Finding intonational boundaries using acoustic cues related to the voice source

Jeung-Yoon Choi
Department of Electrical and Computer Engineering, 2005 Beckman Institute, 405 North Mathews Avenue, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering, 155 Everitt Laboratory, 1406 West Green Street, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

Jennifer Cole
Department of Linguistics, 4088 Foreign Languages Building, 707 South Mathews Avenue, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

(Received 1 July 2004; revised 7 February 2005; accepted 2 July 2005)

Acoustic cues related to the voice source, including harmonic structure and spectral tilt, were examined for relevance to prosodic boundary detection. The measurements considered here comprise five categories: duration, pitch, harmonic structure, spectral tilt, and amplitude. Distributions of the measurements and statistical analysis show that the measurements may be used to differentiate between prosodic categories. Detection experiments on the Boston University Radio Speech Corpus show equal error detection rates around 70% for accent and boundary detection, using only the acoustic measurements described, without any lexical or syntactic information. Further investigation of the detection results shows that duration and amplitude measurements, and, to a lesser degree, pitch measurements, are useful for detecting accents, while all voice source measurements except pitch measurements are useful for boundary detection. [DOI: 10.1121/1.2010288]

PACS number(s): 43.70.Fq, 43.70.Gr, 43.72.Ne [AL] Pages: 2579–2587

I. INTRODUCTION

Speech prosody comprises many forms of nonsegmental and nonlexical information. Prosody is related to emotion,[1] and selective emphasis,[2] and helps to resolve syntactically ambiguous utterances.[3] However, there is much debate as to the structure of prosody, as well as the acoustic manifestations of those structures.[4]

One widely used system, TOBI,[5] has been formulated based on phonological study of prosodic units. The system assigns tones to each prominence and boundary, and a break index to each boundary. Briefly, tones are classified as low (L), high (H), or downstepped-high (!H). Prominences are marked with stars (*), and intermediate boundaries with dashes (-). Full intonational boundaries are marked with both a dash (-) and a percent (%). Tones may be combined to form composite tones, such as L^*+H and L-H%. It is assumed that these entities have corresponding acoustic (and presumably articulatory) correlates that may be observed directly.

If so, in order to determine where the acoustic cues for prosody may be found, it is useful to consider which part of the speech signal contains the most information related to prosody. In previous works, sentential prosody imposed on nonsense syllables was found to facilitate short-term memory for the stimuli. Also, based on prosody, listeners were able to choose the correct interpretation when ambiguous sentences were in reiterant form as well as they did when the sentences were spoken normally.[6]

A few observations may be made from these results. One is that listeners are able to perceive acoustic indicators for prosody independently of syntactic or lexical information. Also, the acoustic indicators should be contained within the temporal, spectral, and amplitudinal dimensions of the signal. Another observation is that, when normal speech is transformed into reiterant speech or nonsense syllables, acoustic features of the vocalic segments appear to be rich channels for the transmission of prosodic information. Although a large number of syllables in natural speech contain vocalic segments in the onset and/or the coda, it is also the case that these segments are often absent, while all syllables include a vocalic nucleus. Therefore, in this study, the focus will be on finding the correspondence between characteristics of vocalic segments in the nucleus with prosodic events.

Vocalic sound can be modeled as the glottal voice source filtered by the vocal tract. The identity of the vocalic segment being produced is mostly determined by the formant structure, which corresponds to the filtering by the vocal tract. Hence, it is reasonable to expect that study of acoustic cues that characterize the glottal voice source for correlates of prosodic events will yield useful results.

The glottal voice source can be modeled as a series of glottal pulses, whose spectrum can be broadly characterized by the spacing and relative amplitudes of the component harmonics. The spacing of the harmonics is determined by the fundamental frequency, or pitch, of the glottal pulse train. The relative amplitudes of the harmonics are affected by the shape of the glottal pulse itself. For example, a larger portion of the fundamental period where there is nonzero airflow (open quotient) leads to a more dominant first harmonic am-

plitude. The faster the glottal pulse returns to zero after the peak, the smaller the amplitude of the first harmonic.[7] These characteristics are results of manipulation of the laryngeal structures that are employed in phonation during vocalic segments.

In a study by Klatt and Klatt,[8] the authors present evidence that declarative sentences may be terminated such that the arytenoid cartilages begin to separate in preparation for breathing, leading to a breathy-voiced offset to the final syllable. This early abduction gesture may be implemented as a general "relaxed" separation of the arytenoids, or a "laryngealized" mode where the abduction is accompanied by a rotational motion of the arytenoids so that medial compression is applied.

For both cases, the spectral noise increases due to the presence of the posterior interarytenoid separation. In the first case ("breathy"), the glottal waveform exhibits a larger open quotient, so that the first-harmonic amplitude ($h1$) is increased and the harmonic spectrum tilts down. The converse is true for the laryngealized case, i.e., $h1$ and spectral tilt both decrease. In the paper by Klatt and Klatt,[8] cues to breathiness increased for unstressed syllables, for final syllables, and at the margins of voiceless consonants. In stressed vowels with a relatively high fundamental frequency, the spectrum showed little evidence of breathiness. On the other hand, many utterances appeared to end in a "breathy-laryngealized" type of vibration, along with diplophonic irregularities in the timing of glottal periods. This observation agrees with findings where aperiodicity associated with creaky voice was more frequent at L-L% than L-boundaries.[9]

An underlying mechanism for such irregular phonation may be due to changes in subglottal pressure. In previous studies, including a study by Slifka,[10] utterance endings were found to be correlated with a drop in subglottal pressure, and irregular phonation with partially spread vocal folds was frequently observed. Additionally, the first stressed syllable in an utterance was found to be correlated with the initial subglottal pressure peak.

