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

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

[*]Department of Electrical Engineering, UCLA
[**]Department of Linguistics, UCLA
[***]Department of Head and Neck Surgery, UCLA

Work supported in part by the NSF

Goal

- To investigate how certain acoustic measures related to the voice source (F_0, H_1^*-H_2^*, *LIN*, *RK*, and E_e) correlate with prosodic events.

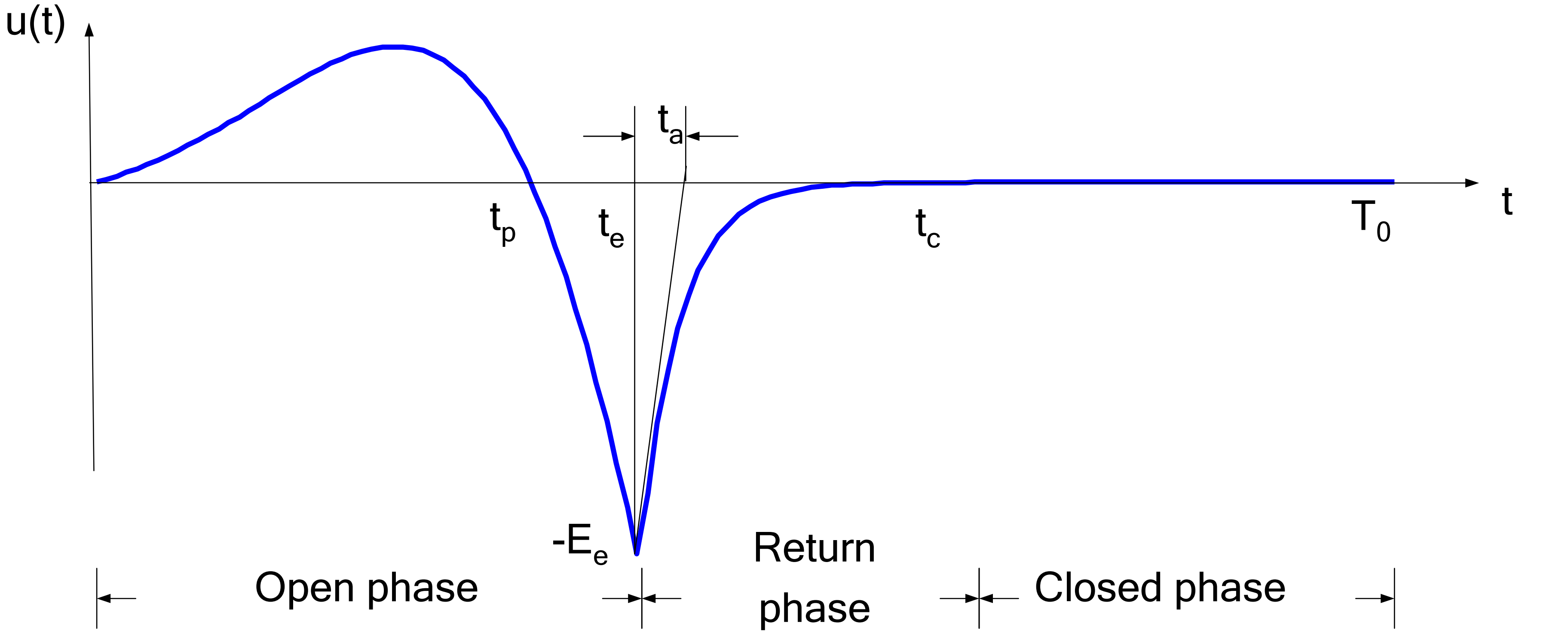
Motivation

- Prosodic events are conveyed in part by the voice source.
- Few studies have analyzed voice source parameters in connected speech (e.g. Fant & Kruckenberg 1994, Sluijter & Van Heuven 1996, Epstein 2002, Kochanski et al. 2005, Choi et al. 2005).
- Speech processing applications would benefit from knowledge of voice source parameter dependencies on prosody.

