Pitch Accent versus Lexical Stress: Quantifying Acoustic Measures Related to the Voice Source

Yen-Liang Shue[1], *Markus Iseli*[1], *Nanette Veilleux*[2], *Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles
{yshue,iseli,alwan}@ee.ucla.edu
[2]Department of Computer Science, Simmons College
nanette.veilleux@simmons.edu

10.21437/Interspeech.2007-690

Abstract

In this paper, we explore acoustic correlates of pitch accent and main lexical stress in American English, and the interaction of these cues with other factors that affect prosody. In a controlled study, we varied presence or absence and type of pitch accent (L^* vs H^*), boundary-related tone sequence (L-$L\%$ vs. H-$H\%$) and gender of the talker, for the sentence "Dagada gave Bobby doodads". The measures were duration, F_0 (fundamental frequency), $H_1^*-H_2^*$ (related to open quotient), and $H_1^*-A_3^*$ (related to spectral tilt). Contour approximations were used to analyze time-course movements of these measures. For "Dagada" we found that, consistent with earlier literature, a) H^* and L^* pitch accents showed different F_0 contours, b) pitch-accented syllables were longer than unaccented ones, c) stressed "ga" syllables had lower $H_1^* - H_2^*$ values than surrounding unstressed syllables, and for male talkers, lower $H_1^* - A_3^*$ values, indicating lesser spectral tilt. Unexpectedly, F_0 maxima associated with an H^* accent occurred most of the time later in the accented syllable than F_0 minima associated with L^*. The cues to lexical stress were consistent with or without pitch accent (e.g. lower $H_1^* - H_2^*$), but they sometimes interacted with gender and/or boundary tones: for example, lower $H_1^* - A_3^*$ in stressed "ga" syllables was only found for female talkers in unaccented cases, and some cues of both accent and stress were less pronounced in the final word "doodads", which also carried boundary-related tones.

Index Terms: voice source, prosody, voice quality

1. Introduction

Prosody describes properties of speech such as rhythm, timing, intonation, and stress. In American English, an important part of prosody relates to the prominence of a word within a phrase. This is usually marked by a pitch accent. Pitch accent, as a prosodic feature, allows a speaker to place contrastive stress on words within a phrase to indicate prominence or significance. Similarly, lexical stress allows a syllable to be more prominent than others within a word. Boundary tones signify groupings and allow a speaker to group words into intonational phrases and the choice of boundary tone can distinguish statements (Low or L-$L\%$) from questions (High or H-$H\%$). Accurate detection of pitch accents, stress, and boundary tones would benefit applications such as automatic speech recognition, speaker identification, and emotion classification.

With a few exceptions, previous studies of prosodic features have typically focused on the fundamental frequency (F_0), intensity, and duration. In [1], a large number of voice source related measures was analyzed using the Boston University Radio Corpus and it was found that there were no spectral harmonic measurements which could distinguish between accented and non-accented syllables. Similarly, [2] found that correlates of pitch accents were: differences in peak fundamental frequency (F_0), peak intensity, and amplitude of voicing. In [3], which studied Dutch speakers, and [4], which studied Swedish sentences, it was found that stressed syllables are generally tenser, have more high frequency energy and lower open quotient of the glottal source. Since pitch-accented syllables are also stressed, it would be expected that these attributes might also apply to pitch-accented syllables. In [5], it was found that these results were statistically significant if a distinction was made between low and high pitch accents. However, in that study, stressed syllables were compared with all other unstressed syllables in the corpus. When the effects of boundary-related tones were taken into account in later analysis, it was found that the results were only significant if the speakers were separated by gender.

In this paper, using a prosodically-labeled corpus, which is carefully constructed to have the same words in different prosodic contexts, we examine how acoustic measures of lexical stress are affected by the presence of pitch accent, gender of the talker, and boundary tones. Acoustic measures are estimated and contours are fitted to these measures based on a weighted least squares error criterion. Analysis of variance (ANOVA) is performed to assess the statistical significance of the results.

2. Materials and Methods

2.1. Corpus and Subjects

The corpus consists of data from [6] along with new recordings of the same sentences so that the total number of speakers is 10: 5 males and 5 females. For each speaker, 10 repetitions were recorded for each of the following sentences, where the bold word is accented:

- **Dagada** gave Bobby doodads.
- Dagada gave Bobby **doodads**.
- **Dagada** gave Bobby doodads?
- Dagada gave Bobby **doodads**?

The declarative and interrogative sentences induce the subjects to place contrasting boundary tones on the same word for the different sentence types.