Also, in a work by Pierrehumbert and Talkin,[11] harmonic structure (and amplitude) measurements for /h/ and glottal stop were found to become more vocalic (i.e., less aspiration, more phonation) at accents, and more consonantal (i.e., more aspiration, less phonation) at boundaries. The observations above suggest that acoustic measurements related to the glottal voice source will be useful in finding the various types of prosodic events.

In addition to the spectral characteristics of the voice source, temporal and amplitudinal characteristics may be observed, such as length and amplitude of the vocalic segment, and the length of surrounding speech/nonspeech units. Previous studies show that durational cues, such as segmental durations, were found to be correlated with the presence of prosodic boundaries.[12–14]

In light of the discussion above, an attempt has been made to find acoustic cues associated with accents and boundaries by examining measurements related to the glottal source for vocalic segments. The glottal characteristics can be found by examining durational, spectral, and amplitudinal measurements. This study will focus on finding these acoustic cues independent of other knowledge, such as syntactic structure, part of speech of constituent words, or the identity of the words or segments.

II. EXPERIMENTS

A. Prosodic units

The prosodic units that are considered in this study are based on the TOBI system and can be described as two types—boundaries and accents. The first type includes markers for discourse, turn, intonational boundaries, intermediate boundaries, and words. The second type includes markers for phrase-level stress or prominence, and are localized to accented syllables. In this study, boundary will be limited to intermediate (ip) and intonational (IP) boundaries.

Each syllable in a section of speech can be assigned to one of six categories: (1) not accented and not at a boundary (0), (2) accented but not at a boundary (A), (3) not accented and at an ip boundary (ip), (4) accented and at an ip boundary (Aip), (5) not accented, at IP boundary (IP), and (6) accented, at IP boundary (AIP). The type of accent and/or boundary may be used to further specify the syllable environment. In the Boston University Radio Speech Corpus,[15] intermediate boundaries with low, downstepped high, and high tones are indicated by L-, !H-, and H-, respectively, and intonational boundaries, comprising one intermediate and one final tone, are indicated by L-L%, L-H%, !H-L%, H-L%, and H-H%. (The !H-H% IP boundary was not observed in this corpus.) Simple accents are indicated by L^*, $!H^*$, and H^*. Complex accents are indicated as combinations of two accent tones such as L^*+H, etc. In this study, prosodic events marked with a question mark (?) in the corpus were ignored, i.e., syllables associated with those incompletely marked events were considered to be prosodically unmarked. Based on counts of these events, about 9% of the syllables marked as neutral (0) are estimated to belong to this category. An exhaustive tally of the different types of prosodic markers in the corpus includes 90 markers that can be assigned to a syllable (1 neutral, 8 accents, 3 intermediate +6 intonational boundaries, and 24+48 accented at intermediate/intonational boundaries). However, the focus on this study will be on examining the six broad classes described above.

B. Acoustic measurements

The acoustic cues related to the voice source examined in this paper to identify the presence of prosodic events can be divided into five categories: duration, pitch, harmonic structure, spectral tilt, and amplitude.

Durational measurements include length of following pause (if any) and speech rate. A pause is defined to be an interval of speech where the probability of voicing is below 0.5, and rms energy is below 150, for longer than 30 ms, as extracted automatically using the *formant* command in the *xwaves* package.[16] If no such interval exists after a syllable, the pause length was assigned as zero. It is expected that pause length will be longer after boundaries.[12] Speech rate, calculated as the reciprocal of the length of the vocalic seg-

ment, is expected to be slower at boundaries (final lengthening effect).[13,14] Here, a vocalic segment is defined as all vowels (including diphthongs), syllabic liquids (e.g., phones labeled as /er/, /el/), and syllabic nasals (e.g., phones labeled as /em/, /en/, /eng/). The start and end times of the vocalic segments were identified using the phone labels in the *.lbl* or *.lba* files (see Sec. II C). Speech rate was not normalized, either for speaker or phone identity.

Pitch measurements include fundamental frequency (f0) measured at the end of a vocalic segment, and the slope and the convexity of the fundamental frequency over the vocalic segment. The fundamental frequency was measured at the last frame of a vocalic segment where the probability of voicing was above 0.95. Pitch contours for boundary tones have been described as occuring from the nuclear pitch to the end of the utterance, so that the ending fundamental frequency value of the last syllable of an intermediate or intonational phrase would be more similar for utterances with similar boundary tones, but with different numbers of syllables from the nuclear pitch to the end of the utterance. The convexity was calculated as the sum of the difference between each signal point and the linear interpolation between the start and end values of the segment. That is,

$$\frac{\sum_{t=t_1}^{t_2} s(t) - h(t)}{t_2 - t_1},$$

where t_1 and t_2 are respectively the start and end times of the vocalic segment, $s(t)$ is the value of the measurement at time t, and $h(t)$ is the linear interpolated function,

$$h(t) = \frac{s(t_2) - s(t_1)}{t_2 - t_1}(t - t_1), \quad \text{for } t_1 \leqslant t \leqslant t_2, \quad t_1 < t_2.$$

The normalized pitch (nf0) at the end of the vocalic segment, and the slope and convexity of the normalized pitch over the segment are also included in the pitch measurements. Normalized pitch was obtained by training a three-mixture Gaussian distribution over all fundamental frequency measurements of an utterance (sentence), where the means were constrained to be 0.5μ, μ, and 2μ, to separate pitch-halved and pitch-doubled measurements. The outlying values are eliminated, and the resulting fundamental frequency values are smoothed. Normalized pitch values are then computed to be between 0 and 2, with 1 corresponding to the mean fundamental frequency of the utterance.[17,18] This measure normalizes with respect to the speaker over the utterance.

