on common amplitude measures. We computed RMS amplitude by filtering with a fourth-order, 60 Hz, time-symmetric Butterworth high-pass filter, squaring,

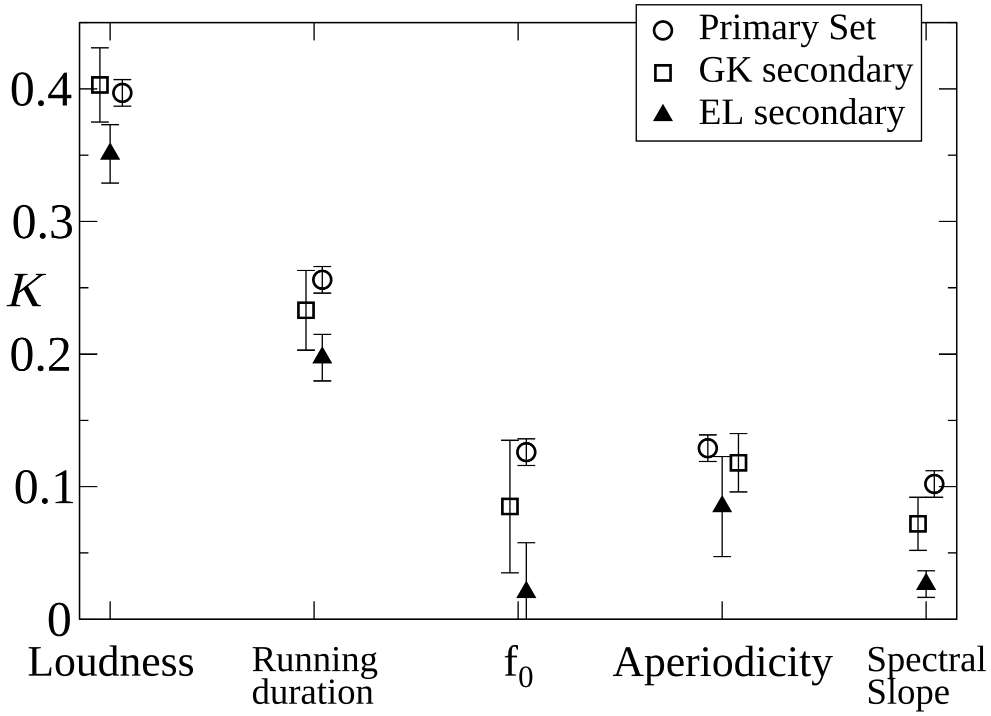

FIG. 12. Classifier performance comparisons between the primary and secondary data sets.

and smoothing with a 15 ms standard deviation Gaussian kernel. Peak-to-peak amplitude was computed by high-pass filtering, then finding the positive peak amplitude by taking the maximum over a 20 ms window centered at each point, and then subtracting a similarly defined negative peak amplitude.

Perhaps surprisingly, there is no substantial difference between the loudness, RMS, and peak-to-peak classifier performance: K_{RMS} is 0.01 lower than K_L, and $K_{\mathrm{peak-to-peak}}$ is 0.03 higher. The similarity between our loudness and RMS intensity results means that our results appear to conflict with the findings of Sluijter and van Heuven (1996) and Sluijter *et al.* (1997), who found that intensity is relatively unimportant. The difference may relate to their experimental conditions, which were rather more formal, to the different language, or some other factor.

IV.C. Combining different acoustic properties

Finally, we built one more classifier to see if information from the other acoustic properties could improve the behavior of a classifier based on loudness. To do this, we took the forest of classifiers and constructed a feature vector for each syllable by counting the fraction of classifiers that labeled it as prominent, for each of the five acoustic measures. The feature vector for each syllable is thus $(F_{f_0}, F_L, F_D, F_A, F_S)$, with each of the F_α in the range $[0, 1]$. It is input for a second-stage classifier, operating on the outputs of the first stage Gaussian Forest classifiers. This is a "bagging" or classifier fusion approach (Breiman, 1996; Kittler *et al.*, 1998; Wolpert, 1992; Huang and Suen, 1995). The second stage classifier is a logistic discriminant classifier (Webb, 1999, pp. 124–132).

The resulting distribution of K across dialects and style is fairly narrow distribution, with $\sigma = 0.07$, $\sigma/K = 0.14$. All dialects seem to be about equally good at marking prominence acoustically. The average K is 0.479 (based on P[chance] for loudness), and $P[\text{correct}] = 0.786$.