Introduction: *Prosody*

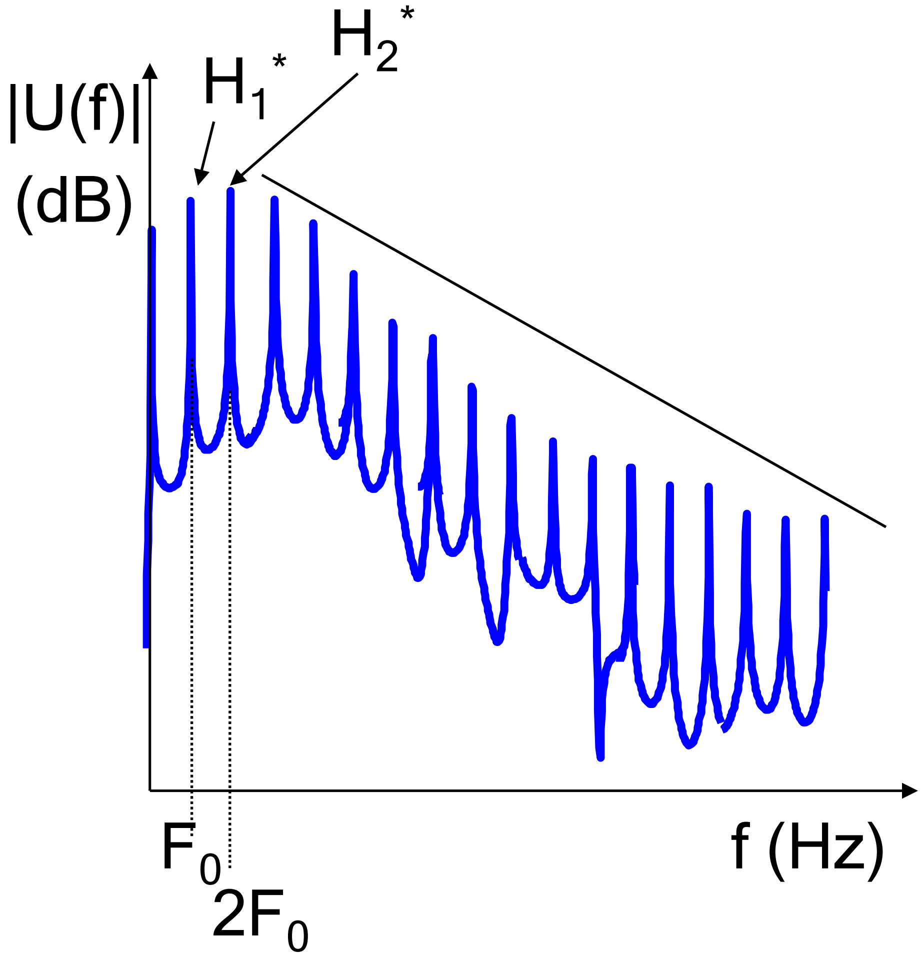
- Prosody broadly refers to intonation, phrasing, timing, and lexical stress in speech.
- *Lexical stress* allows for a particular syllable in a word to be more prominent.
- *Pitch accents* signify prominence of a word within a phrase. Here, both low (L^*) and high (H^*) pitch accents are studied.
- *Boundaries* indicate breaks between groups of words.

Acoustic measures: *LF model measures*



- $F_0 = 1/T_0$
- E_e is proportional to intensity
- ***RK*** = $(t_e - t_p)/t_e$ is related to glottal skew (inversely related to high frequency energy)

Acoustic measures (cont'd)



- H_1^*-H_2^* is related to open quotient (Holmberg 1995)
- ***LIN*** is proportional to high-frequency energy

