ACOUSTICS OF WHISPERED BOUNDARY TONES: EFFECTS OF VOWEL TYPE AND TONAL CROWDING

Willemijn Heeren & Vincent J. van Heuven

Leiden University Centre for Linguistics; Leiden Institute for Brain and Cognition, Leiden University, the Netherlands
w.f.l.heeren@hum.leidenuniv.nl; v.j.j.p.van.heuven@hum.leidenuniv.nl

ABSTRACT

The acoustic realization of boundary tones in whispered speech was investigated. This was done in four different vowels, and in two structures: with or without lexical stress and boundary tone coinciding. The analyses showed a number of cues, both secondary and compensatory ones, that were not fully comparable across vowel contexts, and more clearly present without tonal crowding.

Keywords: whispered speech, speech production, intonation, boundary tones

1. INTRODUCTION

In whispered speech –where voicing is absent– listeners can still perceive differences in intonation, albeit less reliably than in normal speech. For instance, in whisper listeners recognize questions and statements expressed by different boundary tones (H% versus L%), when prosody –rather than syntax– codes the crucial information [4, 5]. In whisper, listeners can also, amongst other things, discriminate intended 'pitch' height [7], differentiate emotional from neutral speech [15], and identify lexical tones, e.g., [1, 12], information that is normally thought to be largely carried by pitch.

The question which acoustic correlates may carry the information associated with intonation in whispered speech, has received relatively little attention, and has mainly been studied at the level of syllables (to eliminate context effects), rather than at the level of multiword phases or sentences, which might be considered more ecologically valid. Also, as some of these studies were done several decades ago, most evidence is qualitative rather than quantitative. In addition to knowing what acoustic correlates may carry prosodic information in whisper, we are interested in the nature of these correlates: are they secondary or compensatory?

If pitch perception in whisper is coded by *secondary* cues, assuming that speech is a redundant signal, these would be cues that are already present in normal, phonated speech. Early support for this hypothesis is found in the perception of 'vocoder whisper', i.e. vocoded normal speech with the periodic excitation signal replaced by a noise source in the resynthesis, in which lexical tones remained identifiable [1,3]. Acoustic studies have found some evidence that intensity [3, 5] and the first formant (F1) might be secondary cues [5]. For /a/, a combination of F1 and F2 upward shifts were found to correlate with intended pitch height in whisper [7], though no comparison with pitch height in normal speech was made. In a follow-up study, listeners discriminated intended height best when both formants, as opposed to only one, were changed [6].

If pitch perception in whisper is coded through *compensatory* cues, this would be in line with the idea that speakers attempt to match their listeners' needs, and put in more effort when needed, e.g. hyperspeech [10] or clear speech [13]. In [5] F2 and its bandwidth showed interactions of speech mode by intonation condition, suggesting that speakers use it to compensate for the lack of pitch.

The present research is an extension of [5], who investigated acoustic correlates of boundary tones, but in only one vowel setting, /ə/. Compensation strategies, however, have been suggested to vary with vowel quality [11]. Here we investigate four different vowels, and in two different structures: lexical stress and boundary tone do or do not coincide on the final syllable of a sentence. In the former case, the coincidence of nuclear accent and boundary tone causes tonal crowding, which poses a potential challenge for speaker and hearer. Also, by making a direct comparison between whispered and phonated speech we can separate potential secondary from potential compensatory cues.

2. MATERIALS AND METHOD

2.1. Materials

A set of 12 English, 8- to 10-syllable declarative sentences of the form Subject-Verb-Object were constructed such that each could be produced as a statement or an interrogative, solely depending on the prosody, i.e. L% or H% boundary tone. Each sentence ended in a syllable containing one of four vowels, /i, æ, ɔ, u/, together spanning the phonological dimensions [+/–high] and [+/–back].

Per vowel, three word types were recorded: 1-syllable (lexical stress, realized as a nuclear accent, and boundary tone coinciding on same syllable), 2-syllable with initial lexical stress (nuclear accent not coinciding with boundary tone), and 2-syllable with final lexical stress (nuclear accent coinciding with boundary tone). Here, we discuss only the former two types: {*wheel* /wil/, *law* /lɔ/, *moon* /mun/, *man* /mæn/}, and {*baseball* /beisbɔl/, *venue* /vɛnju/, *wombat* /wɔmbæt/, *rally* /ræli/}.