To see how much information the other acoustic features are contributing, we can compare the $K = 0.479$ for the combined classifier to to a similar logistic discriminant classifier that is fed only F_L. Such a loudness-only classifier achieves $K = 0.430$. The improvement caused by attaching F_D, F_{f_0}, F_A, and F_S to the feature vector is statistically significant (t=4.2, df=21, $P < 0.001$), but it is not large. The probability of correct classification only increases from 76.6% to 78.6%. This is not completely unexpected: $D(t)$ is correlated with $L(t)$ and the other acoustic measures are generally less effective at classifying syllables than loudness or running duration. We conclude that $D(t)$, $f_0(t)$, $S(t)$ and $A(t)$ do contain some information not present in the loudness, but not very much. This result is not inconsistent with claims that loudness and duration are perceived as a unit (Turk and Sawusch, 1996).

V. CONCLUSION

Prominent syllables are marked by being louder and longer than the previous syllable. Of the two, loudness is the better predictor. However, these two acoustic measures are correlated enough so that distinguishing the effect of the two may not be completely reliable.

Contrary to the common assumption, there is no pattern of f_0 detectable by our analysis that is more than a weak predictor of prominence. Many prominent syllables do indeed have high pitch, but many non-prominent syllables also do. Thus, taking the listeners' point of view, the observation of high pitch does not usually allow the listener to conclude that a syllable is prominent. We found that prominence cannot be usefully distinguished on the basis of local f_0 values, local f_0 changes, or the local variance of f_0. We see no evidence that long f_0 patterns are relevant to the prominence decision.

We do not disagree with the common assumption that dramatic changes in f_0 *can* cause listeners to label syllables as prominent; however, we find that our speakers *do not* normally use this mechanism. They almost never produce the large pitch excursions that are presumably necessary to induce a listener to judge a syllable as prominent. The fact that the labelers were able to consistently mark prominent syllables is clear proof that special f_0 patterns are not necessary near prominent syllables.

All the dialects and styles of speech in our corpus have a similar definition of prominence. We suggest that this definition of prominence could be a feature of most English dialects, as it seems consistent with the work of Silipo and Greenberg (2000) and of Beckman (1986). The definition of prominence also seems independent of the labeler's dialect and academic training.

These results have several implications for linguistics. First, prominence and pitch movements should be treated as largely independent and equally important variables. Prominence has a clear acoustic basis, although metrical expectations may also play some role.

Second, these results raise a puzzle. Individual utterances where prominence seems to be due to large pitch excursions are not hard to find in the literature. Are they simply unusual contours that were selected for their tutorial value, or do they represent another style of speech that is not represented in the IViE corpus? Do people produce large f_0 excursions in certain experiments and not in others?

Third, too much attention may have been focused on f_0. Various authors have assigned f_0 the tasks of communicating emotion, contrastive focus, marking the introduction of new topics and new words, separating declaratives from interrogatives, and helping to separate pairs of words. Perhaps it has been assigned too many tasks. At the least, it seems that f_0 does not normally play a role in signaling the prominent words in a sentence.

ACKNOWLEDGEMENTS

We would like to thank Eleanor Lawson for comments and many hours of labeling, and Chilin Shih for valuable comments. This research was supported by award RES000–23–0149 from the UK Economic and Social Research Council; the IViE corpus was funded by UK ESRC award RES000–23–7145.

Appendix A. SENSITIVITY ANALYSIS - FORM OF WEIGHT FUNCTION

If W were changed, one might expect K to change, since different weight functions emphasize different parts of the syllables. We examined this possibility by picking three new sets of weight functions and re-analyzing the data: [A] All weights (see §II.F) raised to the 0.5 power. This means that the analysis is not focused as strictly on syllable centers: syllable edges contribute more. It also puts more nearly even weights on prominent and non-prominent syllables. [B] All weights raised to the 1.5 power, thus focussing the analysis more tightly toward syllable centers. [C] Changing $W_S(t)$ to $W_S(t) = L^2(t)$, thus including unvoiced regions in the OP fits to the spectral slope data.

None of these results differed much from the default case for any of the acoustic measures: K-values changed by no more than 0.03. We conclude that our weight function is adequate and that changes to them would probably not substantially affect our results.

Appendix B. SENSITIVITY ANALYSIS - ORDER OF POLYNOMIAL FIT

To check that we used the appropriate number of orthogonal polynomials, we ran the same analysis with 20% more or fewer orthogonal polynomials by altering the τ_α. Most changes to the K-values were small, within 0.02 for all acoustic measures, except f_0.