Materials: *The corpus*

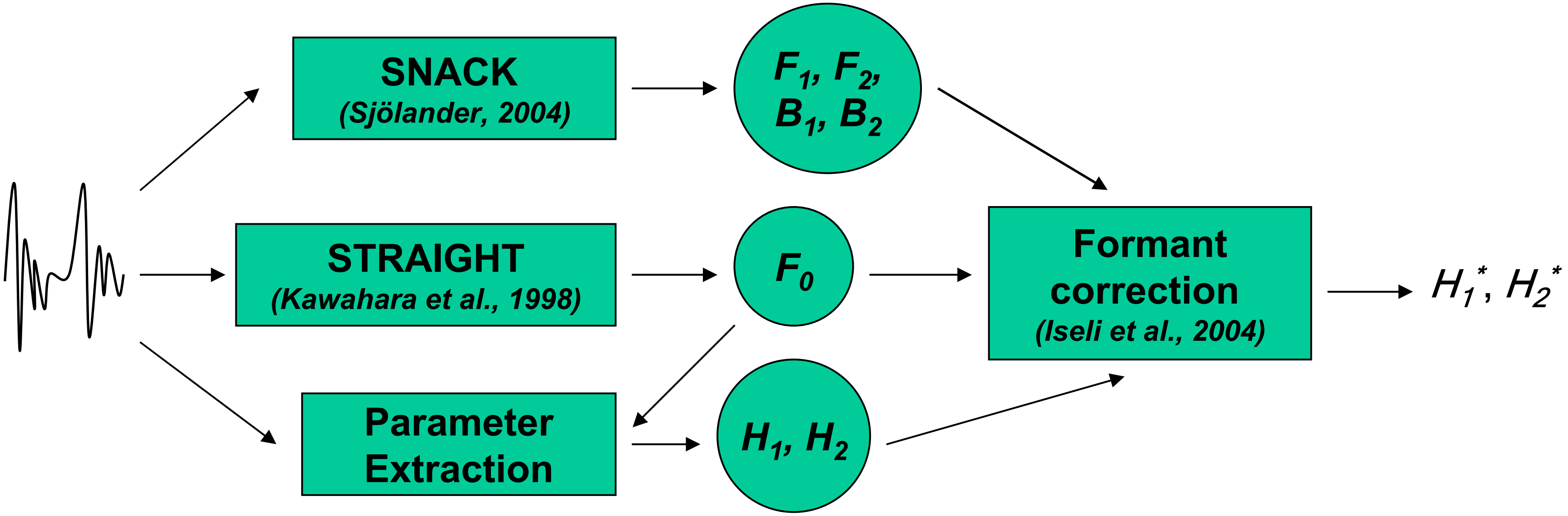
- The corpus (Epstein, 2002) consists of the following eight-syllable sentences which were ToBI labeled:
  - **Dagada** gave Bobby doodads.
  - Dagada gave Bobby **doodads**.
  - **Dagada** gave Bobby doodads?
  - Dagada gave Bobby **doodads**?
- Bold words are focused: *pitch accent* (PA) factor.
- Two sentences are declarative and two are interrogative: *sentence type/boundary* (BOUND) factor.
- Stressed vs. unstressed syllables are studied to examine the *lexical stress* (STR) factor.

Speakers and Material

- Speakers: 3 adult (25-35 years old) native speakers of American English: 2 females (B and S) and 1 male (L)
- Signals collected in a sound booth with a 1.0” B & K condenser microphone, and sampled at 20 kHz (later downsampled to 10 kHz)
- Each sentence was recorded 10 times for each speaker; the first and last recordings were discarded in the analysis.
- Total number of syllables analyzed: 700