2.2. Participants & procedure

Ten speakers of American English participated (5 males, 5 females, aged 19-31, informed consent obtained). They were recorded individually in a silent room at the University of Rochester, USA, using a Marantz PMD 670 solid state recorder and a Shure SM57 microphone (32 kHz, 16 bits). Participants were compensated for their time.

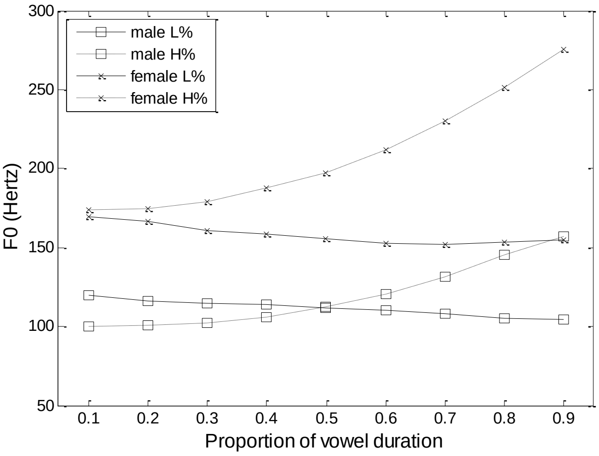
Sentences were presented one at a time in quasi-random order on a computer screen, ordered in blocks of either statements or questions. Half the speakers first produced statements, and the others first did questions. In all cases, phonated versions were recorded before their whispered counterparts. Sentences were presented twice in each condition.

In total, 320 normal and 320 whispered tokens (10 speakers × 2 word types × 2 speech acts × 4 vowel contexts × 2 repetitions) were gathered, from which 29 whispered and 3 phonated tokens were excluded (because of voicing or clipping in the final syllable).

2.3. Annotation & analysis

The recordings were analyzed using Praat [2]. F0 was measured in the phonated versions to check whether speakers produced an intonational difference between the two sentence types. If they did, we assumed that they would attempt to convey the same difference while whispering. Figure 1 shows the mean F0 values (per speaker sex, across speakers and items) over the target vowel for the 1-syllable words. Individual speakers showed the same pattern of results.

Figure 1: F0 course (across vowels) for H% and L% on the 1-syllable words.



On final vowels we measured duration, vowel intensity, energy in three spectral bands, and the formants F1, F2 and F3 (Burg algorithm as implemented in Praat). All measurements were manually checked, and corrected where needed. Spectral energy was measured in the bands .5-1 kHz (B1), 1-2 kHz (B2) and 2-4 kHz (B3), and divided by the total energy over those three bands for normalization. Vowel acoustics, apart from duration, were examined at 80% into the vowel where potential cues may be expected to show, see Fig. 1.

After averaging over repetitions, paired samples t-tests (two-tailed) were conducted within speech modes (normal and whisper) and per word type (1-syllable and 2-syllable) to look for cues to speech act (statement “L%” versus interrogative “H%”).

3. RESULTS

3.1. Vowel duration

In most cases, mean vowel durations were comparable between statements and questions. For whispered 1-syllable words, the means were 285 ms for statements and 288 ms for questions, and 207 and 203 ms for their phonated counterparts. For 2-syllable words, the means in whisper were 174 and 183 ms, respectively, and 129 and 134 ms in phonated speech.

Only for the vowel /i/ were effects of speech act found. In phonated 1-syllable words, final vowels were longer in statements than in questions, t(9)= –2.5, p = .035. In whispered 2-syllable words, final vowels were longer in questions than in statements, t(9) = 3.1, p = .012.

3.2. Intensity

Mean vowel intensities (at 80% relative vowel duration) were generally higher in questions than statements. In whispered 1-syllable words the means were 59 dB for questions and 54 dB for statements, and in the phonated versions the means were 73 and 67 dB. In the 2-syllable whispered words the means were 55 and 51 dB, respectively, and 71 and 65 dB in the phonated versions.

The same effect, i.e. greater intensity for questions than for statements, was found for all whispered vowels, except for front vowels in monosyllabic words (at least t(9) = 2.3, p < .05).

