Analysis of prosodic classes using voice source measurements

Ran Han and Jeung-Yoon Choi.[1]

[1]School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
magnolia@dsp.yonsei.ac.kr, jychoi@dsp.yonsei.ac.kr

Abstract

In this study, we use voice source measurements over single-and multi-syllable regions to find acoustic cues for broad prosodic classes. Voice source measurements include duration, fundamental frequency, harmonic amplitudes, spectral tilt, and speech amplitude measurements. Using the Boston University Radio Speech Corpus, significant measurements are found from ANOVA tests, and the distributions of those measurements are observed. Finally, using all measurements, broad class prosody detection is carried out. The best detection rates for 4 broad accent classes and 5 broad boundary classes are 46.2% and 51.8%. From these results, it can be seen that voice source measurements and multi-syllable measurements may be useful in detecting detailed prosodic classes.

1. Introduction

Prosody is the rhythm, or intonation of speech and may reflect various characteristics of the speaker or the utterance, including emphasis, contrast, and focus.

To represent prosody, the ToBI[1, 2] system has been proposed. It transcribes the intonation and prosodic structure of spoken utterances in a variety of languages. Prosodic units are represented as accents and boundaries, with associated tones. In this study, we focus on the analysis and detection of the various types of accents and boundaries in utterances.

In previous works, detection for presence of accent or boundary are examined, mostly. Using acoustic, lexical, and syntactic evidences, prosodic event detection rates are 89.8% for accent and 93.3% for IP boundary[3, 4, 5, 6, 7]. A few are interested in analyzing for more detailed classes of accent and boundary[8, 9], lately. In this study, we focus on the analysis and detection of the various detailed types of accents and boundaries in utterances using only acoustic evidences.

In previous works using only acoustic cues, Choi *et al.*[10] examined detection methods for presence of accent and boundary over 1, 2, and 3 syllables were examined.

In this study, using the measurements previously suggested, we analyze these cues for broad classes of accent and boundary types. Also, the relative importance of various voice source measurements over 1 through 3 syllables are examined. In addition, we attempt to describe how these measurements are related to broad prosodic class detection compared with the presence-absence tests for accent and boundary.

2. Experiments

2.1. Prosodic units

The prosodic units considered here are based on the ToBI system, which describes various types of accent and boundary as indicators for prosodic events.

Accents are represented by a *. The - indicates an intermediate boundary, and the - % combination indicates a full intonational boundary. Each type of accent or boundary has an associated simple or complex tone. Simple tones include H, !H, and L, which represent high, downstepped high, and low tones, respectively. Complex tones for accents are denoted with a + symbol, such as in L+H*. Neutral syllables are denoted as 0.

The various prosodic notations can also be grouped into broad classes according to primary tone characteristics. The 4 broad accent types are H*, !H*, L*, and 0, while the 5 broad boundary types can be grouped as 0, L-, H-, L%, and H%. The 4 broad accent types are decided according to the presence of a dominant accent tone. Similarly, the 5 broad boundary types are decided according to the tones of boundaries. The vocalic segments with no boundary are one group, intermediate boundaries are divided into two groups (L- for low tones, H- for high and downstepped high tones), and intonational boundaries make up the other two groups, which are grouped according to the final tones of the intonational boundaries.

2.2. Acoustic measurements

The voice source acoustic cues that are examined here are divided into 5 categories: duration, fundamental frequency, harmonic structure, spectral tilt, and amplitude. All of these are calculated on vocalic segments only. Durational measurements include speech rate and length of a following pause. Speech rate is calculated as the reciprocal of the length of the vocalic segment, and 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. These values are from the work cited above.