Method: *Estimation of source-related measures*

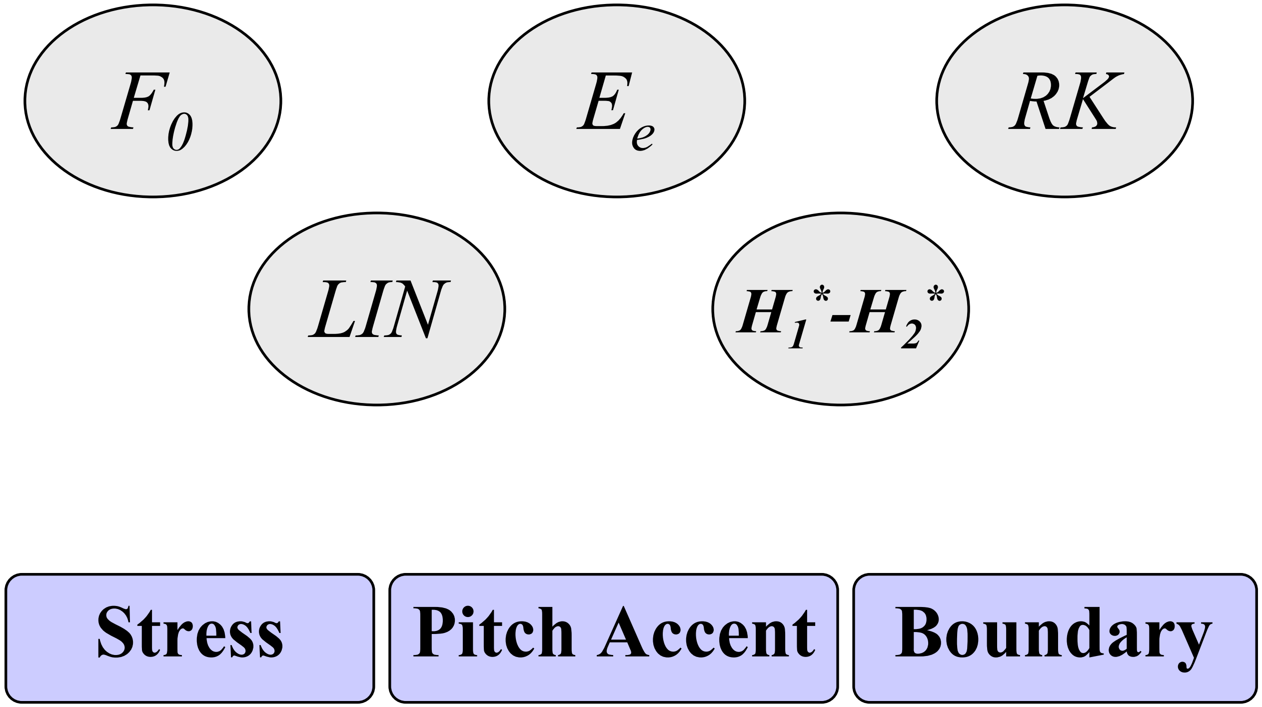
- F_0, E_e, *RK*, and *LIN* estimated by inverse filtering and LF-fitting. Measures are taken over one cycle.
- H_1^*-H_2^* obtained as follows:



Inter- and intra-correlations

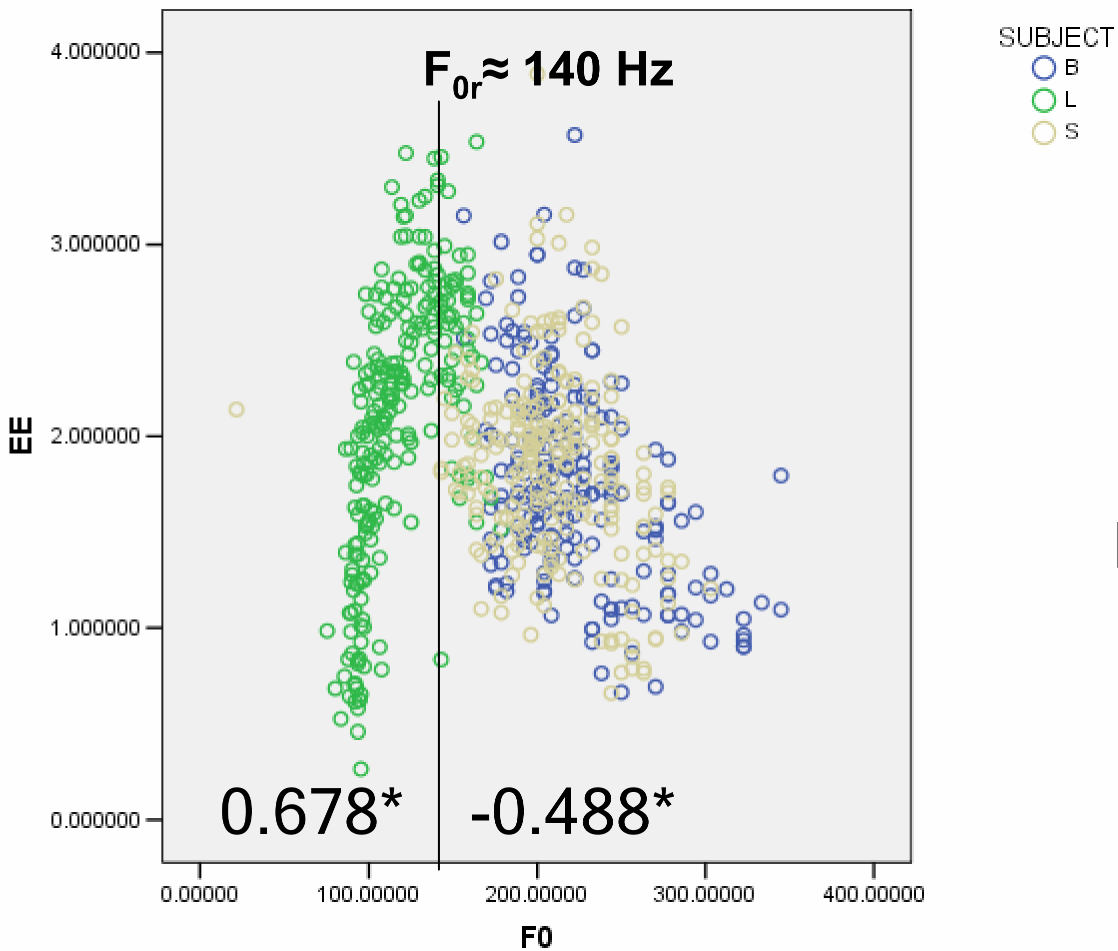
Acoustic features*

Prosodic features:



*all measures are z-score normalized for each utterance

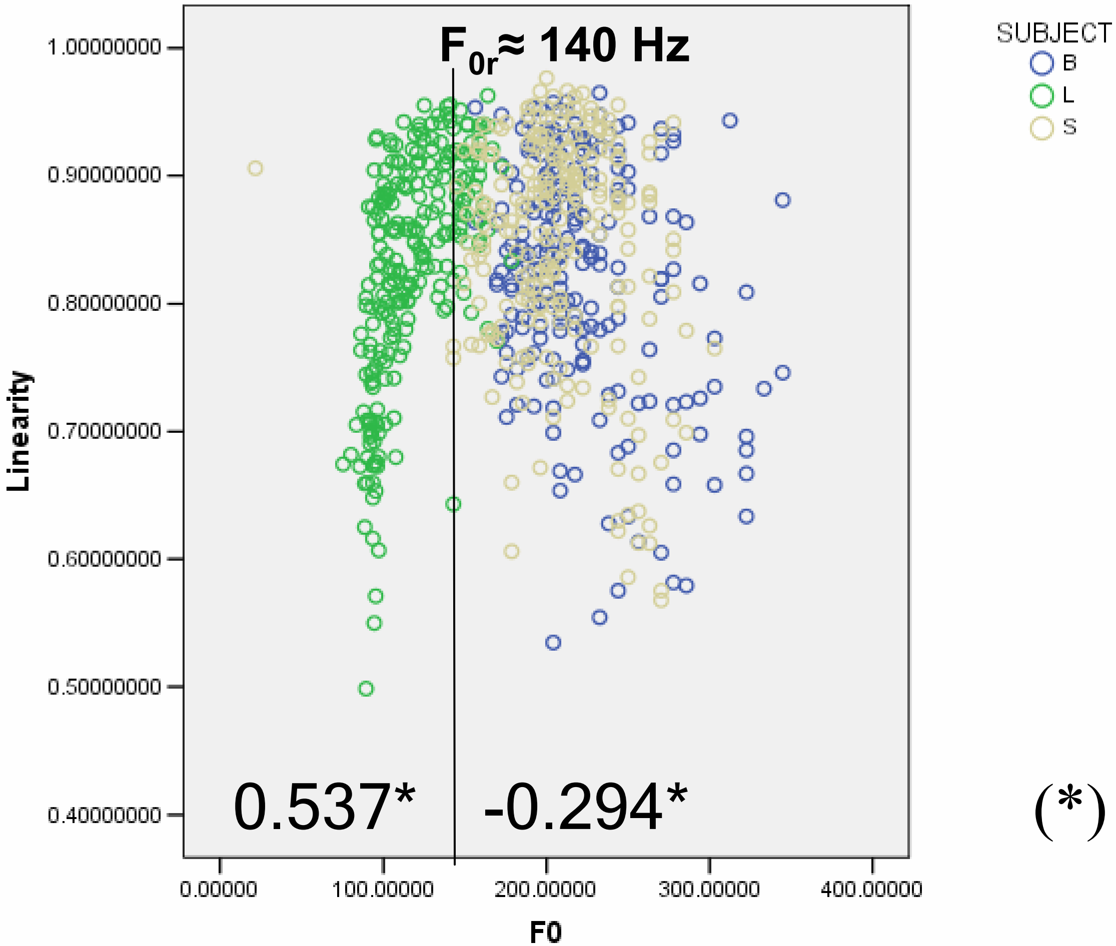
Results: Correlation between E_e and F_0