3.3. Formants

Mean formant values for phonated speech were largely comparable to values reported in [8]. For /i/, F2 for the 1-syllable words was lower, probably because of diphtongisation in the transition to /l/ at 80% into the vowel. At 50% into the vowel means were in line with the literature. F2 for /u/ tended to be higher, i.e. more fronted, in both 1-syllable and 2-syllable words. Figures 2 and 3 show F2-by-F3 plots for whispered vowels in 1-syllable and 2-syllable words, respectively.

For the vowel /æ/, the F2 in final vowels of 2-syllable whispered words was higher in questions (1818 Hz) than in statements (1746 Hz), t(9) = 2.9, p = .017. In phonated 2-syllable words, F3 was about 90 Hz higher in questions than statements, t(9) = 2.4, p = .038.

Figure 2: F2-by-F3 plots for the four whispered vowels in monosyllabic contexts, with either high or low boundary tones.

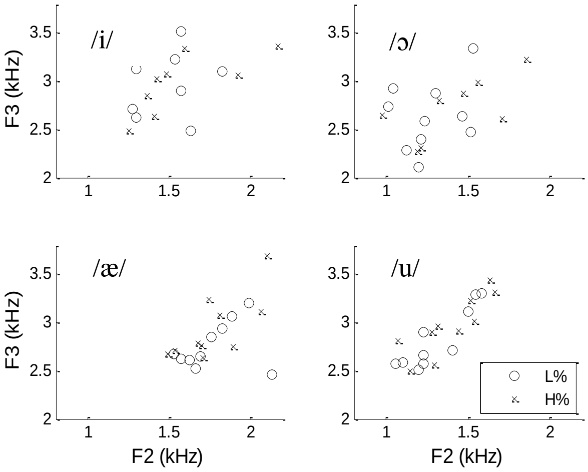
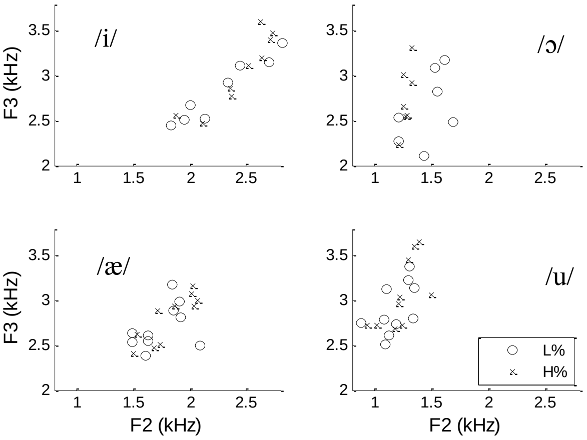


Figure 3: F2-by-F3 plots for the four whispered vowels in disyllabic contexts, with either high or low boundary tones.



For /ɔ/, the F1 in statements was higher than in questions when phonated in a 2-syllable word, t(9) = –3.1, p = .012. In whispered 1-syllable words both F2 and F3 were higher in questions (F2: 1392 Hz, F3: 2963 Hz) than in statements (F2: 1308 Hz, F3: 2820 Hz), t(9) = 2.3, p = .046; t(9) = 3.6, p = .006. The F3 change was also observed in the 2-syllable words, t(9) = 3.1, p = .012.

For /i/, F3 on whispered 2-syllable words was higher in questions (3020 Hz) than statements (2838 Hz), t(7) = 2.8, p = .027.

For /u/, F2 was higher in statements (1396 Hz) than questions (1247 Hz) in whispered 2-syllable words, t(9) = –2.6, p = .030. F2 was higher in questions (1413 Hz) than statements (1274 Hz) in whispered 1-syllable words, t(7) = 3.1, p = .018.

3.4. Spectral energy in bands

The relative energy per band was compared between speech acts, within speech modes. For 1-syllable words only whispered /ɔ/ showed differences between speech acts in B1, t(9) = –3.0, p = .014, and in B2, t(9) = 2.8, p = .021. In B1, energy was higher in statements; in B2 it was higher in questions.

For the 2-syllable words, spectral changes were only found for phonated /æ/. B1 to B3 differed between speech acts. B1 contained relatively more energy in questions than statements, t(9) = 4.3, p < .01. B2 and B3 showed more energy in statements than in questions, t(9) = –4.0, p < .01 and t(9) = –3.8, p < .01, respectively. One marginal difference was found for whispered /u/ in 2-syllable words: its energy in B3 tended to be higher in statements than in questions (p = .065).