Fundamental frequency measurements include end value of fundamental frequency(f0) and normalized fundamental frequency(nf0) at the end of vocalic segments, and their slopes and convexities over 1, 2, and 3 syllables (vocalic segments only). The slope and convexity are calculated as

$$\text{slope}_n = \frac{s(t_{1e}) - s(t_{ns})}{\sum\limits_{m=1}^{n} (t_{me} - t_{ms})},$$

where $s(t_{np})$ is the value of the measurements at time t_{np} at position $p \in \{s(\text{tart}), e(\text{nd})\}$ of the (n-1)th previous syllable,

$$\text{convexity}_n = \frac{\sum\limits_{m=1}^{n} \sum\limits_{t=t_{ms}}^{t_{me}} s(t) - h(t)}{\sum\limits_{m=1}^{n} (t_{me} - t_{ms})},$$

where $s(t)$ is the value of the measurements at time t, and $h(t)$ is the linear interpolated function,

$$h(t) = \frac{s(t_{me}) - s(t_{ms})}{t_{me} - t_{ms}} \cdot (t - t_{ms}),$$

Table 1: One-way ANOVA results for (a) 23 one-syllable measurements, and (b) 28 multi-syllable measurements. The values are the F-ratios for each measurement. The notation end denotes the values of each measurement at the end of vocalic segments and slp and cnv denote their slopes and convexities over 1 syllable. Also the notations slp# and cnv# denote the slopes and convexities of the measurements over 2 and 3 syllables.

classes / interval	broad acc		detailed acc		broad bnd		detailed bnd	
	meas	F-ratio	meas	F-ratio	meas	F-ratio	meas	F-ratio
1 syllable	f0 end	1238.9	pause	187.8	pause	359.7	f0 end	482.7
	rate	729.6	rate	159.1	f0 end	346.9	rate	275.9
	nf0 end	560.7	nf0 end	127.1	rate	314.8	nf0 end	218.9
	nf0 cnv	455.1	nf0 slp	72.1	nf0 end	233.7	f0 cnv	218.9
	nf0 cnv	383.3	f0 slp	52.1	f0 slp	102.7	nf0 cnv	155.6
2&3 syllables	f0 slp2	46.2	f0 slp3	18.4	f0 slp3	51.8	f0 slp2	22.2
	f0 slp3	689.1	f0 slp2	61.5	f0 slp2	118.2	f0 slp3	156.9
	nf0 slp2	311.1	nf0 slp2	40.9	nf0 slp2	77.8	nf0 slp2	117.4
	rms slp2	284.2	a1a3 slp3	27.3	nf0 slp3	64.2	rms slp2	107.9
	rms slp3	284.2	nf0 cnv2	23.1	a1a3 slp3	52.4	nf0 slp3	104.1

for $t_{ms} \leq t \leq t_{me}$, and $t_{ms} < t_{me}$, respectively.

Harmonic structure measurements include end value of h1-h2 at the end of vocalic segments and their slopes and convexities over 1, 2, and 3 syllables. h1 and h2 are the amplitudes of the first and second harmonics, respectively.

Similarly, spectral tilt measurements include end value of h1-a1, h1-a3, a1-a3 at the end of vocalic segments and their slopes and convexities over 1, 2, and 3 syllables. a1 and a3 are the amplitudes of the first and third formants, respectively.

Finally, amplitude measurements include end value of rms energy and their slopes and convexities over 1, 2, and 3 syllables.

The fundamental frequencies, harmonics, formants, and rms values were found automatically using the *formant* and *pitch* commands in *wavesurfer*. The vocalic segments were found using the phone labels in the *.lbl* or *.lba* files of the Boston University Radio Speech Corpus.

2.3. Database

The Boston University Radio Speech Corpus was used to analyze accents and boundaries in utterances. It is composed of radio news stories from 7 speakers, and is divided into 2 sections, labnews and radio. For this study, this corpus divided into 3 subsets - the analysis, training, and test sets. The analysis set includes 22 stories from speaker f1a and 41 stories from speaker f2b, and all of these contain ToBI labels (*.ton* files) and hand-corrected phone labels (*.lbl* files). The training set includes 36 files each from speaker m1b and f2b from the radio section, and the test set includes 24 files each from speaker m2b and f3a from the labnews section. The training and test sets have only automatically generated phone label files (*.lba* files) and ToBI labels (*.ton* files). The three database subsets are disjoint.