Pitch measurements are expected to be lower for L- and L-L% boundaries, and higher for H- and H-H% boundaries. Pitch is also thought to be affected by low or high accent, but this effect was not examined in this study. However, since the majority of the accents in the corpus were of the H^* type (see Sec. II C), it may be assumed that, mostly, pitch will be *increased* by the presence of accent. Pitch slope is expected to fall for L- or L-L% boundaries, remain mostly level for H-L% boundaries, and rise for H-, H-H%, or L-H% boundaries. Pitch convexity is expected to be flat for L-, L-L%, H-, or H-H% boundaries and to be upward (more positive) for a H-L% boundary and downward (more negative) for a L-H% boundary.

Harmonic structure measurements included end value, slope, and convexity of $h1-h2$, where $h1$ and $h2$ are the amplitudes of the first and second harmonics, respectively. The values for the amplitudes of the first harmonic $h1$ were measured as the amplitude of the spectral component closest to the fundamental frequency, using the spectral analysis procedures in *xwaves*. The amplitude of the second harmonic $h2$ was measured as the amplitude of the spectral component closest to twice the fundamental frequency. In this study, the harmonic measurements were made uniformly for all vocalic segments—the influence of the first formant or presence of liquids and nasals were not factored into the analysis. A more detailed procedure could be used to compensate for these effects, but that approach was not explored in this study.

The $h1-h2$ measurement is greater for a larger open quotient, corresponding to a breathy voice; conversely, it is lesser for a smaller open quotient in a laryngealized voice. It is expected to be lower for accented syllables, which were described as being less breathy. At boundaries, a more breathy offset is expected to correspond to a larger $h1-h2$ measurement, while a more laryngealized offset should produce a smaller measurement.

Spectral tilt measurements included end value, slope, and convexity of $h1-a1$, $h1-a3$, and $a1-a3$, where $a1$ and $a3$ are the amplitudes of the first and third formants, respectively. For both breathy and laryngealized phonation, higher noise and harmonic components relative to the lower frequency amplitudes are observed. Consequently, spectral tilt measurements are expected to exhibit smaller differences at breathy/laryngealized L- or L-L% boundaries and larger differences for more modal H- or H-H% boundaries.

Finally, amplitude measurements included the end value, slope, and convexity of the rms, and these are expected to be lesser at boundaries and greater at accents.

The pitch, harmonic, formant, and rms values were found automatically using the *formant*, and *sgram* commands in *xwaves*. The voice source measurements described above were made over each single vocalic interval, found using the phone labels in the *.lbl* or *.lba* files.

C. Database

The acoustic cues characterizing the voice source were obtained for each vocalic segment in analysis, training, and test subsets of the Boston University Radio Speech Corpus.[15] The corpus contains radio news stories from seven speakers and is divided into two sections, *labnews* and *radio*. Prosodic labels are available for five speakers, f1a, f2b, f3a, m1b, and m2b. In this study, an analysis set was picked to include all files in the *labnews* section with both TOBI labels (i.e., *.ton* files) and hand-corrected (automatically generated) phone labels (i.e., *.lbl* files). The analysis set includes 22 stories from speaker f1a and 41 stories from speaker f2b, both of whom are female. The training set includes 36 stories each from one male speaker (m1b) and one female speaker (f2b) from the *radio* section. The test set includes 24 stories each from one male speaker (m2b) and 1 female speaker (f3a) from the

TABLE I. Number of tokens for each prosodic category for the analysis, training, and test subsets of the Boston University Radio Speech Corpus. The counts for ip boundaries and IP boundaries are further divided into subcategories. (The !H-H% IP boundary was not observed in this corpus.)

	Analysis		Training		Test	
	unacc	acc	unacc	acc	unacc	acc
nonbnd	4944	2193	4849	2133	3994	1958
ip bnd	356	130	213	127	178	121
L-	203	32	110	47	56	16
!H-	74	29	42	35	79	49
H-	79	69	61	45	43	56
IP bnd	732	340	583	236	357	121
L-L%	373	188	335	144	214	68
L-H%	329	116	246	81	126	31
!H-L%	6	9	0	3	0	0
H-L%	21	24	2	8	17	20
H-H%	3	3	0	0	0	2

labnews section. The analysis, training, and test subset utterances are all disjoint. The speakers in the test subset are not included in the analysis or training subsets. For the analysis and training subsets, utterances from speaker f2b appear in both subsets, but only utterances from the *labnews* section are included in the analysis set, and utterances from the *radio* section are included in the training set. Only automatically generated phone labels (i.e., *.lba* files), which are not as accurate as the hand-corrected labels, were available for the training and test data. The measurements were made using the start and end times of vocalic segments included in the *.lbl* files (analysis set) or the *.lba* files (training and test sets).

Each syllable in the three data sets was assigned to one of six prosodic categories described in Sec. II A using the TOBI labels in the *.ton* files, the word transcription in the *.wrd* files, and the phone labels in the *.lbl* or *.lba* files. First, accent markers in the TOBI label files were matched with each vocalic segment in the phone label files, to find accented syllables. Next, the final vocalic segment of a word was found using the word transcriptions and the phone labels. These word-final vocalic segments were then matched with intermediate and intonation boundary markers in the TOBI label files, to find vocalic segments at boundaries. Finally, all vocalic segments were assigned to one of the six prosodic categories, by combining the list of vocalic segments in the accented and boundary categories. If a vocalic segment appeared in neither list, it was assigned to the unaccented, nonboundary (0) category.