Compare to midfrequency F_{0r} presented in Fant et al. (1996)

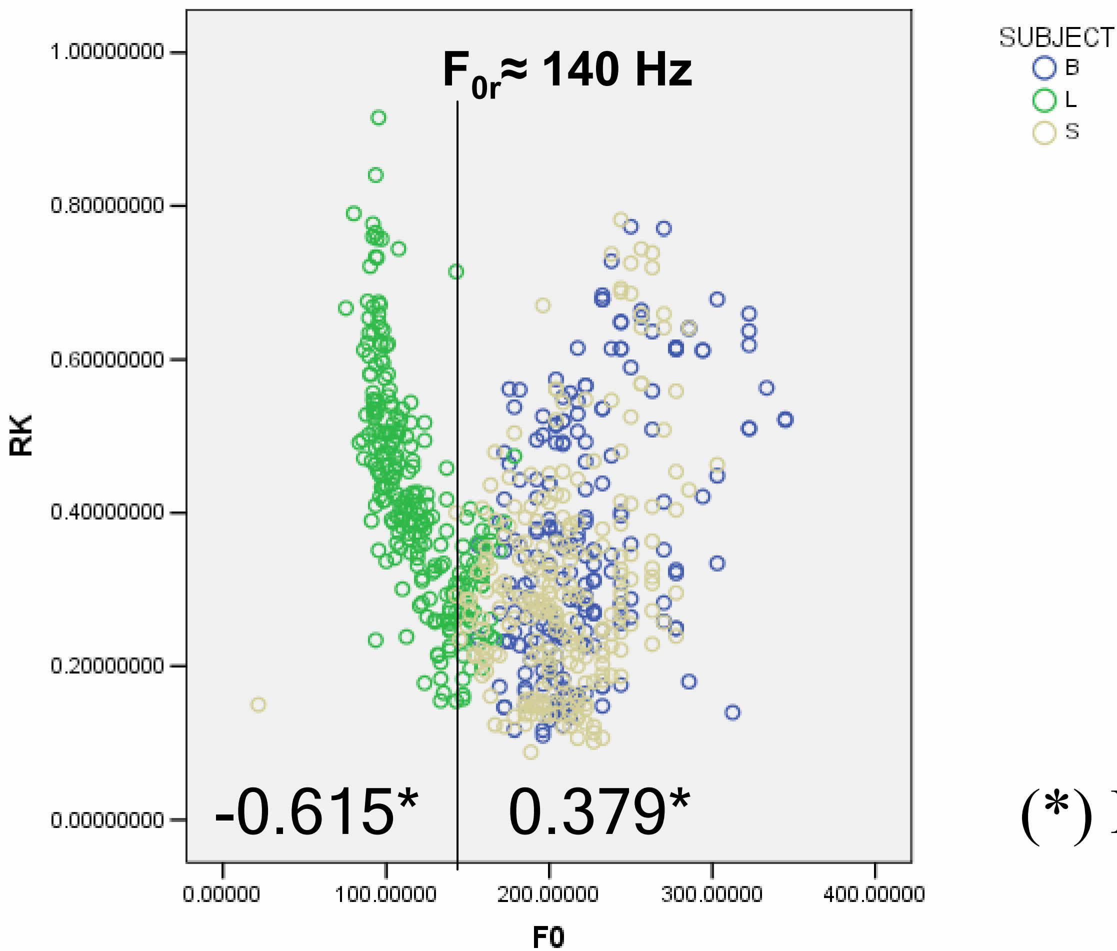
(*) Pearson's Correlation Coefficient (r)