Subjects were native speakers of Western American English between 21-35 years old. Signals were recorded in a sound-attenuated booth with a 1.0" Bruel & Kjaer condenser microphone placed 5 cm from the subjects' lips. The signals were sampled at 20 kHz and downsampled to 10 kHz. The first and

last repetitions of each sentence were discarded for the final analysis.

Two graduate students manually segmented the sentences and used the ToBI [7] transcription standard to label the corpus. For this study, the high and low pitch accents, denoted by H^* and L^*, and the high and low boundary-related tones, denoted by H-$H\%$ and L-$L\%$, on the words "Dagada" and "doodads" were analyzed. Syllables with primary lexical stress as on "ga" in "Dagada" and on "doo" in "doodads" are underlined. For the analysis of "Dagada", 32 files from a male speaker who pronounced the word as "Dagada" were discarded, while for the "doodads" 10 files were discarded as the F_0 tracker did not provide reliable data. The final distribution of prosodic labels was 69/97/122 ($L^*/H^*/noPA$) occurrences for "Dagada" and 81/82/75/72 (L^*H-$H\%/H^*L$-$L\%/L$-$L\%/H$-$H\%$) occurrences for "doodads". Note that $noPA$ indicates no pitch accent and that the labels for "doo" can be L^*/H^* or none, while for "dads" they are either L-$L\%$ or H-$H\%$.

2.2. Voice Source Measures

Three measures related to the voice source were estimated: F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. Asterisks denote that the corresponding spectral magnitudes have been corrected for the effects of the vocal tract [8]. These measures were estimated over the entire duration of each sentence at a time resolution of 1 ms.

F_0 was estimated using the STRAIGHT algorithm [9]. The formant frequency and bandwidth inputs to the vocal tract correction formula [8] were estimated using the Snack Sound Toolkit [10] with the following settings: pre-emphasis factor of 0.96, window length of 25 ms, and window shift of 1 ms.

$H_1^* - H_2^*$ is the corrected difference between the first and second spectral harmonic magnitudes and has been shown to be related to open quotient [11]. The harmonic magnitudes H_1 and H_2 were estimated from the signal spectrum using the F_0 information from the STRAIGHT algorithm. Corrections [8] were then applied to the harmonic magnitudes to compensate for the effects of the first two formant frequencies (F_1 and F_2). $H_1^* - A_3^*$ is the spectral magnitude difference between the first harmonic and the magnitude of the spectrum at the third formant frequency (F_3); this measure is a correlate of spectral tilt [11, 12]. A_3^* was estimated using F_3 values from Snack and corrected for the effects of F_1, F_2, and F_3.

2.3. Contour Fitting and Analysis

For each word, contours were fitted to the three voice source measures according to a weighted least squares error criterion based on the signal energy, $E(n)$. When the energy falls below a certain threshold, as would occur in-between syllables of a word, the voice source measures become less reliable, and hence, less weighting is applied to the error function. The error weighting function, $W(n)$, was determined by $E(n)$, with the threshold, E_{th}, at a quarter of the mean energy of the word. After $E(n)$ drops below the threshold, the weighting function decreases exponentially, as shown in Eq. 1.

$$W(n) = \begin{cases} 1, & E(n) \geq E_{th} \\ e^{-(E_{th}-E(n))/E_{th}}, & E(n) < E_{th} \end{cases} \quad (1)$$

The use of this error weighting function ensures that only the most reliable parts of the voice source measures are used for the contour fitting. Although raw values are not continuous between syllables in a word, silence duration is usually small compared to syllable duration. Using contour approximation allows general trends to be captured.

Similar to what was done in [13], weighted Legendre polynomials were used for the contour approximations due to their orthogonality property. Each Legendre polynomial, $P_i(n)$ is associated with a coefficient, a_i, which enables a data vector, $y(n)$, to be approximated as $y(n) \approx \sum_{i=0}^{N} a_i P_i(n)$, where N is the desired polynomial order. The coefficients a_i provide a simple way to approximate a data vector. For this study, we set $N = 3$ since the longest word in the test corpus consists of three syllables. Eq. 2 shows the error criterion, E_a, used in the optimization of the a_i's.

$$E_a = \sum_n \Big(y(n) - \sum_{i=0}^{3} a_i P_i(n) \Big)^2 \cdot W(n) \quad (2)$$

The orthogonal property of Legendre polynomials enables each coefficient to be optimized separately. For simplicity, we used iterations of the intermediate value theorem to find the optimal a_i's. Iterations were stopped when the a_i values did not change within five decimal places. The four coefficients (a_0, a_1, a_2 and a_3) used in this study represent, respectively, the Legendre polynomials $P_0(x) = 1$ (related to the mean), $P_1(x) = x$ (related to linear slope), $P_2(x) = \frac{1}{2}(3x^2 - 1)$ (related to quadratic convexity/concavity), and $P_3(x) = \frac{1}{2}(5x^3 - 3x)$ (related to cubic behavior).

