Voice Source Correlates of Prosodic Features in American English: A Pilot Study

*Markus Iseli, *Yen-Liang Shue, **Melissa A. Epstein
**Patricia Keating, ***Jody Kreiman, *Abeer Alwan

*Department of Electrical Engineering
{iseli,yshue,alwan}@ee.ucla.edu
**Department of Linguistics
maepstein@alumni.upenn.edu, keating@humnet.ucla.edu
***Department of Head and Neck Surgery
jkreiman@ucla.edu
University of California, Los Angeles
405 Hilgard Ave., Los Angeles, CA, 90095

10.21437/Interspeech.2006-318

Abstract

In this paper, we examine the dependencies of voice source parameters F_0(fundamental frequency), E_e(maximal glottal flow change), RK(glottal symmetry/skew), LIN(value related to source spectral tilt) and $H_1^* - H_2^*$(difference of formant-corrected magnitudes of the first two source spectral harmonics) on prosodic features such as pitch accents, stress, and sentence type and the interdependencies of some of these measures. A small, carefully designed corpus containing a sentence in different prosodic configurations was used in this study. Statistical analysis was performed using two-way ANOVAs to test for the voice source parameter dependencies. Results show that F_0 is positively correlated with E_e and LIN, and negatively correlated with $H_1^* - H_2^*$. Stressed syllables showed lower values of RK and $H_1^* - H_2^*$ compared to stressless syllables. The effect of pitch accent can be seen as a combination of its F_0, and stress. Phrase-final syllables for interrogative sentences yielded a higher F_0 and lower RK and $H_1^* - H_2^*$ compared to declarative sentences. It was found that it is important to differentiate between tones when analyzing prosodic features that involve tones, such as pitch accent and probably boundary.

Index Terms: voice source, prosody, voice quality.

1. Introduction

In connected speech, prosody serves both as a grouping function and a prominence-marking function. The groupings of, for example, words into phrases are indicated by prosodic *boundaries*. The prominence of a word within a phrase is marked in English by particular F0 patterns, called *pitch accents*; for example, a pitch accent can signal a focal accent, for contrastive stress on a word. Likewise, in English words one syllable is more prominent than the others, because English is a language with *lexical stress*. These aspects of prosody convey important information for understanding connected speech on word, phrase, and content levels. Most previous studies of speech prosody have focused on F0, duration, and intensity as acoustic correlates. Only a few studies have analyzed voice source parameters in connected speech, yet speech processing applications would benefit from knowledge of voice source parameter dependencies on prosodic features.

A framework for studying voice source parameters in connected speech was provided in [1], in the context of the Liljencrants-Fant, or LF, model [2]. The framework was evaluated for several Swedish sentences. Results show the importance of several factors that affect source parameters, including speaker- and segment-specific effects, coarticulation and interpolation at boundaries, fundamental frequency (F_0), stress, pitch accent and voice intensity, and phrasal contour effects. For example, contrastive stress boosts overall intensity as well as the high frequency balance, but for non-focused lexical stress these measures are relatively unimportant. In [3, 4] it was shown that the LF source parameters vary systematically as a function of both stress and pitch accent in Swedish. In [5] it was shown that for Dutch speakers, spectral balance, duration, overall intensity, and vowel quality all varied with lexical stress (with and without pitch accent), but especially, stressed syllables were generally longer and had higher spectral balance. Spectral balance here refers to the relative spectral energy above 500 Hz compared to the total energy, and is related to the speed of glottal closure.

Recent publications [6, 7] have used the ToBI framework, which provides labels for the following prosodic events: pitch accent, boundary tone, and break indices. In [6], normalized LF model parameters were shown to vary with the presence of accents and boundary tones in a small set of short read sentences. Epstein suggested that, at least in English, prosodic strengthening is seen in voice measures in much the same way as elsewhere in speech (e.g. [8]). She found tenser voice, utterance-initially and with pitch accent, suggesting greater laryngeal tension in prosodically strong positions.