3. Results

In the previous work, using the 23 one-syllable measurements, the detection rates for accent, IP+ip and IP detection were reported as 71.1%, 70.3%, and 74.6%[10]. (IP+ip denotes detection of both intonational and intermediate boundaries; IP denotes detection of full intonational boundaries only.) After adding 28 multi-syllable measurements, best detection rates increase to 72.0%, 70.7%, and 75.7%[11], respectively.

In our experiments, baseline detection rates are comparable to or exceed the previous works' results - 71.4%, 81.2%, and 87.2% for the 23 one-syllable measurements, and 68.3%, 78.9%, and 83.7% for the 51 one and multi-syllable measurements.

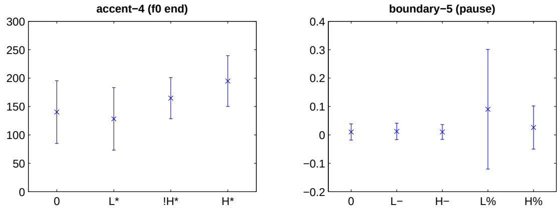

Figure 1: Most significant one-syllable measurements for broad accent and boundary classes. The 'x's are the means, and the bars denote the standard deviations. The distributions show, in order, non-accent (0), the L*, !H*, and H* accent classes, non-boundary (0), the L- and H-intermediate boundary classes, and the L% and H% intonational boundary classes.

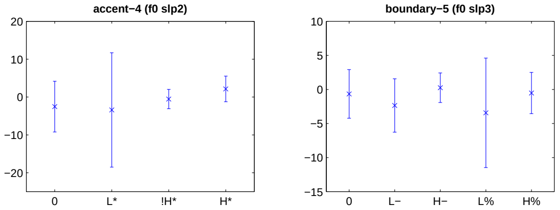

Figure 2: Most significant multi-syllable measurements for broad accent and boundary classes. The 'x's are the means, and the bars denote the standard deviations. The distributions show, in order, non-accent (0), the L*, !H*, and H* accent classes, non-boundary (0), the L- and H-intermediate boundary classes, and the L% and H% intonational boundary classes.

Table 2: Detection rates for 4 broad accent types and for 5 broad boundary types using 1 syllable measurements. The numbers in parentheses next to the subset names indicate the number of measurements included. In the subset section, dur, rms, pitch, harms, and tilt represent duration, rms, pitch, harmonic, and spectral tilt measurements. d, r, p, h, and t also represent those measurements, respectively. glottal or g represents the sum total of harmonic and spectral tilt measurements.

Subset	Accent	Boundary
dur (2)	25.2	30.3
rms (3)	30.0	24.5
pitch (6)	30.6	29.1
harms (3)	25.1	36.4
tilt (9)	25.8	26.3
glottal (12)	32.2	28.9
d+g (14)	33.8	30.4
r+g (15)	35.6	33.2
p+g (18)	37.9	32.8
d+r (5)	33.6	30.9
d+p (8)	42.8	29.3
r+p (9)	37.9	32.6
d+r+g (17)	37.6	34.0
d+p+g (20)	41.8	33.3
r+p+g (21)	40.6	36.5
d+r+p (11)	44.8	33.8
all (23)	43.3	35.7

Using these measurements, this paper focuses on the analysis and detection of broad prosodic class events, namely, the 4 broad types of accents and the 5 broad and 9 detailed types of boundaries.

3.1. Statistical analysis

The measurements for the training subset of the Radio News Corpus were examined using an analysis of variance (ANOVA). A one-way analysis with 4 of accents and 5 types of boundaries were carried out. The F values for measurements considered significant are shown in Tables 1 and 2.