The number of syllables at each prosodic category for the three data sets are shown in Table I. There are more full intonational boundaries (IP) than intermediate boundaries (ip) for all data sets. Of the intermediate boundaries, a greater number of L- boundaries occur for the analysis and training sets; more !H- boundaries are observed for the test set. Of the intonational boundaries, the largest number are L-L% boundaries, followed by L-H% boundaries. For all data sets, there were few or no occurrences of the !H-L%, !H-H%, H-L%, and H-H% boundaries. The training set did not include any H-H% boundaries, and the test set did not include any !H-L% boundaries. The ratio of nonboundary syllables to ip+IP boundary syllables is about 5.96 for the three data sets, implying an average phrase length of about six syllables.

Accents were mostly of the H^* type (i.e., H^*, $L+H^*$) accents, followed by comparable numbers of the $!H^*$ type (i.e., $!H^*$, $H+!H^*$, $L+!H^*$) and L^* type (i.e., L^*, L^*+H, $L^*+!H$) accents. For example, in the analysis set, there were 1885 H^* type accents, 447 $!H^*$ type accents, and 167 L^* type accents. More detailed counts for each type of accent are not presented in this paper.

The distribution of tones found in this corpus may be a characteristic of the prosody of broadcast news—a dialogue may include more questions, leading to a greater number of high boundary tones.

III. RESULTS

A. Distributions

Figure 1 shows the distributions of six representative measurements for different prosodic categories from the analysis set, i.e., neutral, the three ip boundaries, L-, !H-, and H-, and the two frequently occurring IP boundaries, L-L% and L-H%. Overall, the presence of an accent tends to lengthen pauses, decrease speech rate, and increase f0, $h1$-$h2$, $h1$-$a1$, and rms. Pause length remains similar for syllables not at boundaries and at ip boundaries, but increases for IP boundaries. Speech rate becomes slower as boundary level increases. The f0, $h1$-$h2$, and $h1$-$a1$ spectral measurements show slight decrease as the boundary level increases, but the effect is not as pronounced as for durational measurements. Amplitude of the syllable decreases as the boundary level increases.

There is little difference in pause length of the different types of ip boundaries, but f0, $h1$-$h2$, $h1$-$a1$, and rms all show slight increases in the order of L-, !H-, and H-ip boundaries. Pause length is longer for the L-L% IP boundaries, compared to the L-H% IP boundaries. However, speech rate seems to show less distinction between the two groups. The f0 measurements show that the presence of a high boundary tone H% leads to an increase. For the $h1$-$h2$

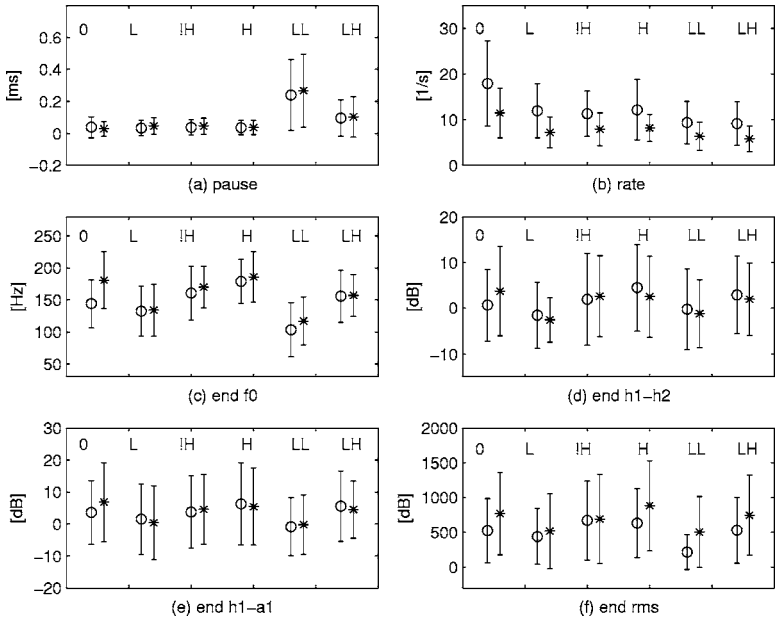

FIG. 1. Distributions of representative measurements for six prosodic categories. The circles/stars are the means, and the bars denote the standard deviations. Each pair of distributions represent unaccented (o) and accented (∗) tokens. The pairs show, in order, non-boundary (0), the L-, !H-, and H-intermediate boundaries (ip), and the L-L% and L-H% intonational boundary (IP) distributions.

and $h1$-$a1$ measurements, accented syllables at boundaries show slightly lower values than unaccented syllables. However, it appears that accented syllables which are not at boundaries show higher values, contrary to expectations, implying that the presence or absence of a boundary affects the production of accented syllables. For example, for this analysis set, the difference in f0 values between accented and unaccented syllables that are not at boundaries seem to be greater than the difference for syllables at boundaries.

Finally, amplitude measurements also show a distinction between the two IP boundaries, with syllables at L-L% IP boundaries showing a lesser amplitude. It must be noted that the standard deviations of the plots are great, with much overlap between the different groups.