In [7], a number of measurements related to the voice source (duration, F_0, harmonic structure, spectral tilt, and amplitude) were made for a relatively large database of American English (the Boston University Radio Corpus). It was reported that duration and amplitude were useful for detecting pitch accents, while voice source measurements were useful for boundary detection. Interestingly, the time course of these measurements (and not their static values) served as good indicators for prosodic events.

Work in part supported by the NSF.

In this paper a statistical analysis of variance (ANOVA) of the five voice source parameters F_0, E_e (LF parameter for value of maximal glottal flow change), RK (glottal symmetry/skew derived from LF model parameters), LIN (related to source spectral tilt), and $H_1^* - H_2^*$ (difference of formant-corrected magnitudes of the first two source spectral harmonics) is performed for several sentences. The ANOVA tests for the dependencies of these parameters against the independent factors: speaker, sentence type (SNT), the presence vs. absence of pitch accent (PA), and stress (STR).

2. Materials and Methods

2.1. Speech Data

The test corpus [6] consists of the following eight-syllable sentences, where the bold word is accented and has narrow focus:

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

These sentences are designed to contain no nasals and to have all vowels surrounded by voiced consonants. By using the same string of words with different pitch accent locations, and different pitch accent and boundary tones (for questions vs. statements), it is possible to directly compare the effects of these prosodic variables on voice source parameters by standard factorial analysis of variance.

Speech signals were recorded from 3 native speakers of Western American English (1 male, 2 females), between 25-35 years old. Signals were collected in a sound booth with a 1.0" Bruel & Kjaer condenser microphone placed 5 cm from the subjects' lips. The signals were sampled at 20 kHz and then downsampled to 10 kHz. Each sentence was recorded 10 times for each speaker and the first and last recordings were then discarded for the final analysis.

The corpus was prosodically labeled so that comparisons can be made across different prominent positions, prominent and non-prominent words and different pitch accent and boundary tones. Table 1 shows the distribution of pitch accents over all syllables, where each pitch accent is denoted by L and H indicating low and high pitch (F_0), respectively. The labeling system was closely based on the ToBI [9] transcription standard. A more detailed description of the data collection procedure and corpus labeling can be found in [6].

2.2. Analysis

Here and throughout the paper, H_1 and H_n refer to the magnitude of the first and n-th spectral harmonic, respectively. Thus for example $H_1 - H_2$ is the difference (in dB) between the fundamental and second spectral harmonic magnitudes. Formant frequencies and bandwidths of the n-th formant are written as F_n and B_n, respectively.

Our algorithms estimate the five voice source parameters F_0, E_e, RK, LIN, and $H_1^* - H_2^*$. The asterisks denote that the corresponding spectral magnitudes (H_1, H_2) are corrected for the effect of the first and second formants [10].

F_0, E_e, and RK are directly related to the parameters in the LF model [2] as shown in Fig. 1. E_e relates to the spectral intensity and is measured as the amplitude of the negative peak of the differentiated glottal pulse. This value is equivalent to the amplitude at the point of maximum discontinuity in the glottal waveform for RK values up to 0.54. $RK = \frac{T_e - T_p}{T_p}$ is the ratio of the closing phase to the opening phase of the pulse and is related to the glottal symmetry. The parameter LIN measures the slope of the source spectrum which is correlated with overall spectral tilt. These 4 parameters were estimated from explicit inverse filtering and LF-fitting by using the signal analysis tool developed at UCLA's Bureau of Glottal Affairs [11]. To prepare the data for this tool, a cycle was taken from the steady state portion of the vowel in each syllable in each word of the corpus. The cycles were then concatenated with themselves 10 times in order to produce a long enough signal for inverse filtering.

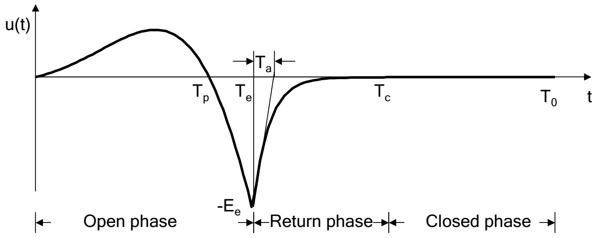

Figure 1: Parameters in the LF-model. $u(t)$ is the differentiated glottal pulse. E_e is the maximum negative amplitude and T_0 is the period. T_p, T_e, T_a, and T_c are the point of zero closure velocity, the point of the maximum negative amplitude, the effective duration of the return phase and the point of the closing phase respectively.