For each word, contours were fitted to the three voice source measures (F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$) and the results were manually checked for all utterances; 29 F_0 contours at the beginning and the end of the utterances had to be corrected. For each prosodic event (H^*, L^*, H^*L-$L\%$, L^*H-$H\%$, H-$H\%$ and L-$L\%$), the means of the coefficients were calculated, enabling a direct comparison between the effects of each prosodic event. Two-way ANOVA tests, from the software package SPSS (v13.0) were then performed on the coefficients, with the fixed factors being speaker and prosodic feature. The p (probability of null hypothesis) values, F (ratio of the model mean square to the error mean square) values, and partial η^2 (measure of effect size) values are reported for some cases.

3. Results

3.1. Pitch Accent

For the word "Dagada", as expected, most talkers showed higher/lower F_0 values for H^*/L^* pitch accented syllables compared to the $noPA$ case. Fig. 1 shows F_0 contours averaged over data from the male talkers for the unaccented and accented pronunciations of the word. Interestingly, for H^*, most talkers showed a minimum value close to the end of the first syllable ("Da") and a maximum value at the beginning of the last syllable ("da"), where the F_0 maximum was about 15 Hz higher for H^* compared to $noPA$. That is, the F_0 maximum did not occur during the stressed "ga" syllable but was delayed to the beginning of the next syllable. The F_0 drop before the actual maximum indicates that these cases should perhaps be labeled with L+H^*, instead of H^*, although this distinction was sometimes difficult to make perceptually. Here, we consider both L+H^* and H^* to be of the same category. For the L^* case, both genders showed an F_0 minimum at the middle of the stressed "ga" syllable, where it was about 15 Hz lower for L^* compared to $noPA$. For 7 out of 9 talkers the delay between F_0 maximum for H^* and F_0 minimum for L^* was about 100 ms. For one female talker, there was no delay, and for another female talker, the delay was 200 ms. The delay may be

due to the dip in F_0 before the H^*, which provides more contrast for the following high pitch accent. ANOVA results on the effects of $noPA$, H^*, and L^* were significant for all speakers and all four polynomial coefficients.

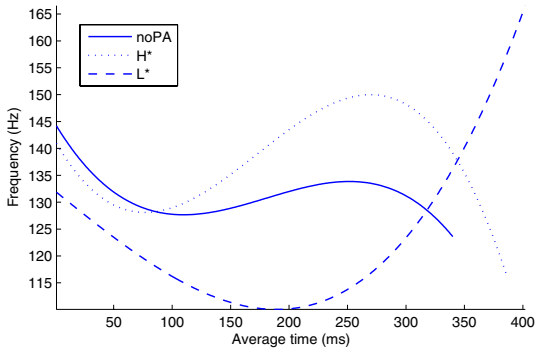

Figure 1: Average stylized F_0 contours "Dagada" (males).

Both genders also exhibited similar F_0 contours for the boundary word "doodads". Fig. 2 shows F_0 contours for female talkers for each of the four prosodic events (L-$L\%$, H-$H\%$, H^*L-$L\%$, and L^*H-$H\%$). With few exceptions, the F_0 contour for H-$H\%$ increased monotonically ($a_1 > 0$), whereas for L-$L\%$ it decreased ($a_1 < 0$). For all talkers the contour for L-$L\%$ always lay below the contour for H-$H\%$ and the contours for accented words (L^*H-$H\%$ and H^*L-$L\%$) lay mostly between the contours for L-$L\%$ and H-$H\%$. The delayed F_0 peak for the H^* case which was observed for "Dagada" was not as pronounced for "doodads", with only a slight delay observed for some talkers. This could be due to the influence of the boundary tone and/or due to the stress structure of the word. We are in the process of collecting and analyzing data from the same speakers but with the words "Dagada" and "doodads" switched in their position to address this question. Interestingly, most speakers showed a slightly lower/higher F_0 before a high/low tone, respectively. This has also been observed in Mandarin [14]. ANOVA analysis on the prosodic events showed that all four coefficients were statistically significant for both male and female speakers. As expected, both words show female F_0 contours with larger values and range than males (M/F: 110-155 Hz/190-260 Hz).