From the results of the one-syllable measurements, pitch and durational measurements are more significant than others. Also, for the multi-syllable measurements, f0 slope over 2 or 3 syllables is the most significant measurement in each case. But in spectral tilt measurements, some are not significant enough in classifying the types of accents or boundaries. (Here, $P<0.001$ for one-syllable measurements, and $P<0.0008$ for multi-syllable measurements.)

3.2. Distributions

Fig. 1 and Fig. 2 show the distributions of 4 broad types of accents and 5 broad types of boundaries from the analysis set. One measurement that was shown to be the most significant in the ANOVA test for each case is shown. Overall, most of the pitch measurements tend to increase or have less negative values in the order of L, !H, and H tones, for both accent and boundary.

Table 3: Detection rates for 4 broad accent types and for 5 broad boundary types using multi-syllable measurements. The numbers in parentheses next to the subset names indicate the number of measurements included. In the subset section, pitch, harms, tilt, and rms represent pitch, harmonic, spectral tilt, and rms measurements. p, h, t, and r also represent those measurements, respectively.

Subset	Accent	Boundary
pitch (8)	19.5	8.1
harms (4)	21.2	20.0
tilt (12)	20.8	11.7
rms (4)	24.0	21.4
p+h (12)	20.8	12.0
p+t (20)	21.3	11.3
p+r (12)	21.9	10.9
h+t (16)	23.8	15.2
h+r (8)	26.9	25.7
t+r (16)	23.6	14.8
p+h+t (24)	22.2	13.1
p+h+r (16)	23.2	14.7
p+t+r (24)	23.0	13.3
h+t+r (20)	24.9	17.8
all(p+h+t+r) (28)	23.6	15.2
all-slopes only (14)	23.9	13.6
all-convs only (14)	46.2	51.8
2 sylls only (14)	22.5	17.2
3 sylls only (14)	23.8	12.3

For the 4 types of accents, end-values of f0 and slopes over 2 syllables are low or negative for L*, but these values increase, in order, for 0 (non-accented), !H*, and H*, so that H* has the highest positive values. For the 5 types of boundaries, pause length is longer for the boundaries, especially intonational boundaries, as might have been expected. The presence of a high tone leads to a less negative slope, and slopes over 3 syllables are more negative at boundaries. Intonational boundaries lead to a more negative slope than intermediate boundaries. In addition, the presence of a boundary leads to low end-values of f0 or more negative slopes over 2 syllables, and this effect is increased at intonational boundaries.

In conclusion, these distributions show that most values of measurements increase or become less negative in the order of L, 0, !H, and H tones, and the presence of a boundary leads to decreased slopes.

3.3. Detection results

All the 51 (23 one-syllable, plus 28 multi-syllable) voice source measurements were next used to detect the various classes of accents and boundaries. Using the training data set, means and covariance matrices for 51-dimensional Gaussian distributions for 4 broad types of accents and 5 broad types of boundaries were computed. For detection, a simple maximum likelihood measure was used.

The detection rates for various types of accents and boundaries using measurements over 1 and those after including 2-and 3-syllables are shown in Tables 2 and 5, respectively.

Using one-syllable measurements, duration, pitch and harmonic structure measurements give the best detection rates for

Table 4: Detection rates for 4 broad accent types and for 5 broad boundary types using 1, 2 and 3 syllable measurements. The numbers in parentheses next to the subset names indicate the number of measurements included. In the subset section, pitch, harms, tilt, and rms represent pitch, harmonic, spectral tilt, and rms measurements. p, h, t, and r also represent those measurements, respectively. d represents duration measurements.

Subset	Accent	Broad
pitch (14)	23.5	9.6
harms (7)	25.1	25.8
tilt (21)	23.1	19.9
rms (7)	27.3	25.1
d+p (16)	29.1	13.6
d+h (9)	34.6	34.2
d+t (23)	24.5	21.7
d+r (9)	34.3	29.6
d+p+h (23)	32.3	18.2
d+p+t (37)	29.8	19.4
d+p+r (23)	33.2	18.9
d+h+t (30)	29.6	24.6
d+h+r (16)	36.8	35.5
d+t+r (30)	29.9	25.8
d+p+h+t (44)	31.4	21.4
d+p+h+r (30)	36.3	23.4
d+p+t+r (44)	32.4	22.3
d+h+t+r (37)	32.9	28.2
all(d+p+h+t+r) (51)	34.3	24.2

each case. Combining these subsets resulted in best detection rates of 30.6%, and 36.4% for 4 broad types of accents and 5 broad types of boundaries, respectively. Overall, using glottal measurements improves detection rates for each case.