B. Statistical analysis

The measurements obtained for all syllables in the training subset of the Radio News Corpus were examined using an analysis of variance (ANOVA). First, a two-way analysis was performed for each of the 23 voice source measurements, with the first factor being boundary level (nonboundary, ip boundary, IP boundary) and the second factor being presence of accent (unaccented, accented). Next, a one-way analysis was carried out, with syllables at all IP boundaries as one group, and the rest as the other group. The F and partial η^2 results are listed in Table II. The critical value for probabilities to be considered significant is $P<0.05$ divided by 46 (23 measurements $\times$ 2 analyses), corresponding to a critical value of $P<0.001$. The degree of freedom between groups is 2 for the first factor (boundary), 1 for the second factor (accent), and 2 for the interaction (boundary $\times$ accent), and 8135 within groups for the two-way analysis. The degree of freedom between groups is 1, and within groups is 8139 for the one-way analysis.

Table II shows that overall, except for the spectral tilt measures related to $h1$-$a3$, the measurements showed significant differences for the two analyses.

TABLE II. ANOVA results (F and partial η^2 values) for 23 voice source measurements for the training data set. Two-way analysis with factor 1: boundary level $\{0,\text{ip},\text{IP}\}\times$ factor 2: accent level $\{0,^*\}$ is shown in columns 2–4. One-way analysis with group 1: IP boundary and group 2: non-IP boundary is shown in the last column. Entries with probabilities greater than ($P>0.001$) (critical value with study correction) are not significant and marked with a dash (-). The degree of freedom between groups for the two-way analysis is 2 for the first factor, 1 for the second factor, and 2 for the interaction; the degree of freedom within groups is 8135, for all measurements. For the one-way analysis, the degree of freedom between groups is 1, and within groups is 8139, for all measurements. Partial η^2 values are shown in parentheses.

Measurements	acc	bnd	acc × bnd	IP vs. not
pause	-	805.6(0.165)	15.4(0.004)	1725.7(0.175)
rate	216.8(0.026)	237.3(0.055)	-	410.9(0.048)
f0 end	90.9(0.011)	92.2(0.022)	33.3(0.008)	122.7(0.015)
f0 slp	-	10.6(0.003)	30.1(0.007)	-
f0 cnv	93.7(0.011)	36.7(0.009)	-	98.3(0.012)
nf0 end	144.3(0.017)	336.7(0.076)	33.9(0.008)	547.1(0.063)
nf0 slp	12.3(0.002)	27.5(0.007)	30.7(0.007)	-
nf0 cnv	103.2(0.013)	37.2(0.009)	-	107.6(0.013)
h1-h2 end	-	11.9(0.003)	-	12.7(0.002)
h1-h2 slp	-	-	-	-
h1-h2 cnv	-	13.3(0.003)	-	35.0(0.004)
h1-a1 end	-	24.8(0.006)	-	41.7(0.005)
h1-a1 slp	-	-	-	-
h1-a1 cnv	-	13.5(0.003)	-	38.0(0.005)
h1-a3 end	-	-	-	-
h1-a3 slp	-	-	-	-
h1-a3 cnv	-	-	-	-
a1-a3 end	-	9.0(0.002)	-	11.5(0.001)
a1-a3 slp	-	-	-	-
a1-a3 cnv	14.0(0.002)	-	-	-
rms end	71.2(0.009)	49.5(0.012)	-	106.6(0.013)
rms slp	-	-	-	-
rms cnv	91.9(0.011)	39.9(0.010)	-	43.9(0.005)

For the two-way analysis, pause length was different for the three boundary levels, but not between unaccented/accented groups. Speech rate was different for all six groups. Except for f0 slope, all pitch measurements showed significant differences across boundary levels and accent levels. The harmonic and spectral tilt measurements show less significant differences, but end measurements for $h1$-$h2$ and $h1$-$a1$, and to a lesser degree, $a1$-$a3$, were significant indicators for boundaries. Only $a1$-$a3$ convexity was significant for accent, however. For amplitude measurements, end rms and rms convexity were significant. Overall, compared to the duration, pitch, and amplitude measurements, harmonic and spectral tilt measurements show small, though significant results.

The interaction between boundary and accent factors are listed in column 4. There is no significant interaction for spectral tilt measurements and amplitude measurements, implying that these measurements show additive effects for boundary and accent. This can be observed in the plots for rms amplitude in Fig. 1(f), from the analysis data.

In the one-way analysis, measurements that were significant indicators for boundary level remained mostly significant for IP versus non-IP discrimination, but pitch slope measurements were not significant.

Partial η^2 values show that effect size is greatest for pause, rate, and rms values, followed by fundamental frequency measurements. The harmonic and spectral tilt measurements show smaller effect sizes, with values less than 0.01, or 1% of the error. (Partial η^2 values are found as $SS_{\text{effect}}/(SS_{\text{effect}}+SS_{\text{error}})$, where SS_{effect} is the type III sum of squares of the measurement, and SS_{error} is the type III sum of squares of the error, respectively.)

C. Boundary detection

The 23 voice source acoustic cues were next used to detect accent and boundary for the training and test sets from the Boston University Radio Speech Corpus. Two boundary detectors were trained and tested: one was trained to detect any IP or ip boundary, the other was trained to detect only IP boundaries, and ignore ip boundaries. The training data set was used to find means and covariance matrices for 23-dimensional Gaussian distributions for nonaccented versus accented tokens (accent detection), for nonboundary versus ip and IP boundary tokens (IP+ip boundary detection), and for non IP versus IP boundary tokens (IP boundary detection). The trained parameters were then used to assign each token in the test set to one group for each task using a simple maximum likelihood measure.