Results: Correlation between *LIN* and F_0



(*) Pearson's *r*

Results: Correlation between *RK* and F_0



(*) Pearson's *r*

Other statistically-significant *intra-correlations*

- For all F_0:
  - E_e is positively correlated with *LIN* (r = 0.708)
  - *RK* is negatively correlated with *LIN* (r = -0.711)
  - *RK* is negatively correlated with E_e (r = -0.593)

Results: *Intercorrelations*

	F_0	E_e	*LIN*	*RK*	H_1^*-H_2^*
STR *no ↔ yes*	⤡ (BOTH)			⤢ (MALE) ⤡ (FEMALES)	
PA *no ↔ yes*			⤢ (BOTH)	⤡ (BOTH)	
PA *L* ↔ H**	⤢ (BOTH)	⤢ (BOTH)	⤢ (BOTH)	⤡ (BOTH)	
BOUND *dec ↔ int*	⤢ (BOTH)		⤢ (BOTH)	⤡ (MALE) ⤢⤡ (FEMALES)	⤡ (BOTH)

- Color code: MALE, FEMALES, BOTH
- Correlations shown are statistically significant at $p < .01$

Differences from our published Interspeech'06 paper

In the published paper, measures were not z-score normalized and we did not separate the results of female versus male speakers.

As a result of the normalization, H_1^*-H_2^* is no longer a correlate of stress nor of pitch accent and E_e is no longer a correlate of sentence type. Instead, F_0 is shown to be a correlate of lexical stress.

In addition, there was a gender (or perhaps F_0) related dependency for *RK* relative to stress and sentence type.

Summary and Conclusions

For our data set:

- *Lexical Stress* – results in lower F_0 and in lower/higher *RK* for the male/female talkers.
- *Pitch accent* –It is important to distinguish between low and high tones. For all talkers, F_0, intensity, and high-frequency energy (as measured by *LIN* and *RK*) are higher for H^* compared to L^*.
- *Boundaries* – interrogative sentences have higher F_0 and *LIN*, and lower open quotient (as measured by H_1^*-H_2^*) than declarative sentences. *RK* was speaker specific.

Comparison with other work

- Choi et al, 2005: H_1-H_2 and spectral tilt measures not useful for identifying *accents*. Amplitude is larger for *accented* syllables. We agree that H_1^*-H_2^* measures are not correlated with stress nor pitch accent, and that E_e is correlated with pitch accent. However, we find that spectral tilt and glottal skew are correlated with pitch accent (they didn't distinguish between L^* and H^*).

Comparison with other work (cont'd)

- Sluijter & Van Heuven, 1996: *Stressed* syllables have more high frequency energy, and *accented* syllables have higher intensity. Here, only the female speakers showed smaller glottal skew for stressed syllables. Moreover, E_e is higher for H^* when compared to L^*.
- Fant & Kruckenberg, 1996: In Swedish, F_0 is a *stress* correlate. F_0, intensity, and high-frequency emphasis, are correlated with *pitch accent*. Here, we also find that F_0 is a correlate for stress, and in addition, female speech shows high-frequency emphasis. For pitch accent, when distinguishing between H^* and L^*, we find similar results.

Summary and Conclusions (cont'd)

- The absolute value of F_0 affects how E_e, *LIN*, and *RK* are correlated with F_0.
- Among the five parameters studied, *RK* was the most speaker dependent.

In the future, we will examine whether these results generalize to a larger database.

Thank you