However, the classifiers built from f_0 data showed a trend toward better performance as they were simplified: K_{f_0} increased by 0.055 as τ_{f_0} was increased from 75 ms to 112 ms. To see whether this increase would continue as the f_0 classifiers became even simpler, we recalculated with τ_{f_0}=141 ms and saw no further increase in K. It seems that the optimal classifier for f_0 therefore involves 4 ($\tau_{f_0} = 141$ ms) or 5 ($\tau_{f_0} = 112$ ms) orthogonal polynomials in the analysis window.

These tests show that the results do not depend strongly on the number of polynomials used. Fitting the data more accurately would not substantially improve the classification results for any acoustic measure. Indeed, this check suggests that only the simplest f_0 patterns (those describable by low-order polynomials) carry prominence information.

Appendix C. LOUDNESS NORMALIZATION

The loudness normalization (§II.E) is arguably too severe: by setting the RMS loudness in the window to a constant, it means that the classifier cannot recognize that the analysis window as a whole might be unusually loud. To check whether this is an important limitation, we also computed K-values where we normalized the loudness by dividing by the speaker's overall RMS loudness. This normalization removes inter-speaker differences but preserves all other loudness differences. The average K was little different; it was reduced by 0.03 ± 0.01. The extra information available in this analysis was probably overwhelmed by the increase in loudness variability associated with inter-utterance differences.

References

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

Beckman, M. E. and Edwards, J. (**1994**), "Articulatory evidence for differentiating stress categories," in *Phonological Structure and Phonetic Form*, edited by P. Keating (Cambridge University Press), Papers in Laboratory Phonology III, pp. 7–33.

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`.

Bolinger, D. (**1958**), "A theory of the pitch accent in English," Word: Journal of the International Linguisic Association **7**, 199–210, reprinted in D. Bolinger, Forms of English: accent, morpheme, order, Harvard University Press, Cambridge, MA (1965).

Breiman, L. (**1996**), "Bagging predictors," Machine Learning **26**(2), 123–140.

Clark, J. and Yallop, C. (**1995**), *An Introduction to Phonetics and Phonology* (Blackwell Publishers, Ltd., Oxford), 2nd ed.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (**1985**), "Acoustical aspects of contrastive stress in question/answer contexts," J. Acoustical Society of America **77**(6), 2142–2156.

Cruttenden, A. (**1997**), *Intonation* (Cambridge University Press, Cambridge), 2nd ed.

Eady, S. J. and Cooper, W. E. (**1986**), "Speech Intonation and focus location in matched statements and questions," J. Acoustical Society of America **80**(2), 402–415.

Fletcher, H. and Munson, W. A. (**1933**), "Loudness, its definition, measurement, and calculation," J. Acoustical Society of America **5**, 82–108.

Fletcher, J., Grabe, E., and Warren, P. (**2004**), "Intonational variation in four dialects of English: the high rising tune," in *Prosodic typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford).

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.

Gelman, A. B., Carlin, J. S., Stern, H. S., and Rubin, D. B. (**1995**), *Bayesean Data Analysis* (Chapman and Hall/CRC), first ed.

Grabe, E., Kochanski, G., and Coleman, J. (**to appear**), "Quantitative modelling of intonational variation," in *Proceedings of SASRTLM 2003 (Speech Analysis and Recognition in Technology, Linguistics and Medicine)*, URL `http://kochanski.org/gpk/papers/2004/2003SASRLTM`.

Grabe, E., Post, B., and Nolan, F. (**2001**), "Modelling intonational Variation in English. The IViE system," in *Proceedings of Prosody 2000*, edited by S. Puppel and G. Demenko, pp. 51–57.

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**(5), 3009–3022.

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.

Ho, T. K. (**1998**), "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(8), 832–844, URL `http://csdl.computer.org/comp/trans/tp/1998/08/i0832abs.htm`.

Hochstrasser, U. W. (**1972**), "Orthogonal Polynomials," in *Handbook of Mathematical Funtions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York), pp. 771–802.

Huang, Y. S. and Suen, C. Y. (**1995**), "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," IEEE Trans. Pattern Analysis and Machine Intelligence **17**(1), 90–94.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (**1998**), "On combining classifiers," IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3), 226–239.