The detection rate versus insertion rate (i.e., receiver operating curve) for accent detection on the training and test data are shown in Fig. 2. Equal error detection rates are about 74.4% for the training set and 70.4% for the test set using all the measurements. The detection rate using only pause, rate, and amplitude (dur+rms) is 71.1%. This curve is included as a baseline to compare the results of including the pitch, harmonic, and spectral tilt measurements. For the case

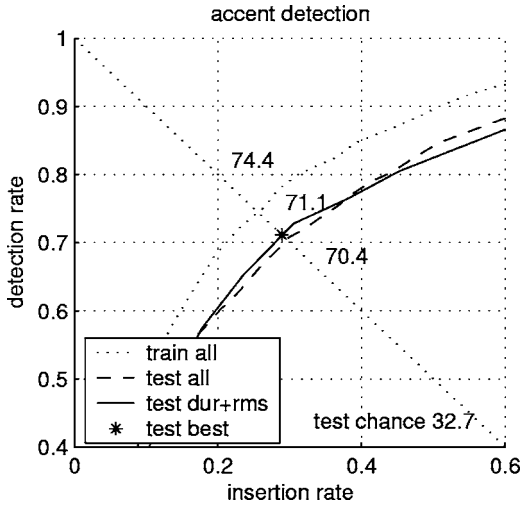

FIG. 2. Detection rate versus insertion rate for accent detection. The diagonal dotted line indicates equal error. Using 23 voice source measurements, equal error detection rates are about 74.4% and 70.4% for the training and test sets, respectively. A best detection rate of 71.1% is obtained using a selected subset of the measurements, and chance is 32.7%, for the test set. A baseline detection rate of 71.1% is obtained using only pause, rate, and amplitude measurements. For accent detection, the baseline and best detection rates are the same.

of accent detection, the baseline performance using dur+rms was better than the results using additional measurements.

The results for detecting both IP and ip boundaries on the training and test data are shown in Fig. 3. Equal error detection rates are about 75.3% and 69.0% for the training and test sets, respectively. The dur+rms detection rate is 69.4%. Finally, as shown in Fig. 4(c), for IP boundary detection, equal error detection rates are about 79.8% and 74.2% for the training and test sets, respectively, with a dur+rms detection rate of 73.7%. For the IP+ip and IP boundary de-

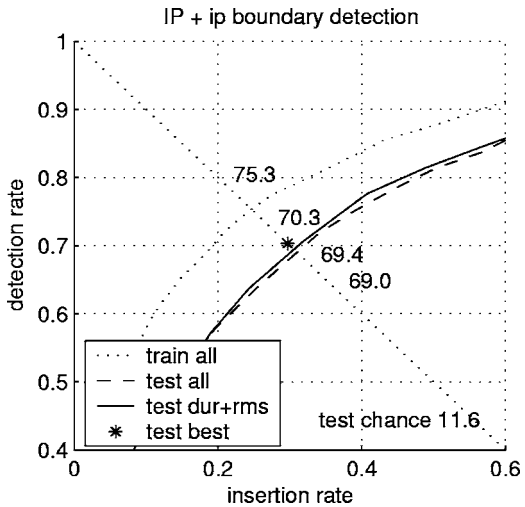

FIG. 3. Detection rate versus insertion rate for IP+ip boundary detection. The diagonal dotted line indicates equal error. Using 23 voice source measurements, equal error detection rates are about 75.3% and 69.0% for the training and test sets, respectively. A best detection rate of 70.3% is obtained using a selected subset of the measurements, and chance is 11.6%, for the test set. A baseline detection rate of 69.4% is obtained using only pause, rate, and amplitude measurements.

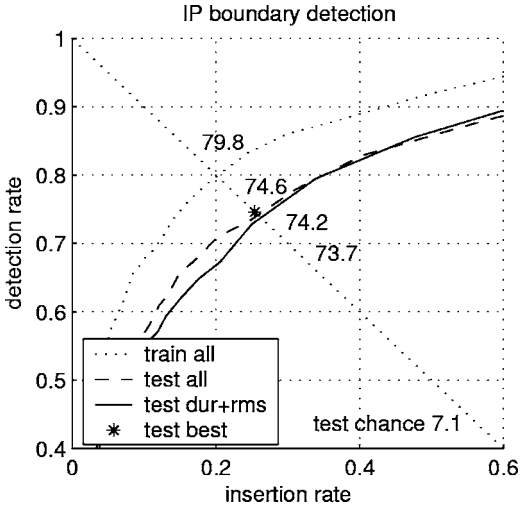

FIG. 4. Detection rate versus insertion rate for IP boundary detection. The diagonal dotted line indicates equal error. Using 23 voice source measurements, equal error detection rates are about 79.8% and 74.2% for the training and test sets, respectively. A best detection rate of 74.6% is obtained using a selected subset of the measurements, and chance is 7.1%, for the test set. A baseline detection rate of 73.7% is obtained using only pause, rate, and amplitude measurements.

tection experiments, using additional measurements results in improved detection rates, compared to the dur+rms detection rates.

To further explore the contributions from various components, detection experiments were carried out using subsets of the 23 voice source measurements. Equal error detection rates for the various subsets are shown in Table III for detection of accent, IP+ip boundary detection, and IP boundary detection. The dur subset (1) includes pause length and speech rate; the rms subset (2) includes end rms, rms slope, and rms convexity; and the pitch subset (3) includes end f0, f0 slope, f0 convexity, end nf0, nf0 slope, and nf0 convexity. The harms subset (4) includes end $h1$-$h2$, $h1$-$h2$ slope, and $h1$-$h2$ convexity, while the tilt subset (5) includes end value, slope and convexity of h1-a1, h1-a3, and a1-a3 measurements. These subsets (i.e., dur, rms, pitch, harms, tilt) were considered the basic component subsets. The harms and tilt subsets were further combined into the glottal subset (6).