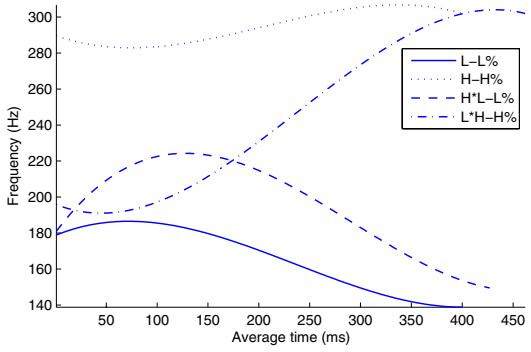

Figure 2: Average stylized F_0 contours for "doodads" (females).

The duration of the word "Dagada" in accented cases was always longer compared to the unaccented cases and was confirmed with ANOVA analysis ($p/F/\eta^2$ = 0.00/139.7/0.52), which tested the significance of the durational change in "ga" with accentedness as a factor. The same trend was also found for "doodads", but with a smaller effect size (ANOVA: $p/F/\eta^2$ = 0.00/30.8/0.09). A similar result was reported in [15]. $H_1^* - H_2^*$ and $H_1^* - A_3^*$ were not found to be distinctive for pitch accent.

3.2. Lexical Stress

In "Dagada", $H_1^* - H_2^*$ values seem to correlate well with lexical stress regardless of pitch accent. All talkers showed similar convex ($a_2 > 0$) contour shapes with a minimum during the stressed syllable "ga". Fig. 3 shows these contours for male talkers for each of the three prosodic events ($noPA$, H^*, L^*). For all talkers and independent of accentedness, $H_1^* - H_2^*$ was larger at the onset and the offset of the word than on the middle, stressed syllable "ga" indicating a smaller open quotient and tenser voice quality for the stressed syllables. An ANOVA test on the raw $H_1^* - H_2^*$ mean values against the fixed factors speaker and syllable position within the word was significant with $p/F/\eta^2$ = 0.00/68.17/0.15. On average, the stressed syllable "ga" was about 2.5 dB and 4 dB lower than the surrounding syllables for males and females, respectively. As expected, "Dagada" $H_1^* - H_2^*$ contours showed higher mean values (M/F: 2.5 dB/4.9 dB) and a larger range (M/F: 0.5-5.5 dB/2-9 dB) for females when compared to male speakers [16].

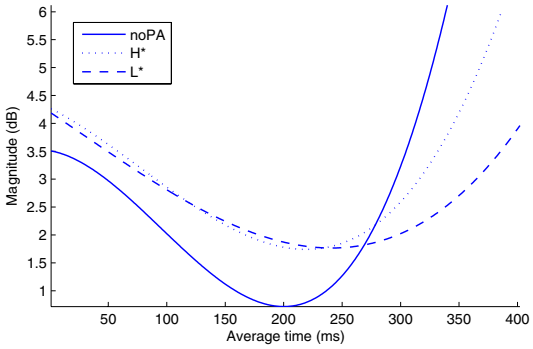

Figure 3: Average stylized $H_1^* - H_2^*$ contours for "Dagada" (males).

As for "doodads", the results for $H_1^* - H_2^*$ contours seem to be speaker and gender dependent. On average, $H_1^* - H_2^*$ contours for L-$L\%$ lay above those of H-$H\%$ in female speech but the opposite was true for male speech. Contour minima and maxima could be found anywhere within the word and it was difficult to associate their locations with stress. This lack of consistency could be due to boundary tone effects.

The $H_1^* - A_3^*$ contours appear to be gender dependent. For "Dagada", the average contours for both genders exhibited a parabolic shape. With the exception of one talker, male speech showed convex curves ($a_2 > 0$) for all three prosodic cases. For 3 out of the 5 female talkers, the opposite ($a_2 < 0$) was true for the accented cases. For almost all talkers the minima/maxima values occurred during the stressed "ga" syllable with male speakers showing a minimum for lexical stress regardless of pitch accent. Fig. 4 shows these contours for one male subject. The figure also shows segment boundaries for the accented and unaccented cases. This indicates a more abrupt

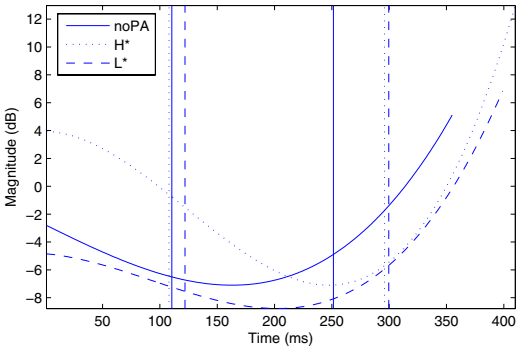

Figure 4: Stylized $H_1^* - A_3^*$ contours for "Dagada" for a male talker showing syllable boundaries for an instance of each prosodic case.

closure of the vocal folds on stressed syllables and agrees with [2], [3], [4], and [5]. As indicated earlier, for some female talkers, unaccented cases also had minima in "ga" but maximum values were observed when the stressed syllable was accented.