The parameter $H_1^* - H_2^*$ is related to the open quotient [12]. The amplitudes of the harmonics were estimated from the signal spectrum by using F_0 information provided by the STRAIGHT algorithm [13]. The correction formula [10] then performs an implicit inverse filtering to remove the effects of the first two formant frequencies. Input to the correction formula requires F_1 and F_2 and their bandwidths B_1 and B_2 which were estimated using the "Snack Sound Toolkit" software [14] with a pre-emphasis factor of 0.9, analysis window length of 25 ms and window shift of 1 ms. The short window shift time was selected to be compatible with the F_0 values from STRAIGHT.

In total, there were 768 syllables in the corpus, however 68 were deemed by the inverse filtering program to be non-LF-fittable and discarded. For each syllable, the five voice source parameters were estimated. Only the raw measures were analyzed and no normalization was done as in [6]. The syllables were classified into categories depending on the speaker, and prosodic features such as pitch accent (PA: the stressed syllables of the words in bold are compared with all others), stress (STR: the syllables "ga", "gave", "Bob", and "doo" are compared with all others), and sentence type (SNT: the syllables in the declarative pair of sentences are compared with the syllables in the interrogative pair). The distributions for each of these categories are shown in Table 1. Statistical analysis was performed using the two-way ANOVA test in the software package SPSS (v13.0). Factors for each of the two-way analyses consist of Speaker plus one other factor chosen from the prosodic features. Speaker effects were generally significant but will not be reported here.

Table 1: *Distribution of prosodic features with respect to the number of syllables.*

Pitch accent (PA)			Stress (STR)	
H	L	None	Yes	No
84	45	547	365	335

Sentence type (SNT)	
Declarative	Interrogative
331	369

3. Results

The following tables show statistically significant dependencies of voice source parameters on prosodic features. Tests where the null hypothesis has a probability of $p < 0.05$ are statistically significant. Statistically insignificant results ($p > 0.05$) are marked with a "−". The prosodic features are pitch accent (PA, low vs. no vs. high), stress (STR, unstressed vs. stressed), and sentence type (SNT, declarative vs. interrogative).

PA: Significant pitch accent dependencies of the source parameters are shown in Table 2. Pitch accents at boundaries were excluded for this analysis to remove the effects of the sentence type. Pitch accented syllables are always stressed and since they usually show significant changes in the pitch contour, the analysis is divided into low (L), no (0), and high (H) pitch accent.

Table 2: *Significant dependencies of voice source parameters on pitch accent (**PA**) comparing low(L) tone pitch accent vs. no(0) pitch accent vs. high(H) tone pitch accent. The probability of the null hypothesis (p) and the means are shown. $H_1 - H_2$ is included for comparison only.*

PA: $L \leftrightarrow 0 \leftrightarrow H$	F_0(Hz)	E_e
p means	0.000 143 ⤢ 179 ⤢ 195	0.000 1.79 ⤢ 1.90 ⤢ 2.09
$L \leftrightarrow 0 \leftrightarrow H$	RK	LIN
p means	0.000 .418 ⤡ .375 ⤡ .276	0.000 .812 ⤢ .830 ⤢ .867
$L \leftrightarrow 0 \leftrightarrow H$	$H_1^* - H_2^*$(dB)	$H_1 - H_2$(dB)
p means	0.027 5.00 ⤡ 4.92 ⤡ 4.45	0.225 0.98 ⤡ 0.81 ⤢ 1.66

A strong dependency on pitch accent of F_0, E_e, RK, LIN, and to a lesser extent, $H_1^* - H_2^*$, can be seen. The results show the importance of dividing the analysis of PA into L, 0, and H pitch accents, since the voice source parameter values drop or rise depending on the tone transition. Therefore when comparing accented vs. unaccented syllables it has to be specified if the accented syllable is L or H. The change of F_0 values with L or H is obvious.