From the table, minimum detection rates using the basic component subsets occurred for the tilt, pitch, and harms subsets for accent, IP+ip, and IP detection, respectively. Maximum detection rates were found for the dur subset for all tasks, showing that using pause length and speech rate gives detection rates that approach detection rates using all measurements. For IP+ip and IP boundary detection, pitch and harms measurements are the least useful, while tilt information seems to be somewhat useful. Combining harms and tilt measurements into the glottal subset leads to improvement in detection rates for all three tasks.

Adding glottal measurements to the dur, rms, and pitch subsets (subsets 7–9) increases detection rates for accent detection. However, adding pitch measurements to the glottal measurements decreases performance. Combining basic component subsets (10–12) again results in better detection rates for accent detection, and worse detection rates for IP +ip and IP boundary detection when pitch measurements are

TABLE III. Equal error detection rates using subsets of the 23 voice source measurements for accent, IP+ip, and IP boundary detection. The numbers in parentheses next to the subset names indicate the number of measurements included. Minimum detection rates for each task are shown in boldface; maximum detection rates are shown in underlined boldface. The numbers in parentheses next to the detection rates show minimum detection rate changes from component subsets. A negative number indicates a worse detection rate than a component subset.

Subset	acc	IP+ip	IP
1 **d**ur (2)	66.6	68.6	71.7
2 **r**ms (3)	65.5	62.2	68.0
3 **p**itch (6)	63.2	**51.7**	**55.9**
4 harms (3)	61.8	57.2	**55.8**
5 tilt (9)	**61.0**	61.2	60.2
6 **g**lottal (12)	64.0(2.2)	62.6(1.4)	62.3(2.1)
7 **d+g** (14)	67.4(0.8)	69.7(1.1)	72.9(1.2)
8 **r+g** (15)	69.3(3.8)	65.6(3.0)	67.6(−0.4)
9 **p+g** (18)	67.3(3.3)	56.7(−5.9)	58.8(−3.5)
10 **d+r** (5)	**71.1**(4.5)	69.4(0.8)	73.7(2.0)
11 **d+p** (8)	69.0(2.4)	65.6(−3.0)	72.3(0.6)
12 **r+p** (9)	67.2(1.7)	55.4(−6.8)	58.8(−9.2)
13 **d+r+g** (17)	70.3(−0.8)	**70.3**(0.6)	**74.6**(0.9)
14 **d+p+g** (20)	69.4(0.4)	68.4(−1.3)	73.3(0.4)
15 **r+p+g** (21)	68.6(−0.7)	59.1(−6.5)	60.4(−7.6)
16 **d+r+p** (11)	70.7(−0.4)	67.4(−2.0)	73.8(0.1)
17 all (23)	70.4(−0.7)	69.0(−1.3)	74.2(−0.4)

added. The best detection rate for accent detection is obtained using dur and rms measurements only, as shown above in Fig. 2. Further addition of glottal measurements (13–15) show slight decrease in performance for accent detection, but produce the best results for IP+ip and IP boundary detection, using dur, rms, and glottal measurements. Finally, combining dur, rms, and pitch information results in better performance for only IP boundary detection, and using all 23 measurements provides somewhat suboptimal results compared to using only a selected subset of measurements.

The results of the table seem to indicate that duration and rms measurements (and to a lesser degree, pitch measurements) are most useful for accent detection, while IP +ip and IP boundary detection benefit from all measurements except pitch measurements.

Next, the errors for each detection task were further examined for each of the six broad prosodic categories, and the results are listed in Table IV. In accent detection, the highest

TABLE IV. Error analysis of six prosodic categories for three types of detection tasks. Each column lists detection error rates for neutral (0), accented (A), intermediate boundary (ip), accented intermediate boundary (Aip), intonational boundary (IP), and accented intonational boundary (AIP) tokens, respectively. The tasks are detection of accent, boundary (IP+ip), and IP boundaries.

Task	0	A	ip	Aip	IP	AIP
acc	27.5	31.2	44.4	18.2	49.3	9.9
IP+ip	26.0	36.5	56.2	34.7	27.2	10.7
IP	13.2	12.2	20.2	33.1	40.9	26.5

error rates are found for ip and IP syllables, and the lowest for AIP syllables. These results show that unaccented syllables at boundaries have a slight tendency to be mislabeled as accented. For IP+ip boundary detection, the highest error rates are for ip, A, and Aip syllables, i.e., intermediate boundary syllables and accented nonboundary syllables are more difficult to classify correctly as occurring at a boundary. The least error rate is found for accented intonational boundary syllables. For IP boundary detection, the largest error rate occurs for IP and Aip syllables—unaccented intonational boundary and accented intermediate boundary syllables were the most difficult to recognize correctly.

IV. SUMMARY AND DISCUSSION

In this paper, acoustic cues related to the voice source, including harmonic structure and spectral tilt, were examined in detecting prosodic events, in particular, intonational boundaries. The measurements comprise five categories: duration, pitch, harmonic structure, spectral tilt, and amplitude.

Distributions and statistical analysis of the measurements from the analysis data show that the following pause length increases at intonational boundaries, and speech rate decreases in the presence of both accents and boundaries. Pitch tends to fall for L- and L-L% boundaries and rise for H- and H-H% boundaries. These results are in agreement with previous studies on the relationship between durational and pitch measurements on prosody.[12–14] Pitch also tends to rise for accents, probably due to the dominant number of H^* type accents in this corpus. Further investigation for different types of accents will be needed to accurately correlate the effect of accent on pitch, since f0 cues for different types of boundary tones differ, and thus should be better modeled with a mixture, rather than a single distribution. Although the absolute and normalized f0 measures of this study both seem to be useful for finding accents, more detailed normalization techniques that take into account phone identity or average local f0 values may provide better results.