More consistency was found for the $H_1^* - A_3^*$ contours for "doodads" which, on average, had concave parabolic shapes. With the exception of two female talkers, the contours showed a low value of $H_1^* - A_3^*$ which increased to a maximum around mid-word and then decreased at the end of the word (end of the utterance). This result again suggests that stressed syllables have lower spectral tilt (more high frequency energy) and agrees with previous work. Compared to declarative sentences (L-$L\%$ and H^*L-$L\%$), interrogative sentences (H-$H\%$ and L^*H-$H\%$) had, on average, a lower $H_1^* - A_3^*$ contour on the phrase-final syllable "doo"; a similar observation was made in [5].

4. Summary

Not surprisingly, pitch accents were clearly marked by differences in F_0 contours. For "Dagada", averaged contours revealed that for both genders, the L^* event caused the F_0 minima to occur at the middle of the accented syllable, while for the H^* case, F_0 maxima appear towards the end of the accented syllable. This delayed peak, which was observed for almost all speakers for "Dagada" but not for "doodads", has implications for analyses which use mid-syllable values. For all speakers, the syllable and hence, word duration was longer for the accented cases than for non-accented cases.

For "Dagada", lexical stress was clearly marked by the convex shape of the $H_1^* - H_2^*$ contours which indicate a tenser voice (lower open quotient) on the stressed syllable; this measure seemed to be independent of pitch accent. However, this trend was not found for "doodads" possibly due to the influence of boundary tones. The spectral tilt measure ($H_1^* - A_3^*$) was seen to be gender dependent for "Dagada", with the contour decreasing for the stressed syllable for male speech, while for female speech, this was true only for the unaccented case. For "doodads", the boundary-related tone, especially $H - H\%$, generally caused the $H_1^* - A_3^*$ contours to decrease towards the end of the word, denoting lower spectral tilt or an increase in high-frequency energy.

These results suggest that acoustic cues of lexical stress can be affected by the presence of a pitch accent, boundary tone, and in some cases, gender of the talker. In the future, we will further explore the interaction between acoustic measures related to prosodic events as well as examine intra-speaker variations.

5. Acknowledgements

We thank Patricia Keating and her students for the database and for many helpful discussions and Stefanie Shattuck-Hufnagel for her insightful comments and inspiration. Work supported in part by the NSF and a Radcliffe Fellowship to Abeer Alwan.

6. References

[1] J-Y.Choi, M. Hasegawa-Johnson, and J. Cole, "Finding intonational boundaries using acoustic cues related to the voice source," *JASA*, vol. 118, no. 4, pp. 2579–2587, 2005.

[2] A. Okobi, "Acoustic Correlates of Word Stress in American English," Dissertation, MIT, 2006.

[3] A. Sluijter and V. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *JASA.*, vol. 100, no. 4, pp. 2471–2485, 1996.

[4] G. Fant, "The voice source in connected speech," *Speech Communication*, pp. 125–139, 1997.

[5] M. Iseli, Y.-L. Shue, M. Epstein, P. Keating, J. Kreiman, and A. Alwan, "Voice source correlates of prosodic features in american english: A pilot study," in *Proc. of ICSLP*, Pittsburgh, PA, September 2006, pp. 2226–2229.

[6] M. Epstein, "Voice Quality and Prosody in English," Dissertation, University of California, Los Angeles, 2002.

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, and J. Pierrehumbert, "ToBI: a standard for labeling english prosody," in *Proc. ICSLP*, vol. 2, Banff, Alberta, Canada, Oct. 1992, pp. 867–870.

[8] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 669–672.

[9] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *ICSLP Proc.*, 1998.

[10] K. Sjölander, "Snack sound toolkit," KTH Stockholm, Sweden, 2004, http://www.speech.kth.se/snack/.

[11] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *JSHR*, vol. 38, pp. 1212–1223, 1995.

[12] H. M. Hanson, "Glottal characteristics of female speakers," Dissertation, Harvard U., Cambridge, MA, 1995.

[13] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *JASA*, vol. 118, no. 2, pp. 1038–1054, 2005.

[14] Y. Xu, "Contextual tonal variation in mandarin chinese," *J. Phonetics*, vol. 25, pp. 61–83, 1997.

[15] A. Turk and L. White, "Structural influences on accentual lengthening in english," *J. Phonetics*, vol. 27, pp. 171–206, 1999.

[16] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *JASA*, vol. 106, pp. 1064–1077, 1999.