After including 2 or 3 syllable measurements, rms measurements show the best detection rates for all types. Combining all of these measurements subsets in various ways, the best detection rates for each type are 46.2% and 18.4% respectively. It is confirmed that the convexity measurements are useful in detecting among various types.

Finally, using 1-, 2-, and 3-syllable measurements, harmonic measurements, and rms measurements yield the best detection rates for each case. Combining all of these measurement subsets in various ways, the best detection rates for each type are 36.8% and 31.1% respectively. All of these results are from detection using duration, harmonic and rms measurements, which show these measurements are useful in detecting prosodic events.

Overall, we can see from Table 5 that 46.2% from using all convexity measurements only over 2 and 3 syllables gives the best detection rate for the 4 broad classes of accents. For 5 broad types of boundaries, 51.8% from using all convexity measurements only over 2 and 3 syllables gives the best detection rate. Measurements over 2 and 3 syllables are more useful in detection, and they include voice source measurements related to harmonic structure or spectral tilt measurements.

4. Summary

In this study, analysis of acoustic cues for the broad prosody classes are investigated. The measurements that are considered comprise 5 categories: duration, fundamental frequency, harmonic structure, spectral tilt, and amplitude.

From ANOVA tests, durational and fundamental frequency measurements are found to be significant, and distributions of those measurements are presented. Of these, most pitch measurements tend to increase or have less negative values in the order of L, !H, H tones, for both accents and boundaries.

The best detection rate for 4 broad types of accents and 5 broad types of boundaries are 46.2% and 51.8% respectively. From the results, 2 and 3 syllable measurements are more helpful in improving detection rates. Also, voice source measurements that include harmonic and spectral tilt measurements help to improve detection rates. In general, it was found that using voice source measurements and multi-syllable measurements are useful for broad class prosody detection as well as for the presence-absence tests for accents and boundaries.

Table 5: The best detection rate of each case.

Interval	Accent		Boundary	
	meas	rate	meas	rate
1 syllable	d+r+p	44.8	r+p+g	36.5
2,3 syllables	all-cnv only	46.2	all-cnv only	51.8
1,2,3 syllables	d+h+r	36.8	d+h+r	35.5

5. References

[1] M.E. Beckman, J. Hirschberg, "The TOBI Annotation Conventions", The Ohio State University, 1994.

[2] M.E. Beckman, G.A. Elam, "Guidelines for ToBI labelling (version 3.0)", The Ohio State University, 1997.

[3] C.W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns", IEEE Transactions on Audio and Speech Processing, vol.2, pp. 469-481, 1994.

[4] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model", Proc. of ICASSP, USA, pp. 509-512, 2004.

[5] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, ' 'Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework", IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 797-811, 2008.

[6] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical and syntactic evidence", IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 216-228, 2008.

[7] J-H Jeon and Yang Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features", Proc. of ICASSP, Taiwan, pp. 4565-4568, 2009.

[8] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis", Computer Speech and Language, vol. 10, pp. 155-185, 1996.

[9] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric F0 features", Proc. of ICASSP, USA, pp. 4545-4548, 2008.

[10] J-Y Choi, M. Hasegawa-Johnson, and J. Cole, "Finding intonational boundaries using acoustic cues related to the voice source", J. Acous. Soc. Am. 118, 2579-2587, 2005.

[11] J-Y Choi, "Voice quality contour measurements for prosody detection", 157th Meeting of the Acoustical Society of America, Portland, Oregon, 18-22 May, 2009.