Two harmonic structure measurements, end $h1$-$h2$ and $h1$-$h2$ convexity, were useful for finding boundaries. Spectral tilt measurements end $h1$-$a1$, $h1$-$a1$ convexity, and end $a1$-$a3$ were also good indicators for boundaries. However, no harmonic structure measurement was significant in discriminating accented from nonaccented syllables, and of the spectral tilt measurements, only $a1$-$a3$ convexity was useful. In the analysis set for this study, the accented syllables at boundaries showed smaller values of $h1$-$h2$ and spectral measurements, which indicated more laryngealized or creaky voicing, than unaccented syllables. Pitch measurements were similar for accented and unaccented syllables at boundaries. However, compared to nonboundary unaccented syllables, nonboundary accented syllables showed greater $h1$-$h2$ and spectral tilt measurements (less creaky) but with much higher pitch values. This result seems to indicate that, in general, syllables at boundaries are more creaky than syllables that are not at boundaries, and the presence of an accent reinforces creakiness at boundaries, but the higher pitch associated with an accent for nonboundary syllables results in less creaky accented syllables, compared with unaccented nonboundary syllables. Finally, amplitude measurements were larger for accented syllables, and smaller for boundary syllables. Compared to duration, pitch, and amplitude measurements, harmonic and spectral tilt measurements exhibited small but significant effects.

Using the 23 acoustic measurements related to the voice source, detection experiments showed equal error rates around 70% detection for accent recognition. Finding both intermediate and intonational boundaries resulted in around 69% detection, and finding intonational boundaries yielded around 74% detection. Using subsets of the 23 voice source measurements shows that duration and amplitude measurements, and to a lesser degree, pitch measurements, are most useful for accent detection, while all measurements except pitch measurements are useful for finding intermediate and intonational boundaries. Analysis of the errors indicates that syllables at boundaries are more easily misclassified as accented. Also, compared to intonational boundary syllables, intermediate boundary syllables are more susceptible to being misclassified as nonboundary syllables.

In this paper, measurements for voice source characteristics were made over single syllables. Further studies will focus on extending the window of analysis to include two and three syllables, to examine the effect of longer-term changes in the source acoustic cues. Experiments with the Switchboard Telephone Speech Corpus[19] of spontaneous speech are also planned, to examine whether the results of this study will generalize to a more natural style of speech.

ACKNOWLEDGMENTS

The authors would like to thank Ken Chen for providing the pitch normalization algorithm, Taejin Yoon for help with statistical analysis, and Janet Slifka for valuable discussions. This work was supported by a grant from the University of Illinois Critical Research Initiative.

[1] V. Auberge, N. Audibert, and A. Rilliard, "Acoustic morphology of expressive speech: What about contours?" *Proceedings of Speech Prosody 2004*, Nara, Japan (2004), pp. 201–204; http://www.isca-speech.org/archive/sp2004/index.html

[2] Y. Xu, C. Xu, and X. Sun, "On the temporal domain of focus," *Proceedings of Speech Prosody 2004*, Nara, Japan (2004), pp. 81–84; http://www.isca-speech.org/archive/sp2004/index.html

[3] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," J. Acoust. Soc. Am. **90**, 2956–2970 (1991).

[4] D. Hirst, "The phonology and phonetics of speech prosody: Between acoustics and interpretation," *Proceedings of Speech Prosody 2004*, Nara, Japan (2004), 163–169; http://www.isca-speech.org/archive/sp2004/index.html

[5] M. Beckman and G. Ayers, "Guidelines for TOBI labeling (version 3.0)," The Ohio State University (1997).

[6] L. Larkey, "Reiterant speech: An acoustic and perceptual validation," J. Acoust. Soc. Am. **73**, 1337–1345 (1983).

[7] K. N. Stevens, *Acoustic Phonetics* (Massachusetts Institute of Technology, Cambridge, MA, 1998).

[8] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857 (1990).

[9] S. Chavarria, T. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the Switchboard corpus," *Proceedings of Speech Prosody 2004*, Nara, Japan (2004), pp. 333–336; http://www.isca-speech.org/archive/sp2004/index.html

[10]J. Slifka, "Respiratory constraints on speech production at prosodic boundaries," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.
[11]J. Pierrehumbert and D. Talkin, "Lenition of /h/ and glottal stop," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by G. Doherty and D. R. Ladd (Cambridge University Press, Cambridge, UK, 1992), pp. 90–119.
[12]C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717 (1992).
[13]A. E. Turk, and L. White, "Structural influences on accentual lengthening in English," J. Phonetics **27**, 171–206 (1990).
[14]J. Edwards, M. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," J. Acoust. Soc. Am. **89**, 369–382 (1991).
[15]M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio Speech Corpus," Linguistic Data Consortium (1995); http://www.ldc.upenn.edu/Catalog/LDC96S36.html
[16]Entropic Research Laboratory, Inc., *xwaves+manual* (1996).
[17]K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," *Proceedings of Speech Prosody 2004*, Nara, Japan (2004), pp. 509–512; http://www.isca-speech.org/archive/sp2004/index.html
[18]K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," *Proceedings of Int. Conf. on Spoken Lang. Proc. 1998*, Sydney, Australia (1998), pp. 3189–3192.
[19]J. J. Godfrey and E. Holliman. "The Switchboard-1 Telephone Speech Corpus Release 2," Linguistic Data Consortium (1997); http://www.ldc.upenn.edu/Catalog/LDC97S62.html