In [6] it is stated that prominent and phrase initial syllables display a tenser voice quality than their non-prominent and phrase-final. The 106 prominent syllables studied in [6] were a subset of the 153 pitch-accented syllables studied here. Citation from [6]: "Both prominent words and phrase-initial words displayed a tenser voice quality than their non-prominent and phrase-final counterparts. A tense voice quality is associated in theory with greater compression of the vocal fold and greater force of closure of the arytenoids. Acoustically, tense voice quality is correlated with low values of open quotient and glottal skew, and high values of spectral intensity and spectral linearity". However, our findings show that the source parameter values behave completely opposite depending on tone.

In [7], $H_1 - H_2$ (not corrected for formant influences) is reported to increase for pitch accented syllables, which would contradict our findings. We could somewhat reproduce this result by looking at the uncorrected value $H_1 - H_2$ (see Table 2), but the result is non-significant and only holds true when comparing L with H tones.

STR: Significant stress dependencies of the source parameters are shown in Table 3. Syllables at boundaries were excluded from this analysis. Since stressed syllables are a subset of pitch accented syllables, some of the results are expected to be similar.

Table 3: *Significant dependencies of voice source parameters on stress (**STR**) comparing unstressed(no) vs. stressed(yes) syllables.*

STR: $no \leftrightarrow yes$	RK	$H_1^* - H_2^*$(dB)
p means	0.050 .375 ⤡ .352	0.013 5.28 ⤡ 4.39

Table 3 shows that only RK and $H_1^* - H_2^*$ are significantly dependent on stress. The slight decrease in RK for stressed syllables signifies a slightly more skewed glottal pulse shape which is characterized by more high frequency components. This result is in line with the increase of spectral balance for Dutch stressed syllables described in [5].

SNT: Significant sentence-type dependencies of the source parameters are shown in Table 4. Since the largest pronunciation difference is expected to be on the phrase-final syllable (IP boundary), only this final syllable, which was stressless (i.e. "dads" in "doodads") was analyzed.

Table 4: *Significant dependencies of voice source parameters on sentence-type (**SNT**) comparing declarative(dec) vs. interrogative(int) sentence type.*

SNT: $dec \leftrightarrow int$	F_0(Hz)	E_e	RK
p means	0.000 143 ⤢ 229	0.006 1.50 ⤢ 1.73	0.000 .473 ⤡ .346
$dec \leftrightarrow int$	$H_1^* - H_2^*$(dB)	LIN	
p means	0.000 5.73 ⤡ −.70	0.000 .759 ⤢ .852	

It is clear that F_0 values at the phrase-final syllable increase for interrogative sentences. Table 4 is very similar to Table 2. This result could be explained with Fig. 2.

Interdependencies: The interdependency of F_0 was checked against the other voice source parameters. For this purpose F_0 was split into low (L) and high (H) values, using the average F_0 for each speaker as the splitting point. Only unstressed syllables were analyzed. The results are shown in Table 5. The result for $H_1^* - H_2^*$, which is negatively correlated with F_0, contradicts the findings of [15] which states that there is a "weak positive correlation between OQ and F_0" in Dutch. The OQ (open quotient) has been found to be positively correlated with the $H_1^* - H_2^*$ measure[12].

Table 5: *Significant dependencies of voice source parameters on fundamental frequency (**F0**) comparing low(L) vs. high(H) F_0. The threshold for the L vs. H decision was the average F_0 value for each speaker.*

F0: $L \leftrightarrow H$	E_e	$H_1^* - H_2^*$(dB)	LIN
p	0.005	0.000	.007
means	1.84 ⤢ 2.12	6.52 ⤡ 3.81	$.819$ ⤢ $.854$

A speculative explanation of the results seen so far (see Fig. 2) is: F_0 is interdependent with E_e, LIN, and $H_1^* - H_2^*$. Stress is correlated with RK and $H_1^* - H_2^*$. Results for pitch accent dependency seem to be a combination of the results for F_0 and stress dependency.