Kochanski, G., Shih, C., and Jing, H. (**January 2003**), "Hierarchical structure and word strength prediction of Mandarin prosody," International J. Speech Technology **6**(1), 33–43, URL `http://dx.doi.org/10.1023/A:1021095805490`.

Kochanski, G. P. and Shih, C. (**October 2000**), "Stem-ML: Language Independent Prosody Description," in *Proceedings of the Sixth International Conference on Spoken Language Processing*, vol. 3, pp. 239–242, URL `http://prosodies.org/papers/2000/stemml_2000.pdf`.

Ladd, D. R. (**1996**), *Intonational Phonology* (Cambridge University Press, Cambridge).

Lieberman, P. (**April 1960**), "Some Acoustic Correlates of Word Stress in American English," J. Acoustical Society of America **32**(4), 451–454.

O'Connor, J. D. and Arnold, G. F. (**1973**), *Intonation of Colloquial English* (Longman Group, Ltd, London), 2nd ed.

Passy, P. (**1891**), *Etude sur les changements phonétiques et leurs caractères généraux* (Firmin-Didot, Paris).

Passy, P. (**1906**), *Petite phonétique comparée des principales langues Européenes* (B. G. Teubner, Leipzig & Berlin).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**), *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, New York), 2nd ed.

Rietveld, A. C. M. and Gussenhoven, C. (**1985**), "On the relation between pitch excursions and prominence," J. Phonetics **13**, 299–308.

Roca, I. and Johnson, W. (**1999**), *A Course in Phonology* (Blackwell Publishers, Ltd., Oxford).

Silipo, R. and Greenberg, S. (**August 1999**), "Automatic Transcription of Prosodic Stress for Spontaneous English Discourse," in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS99)*, pp. 2351–2354.

Silipo, R. and Greenberg, S. (**May 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. (**October 1996**), "Spectral balance as an acoustic correlate of linguistic stress," J. Acoustical Society of America **100**(4 part 1), 2471–2485.

Sluijter, A. M. C., van Heuven, V. J., and Pacilly, J. J. A. (**January 1997**), "Spectral balance as a cue in the perception of linguistic stress," J. Acoustical Society of America **101**(1), 503–513.

Stevens, S. S. (**1971**), "Perceived Level of Noise by Mark VII and Decibels," J. Acoustical Society of America **51**(2 (part 2)), 575–602.

Sweet, H. (**1906**), *A Primer of Phonetics* (Clarendon Press).

't Hart, J., Collier, R., and Cohen, A. (**1990**), *A perceptual study of intonation: an experimental-phonetic approach to speech melody* (Cambridge University Press, Cambridge).

Tamburini, F. (**2003**), "Prosodic Prominence Detection in Speech," in *Seventh International Symposium on Signal Processing and its Applications*, pp. 385–388.

Terken, J. (**April 1991**), "Fundamental frequency and percieved prominence of accented syllables," J. Acoustical Society of America **89**(4), 1768–1776.

Terken, J. and Hermes, D. J. (**2000**), "The perception of prosodic prominence," in *Prosody: Theory and Experiment, studies presented to Gösta Bruce* (Kluwer Academic Publishers, Dordrecht), pp. 89–127.

Trager, G. L. and Smith, H. L. (**1951**), *An outline of English structure*, no. 3 in Studies in Linguistics: Occasional Papers (American Council of Learned Societies, Washington).

Turk, A. E. and Sawusch, J. R. (**1996**), "The processing of duration and intensity cues to prominence," J. Acoustical Society of America **99**(6), 3782–3790, URL `doi:10.1121/1.414995`.

Webb, A. (**1999**), *Statistical Pattern Recognition* (Arnold, London; New York).

Welby, P. (**2003**), "Effects of Pitch Accent Position, Type, and Status on Focus Projection," Language and Speech **46**(1), 53–81.

Wichmann, A., House, J., and Rietveld, T. (**sep 1997**), "Peak displacement and topic structure," in *Intonation: Theory, Models and Applications Proceedings of ESCA workshop on Intonation*, edited by B. et al.

Wolpert, D. H. (**1992**), "Stacked Generalization," Neural Networks **5**(2), 241–260.

Yoon, T., Chavarria, S., Cole, J., and Hasegawa-Johnson, M. (**October 2004**), "Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation Using ToBI," in *Proceedings of the ICSA International Conference on Spoken Language Processing*, pp. 2729–2732, URL `http://prosody.beckman.uiuc.edu/pubs/Yoon-etal-ICSLP2004.pdf`, Interspeech 2004.