4. DISCUSSION

We investigated acoustic cues to boundary tones in whispered speech in four different vowels, and in two prosodic structures (nuclear accent coinciding with boundary tone or not). Also, by directly comparing whispered and phonated speech we intended to look for potential secondary and compensatory cues.

Firstly, the acoustic cues to boundary tones in whisper seemed to be more pronounced when lexical stress and boundary tone did not coincide on the same syllable. The crowding of nuclear accent and boundary tone seemed to affect the speakers' ability to produce differences correlated with speech act. Secondly, the potential cues to boundary tone were not constant across vowel contexts, as first proposed by [11].

Overall intensity was confirmed as a secondary cue, since questions were louder than statements for both speech modes [5], in most vowel contexts. The secondary role for F1 that had been reported earlier [5, 7], could not be confirmed by our data.

Formant shifts were found to correspond with intended boundary tone in whisper. Many of these had no parallel in phonated speech, and can therefore be interpreted as potential compensatory cues to intended pitch height. The common denominator was F2 changing with boundary tone in most whispered vowels. This was in line with the findings of [5, 7]. Between whispered monosyllabic and disyllabic words the direction of F2 changes in /u/ varied with speech act. Through analysis of the remaining recordings and additional data we hope to better understand this variation.

F3 contributed in /i/- and /ɔ/-contexts, and earlier it had already been suggested to contribute to intended pitch in whispered /a/ [4, 11]. Note that there also were a few formant changes in phonated speech that varied with intended pitch that were not paralleled in whisper: F3 in /æ/ and F1 in /ɔ/.

Duration provided compensatory information in only one out of four vowel contexts. In general, the slower pronunciation of whispered speech might still be helpful for listeners [9], e.g., to pick up other cues. Alternatively, or in addition, it may indicate the speakers' relative difficulty with whisper as a speech mode.

The distribution of energy over wider spectral bands did not correspond with boundary tones. This may be explained by the fact that in whisper we expect the expression of boundary tones to be accomplished mainly through changes in filter characteristics, whereas spectral tilt has earlier been found to be mainly influenced by changes in the source [14].

5. ACKNOWLEDGEMENTS

This research was supported by a VENI grant from The Netherlands Organisation for Scientific research, awarded to the first author.

6. REFERENCES

[1] Abramson, A.S. 1972. Tonal experiments with whispered Thai. In Valdman, A. (ed.), *Papers on Linguistics and Phonetics to the Memory of Pierre Delattre*, 29-44.

[2] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

[3] Denes, P. 1959. A preliminary investigation of certain aspects of intonation. *Lang. Speech* 2(2), 106-122.

[4] Fonagy, J. 1969. Accent et intonation dans la parole chuchotée. *Phonetica* 20, 177-192.

[5] Heeren, W., van Heuven, V.J. 2009. Perception and production of boundary tones in whispered Dutch. *Proc. Interspeech 2009* Brighton, 2411-2414.

[6] Higashikawa, M., Minifie, F.D. 1999. Acoustic-perceptual correlates of "whisper pitch" in synthetically generated vowels. *J. Speech, Lang. Hear. Res.* 42, 583-591.

[7] Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H. 1996. Perceived pitch of whispered vowels-relationship with formant frequencies: a preliminary study. *J. Voice* 2, 155-158.

[8] Hillenbrand, J.M, Getty, L.A., Clark, M.J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 3099-3111.

[9] Krause, J.C., Braida, L.D. 2002. Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.* 112, 2165-2172.

[10] Lindblom, B. 1996. Role of articulation in speech perception: clues from production. *J. Acoust. Soc. Am.* 99, 1683-1692.

[11] Meyer-Eppler, W. 1957. Realization of prosodic features in whispered speech. *J. Acoust. Soc. Am.* 19, 104-106.

[12] Miller, J.D. 1961. Word tone recognition in Vietnamese whispered speech. *Word* 17, 11-15.

[13] Picheny, M.A., Durlach, N.I., Braida, L.D. 1986. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *J. Speech Hear. Res.* 29, 434-446.

[14] Sluijter, A., van Heuven, V.J. 1996 Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am.* 100, 2471-2485.

[15] Tartter, V.C., Braun, D. 1994. Hearing smiles and frowns in normal and whisper registers. *J. Acoust. Soc. Am.* 96, 2101-2107.