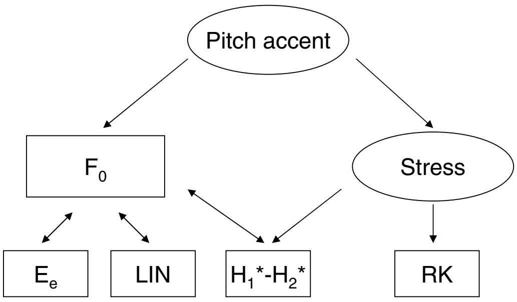

Figure 2: Speculative interpretation of interdependencies of voice source parameters and prosodic features. Voice source parameters are depicted in rectangular boxes and prosodic features are shown in ovals. The arrows indicate the direction of dependencies.

Finally, Table 6 shows a summary of the presented results.

Table 6: *Summary of the statistical significant dependencies ($p < 0.05$) of the analyzed voice source parameters (leftmost column) on prosodic features and F_0 (top row). Statistically insignificant results are marked with a "$-$".*

	PA	STR	F0	SNT
	$L \leftrightarrow 0 \leftrightarrow H$	$no \leftrightarrow yes$	$L \leftrightarrow H$	$dec \leftrightarrow int$
F_0	⤢ ⤢	$-$	⤢	⤢
E_e	⤢ ⤢	$-$	⤢	⤢
RK	⤡ ⤡	⤡	$-$	⤡
$H_1^* - H_2^*$	⤡ ⤡	⤡	⤡	⤡
LIN	⤢ ⤢	$-$	⤢	⤢

4. Conclusions

Statistical analysis of the significant interdependence of voice source parameters is summarized in Fig. 2 and Table 6. For our data set F_0 was positively correlated with E_e and with LIN, and negatively correlated with $H_1^* - H_2^*$. Stressed syllables showed lower values of RK and $H_1^* - H_2^*$ compared to stressless syllables. The effect of pitch accent can be seen as a combination of its F_0, and stress. Phrase-final syllables for interrogative sentences yielded a higher F_0 and lower RK and $H_1^* - H_2^*$ compared to declarative sentences. It was found that it is important to differentiate between tones when analyzing prosodic features that involve tones, such as pitch accent and probably boundary.

A reliable pitch accent and stress detection, based on source parameter estimation, could be helpful for emotion classification, speech recognition, speaker identification, and medical applications. Future work will build on the results presented here to develop such a detector.

5. References

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

[2] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Trans. Lab. Q. Prog. Stat. Report 4*, pp. 1–13, 1985.

[3] G. Fant and A. Kruckenberg, "Notes on stress and word accent in swedish," *Speech Trans. Lab. Q. Prog. Stat. Report 2–3*, pp. 125–144, 1994.

[4] ——, "Voice source properties of speech code," *TMH-QPSR. Report 4*, pp. 45–56, 1996.

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

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

[7] J-Y.Choi, M. Hasegawa-Johnson, and J. Cole, "Finding intonational boundaries using acoustic cues related to the voice source," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2579–2587, 2005.

[8] P. Keating (in press), "Phonetic encoding of prosodic structure," to appear in J. Harrington and M. Tabain (eds) *Speech Production: Models, Phonetic Processes and Techniques*, Psychology Press: New York, New York.

[9] 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.

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

[11] The Bureau of Glottal Affairs, "Inverse Filter Software," UCLA, Available as open source shareware at http://www.surgery.medsch.ucla.edu/glottalaffairs/software.htm.

[12] 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," *J. Speech Hear. Res.*, vol. 38, pp. 1212–1223, 1995.

[13] H. Kawahara, A. de Cheveign, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *Proceedings ICSLP'98*, Sydney, Australia, December 1998.

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

[15] J. Koreman, "Decoding linguistic information in the glottal airflow," Ph.D Thesis, University of Nijmegen, 1996.