Tuning in to whispered boundary tones

Willemijn Heeren [1], *Sarah Bibyk* [2], *Christine Gunlogson* [3], *Michael K. Tanenhaus* [2]

[1] Leiden University Centre for Linguistics, Leiden University, The Netherlands
[2] Department of Brain and Cognitive Sciences, University of Rochester, NY, USA
[3] Department of Linguistics, University of Rochester, NY, USA

w.f.l.heeren@hum.leidenuniv.nl, {sbibyk,mtan}@bcs.rochester.edu, gunlog@ling.rochester.edu

Abstract

Very little is known about how listeners incorporate "intonational" information in whispered speech during online language processing. We present data showing that listeners can incorporate information about boundary tones in whispered speech rapidly, but this process is complicated by additional structural biases as well as by the fact that speakers do not produce cues to boundary tones consistently in whisper. Listeners, however, are able to adapt to these differences in order to correctly identify different boundary tones in whisper.
Index Terms: boundary tones, online processing, whispered speech

1. Introduction

In earlier work we investigated the online processing of high (H%) versus low (L%) boundary tones in normal speech, and found that they are processed as quickly as pitch accents [1-5], and with very few interpretation errors. The main acoustic cue to boundary tones, and intonation in general, is thought to be the speaker's fundamental frequency (f0). But when a speaker whispers, f0 is not being produced. This paper begins to address how listeners process boundary tones produced in whispered speech when f0 is absent.

Offline studies have shown that listeners can identify and discriminate boundary tones in whispered speech [6,7]. Performance is worse than in normal speech, but well above chance level, indicating that there are prosodic cues to boundary tones available in whispered speech as well. However, it is unclear how listeners process whispered boundary tones online and which cues they use in doing so. Using a targeted language game, an acoustic analysis of multiple speakers' boundary tones, and a crowd-sourcing perception experiment, the online processing of and adaptation to whispered boundary tones were investigated.

2. Online processing of whispered boundary tones

In earlier research, we developed a "targeted language game" using the visual world eye-tracking paradigm [8] that is sensitive to the time-course of processing boundary tones: [9]. Here we used that paradigm to study the online processing of whispered speech. The participant played a card game against the computer by means of a verbal interaction, and the game was designed in such a way so that on critical trials only the boundary tone indicated whether the computer's move was a statement (signaled by L-L%) or a yes-no question (H-H%). Syntactic cues were removed by having the computer use elliptical sentences of the form '*Got a <card category>*', which could be elliptical versions of either '*I have got a <card category>.*' or '*Have you got a <card category>?*'. The game elicited different actions (thus different fixation patterns) from the participant in response to questions vs. statements, allowing us to assess listeners' online categorization of boundary tones.

The goal of the game was for the opponents to discard cards from their (virtual) hands by matching them to a match card, a face-up card in the middle of the screen (Fig. 1a). Each player also had a stack of block cards, the top one of which could be used to block the other's matches (Fig. 1b). Upon perceiving a question from the computer, the player would look at the playing cards, and upon perceiving a statement, the player would look at the block card. There were four card categories, each represented by a black and white line drawing shown on the match, playing or block card (shoe [ʃuː], wheel [wiːl], candy [kændi], window [wɪndoʊ]). The center of the match card was placed equidistant to the center of the mean size of the playing card set, and the center of the block card.

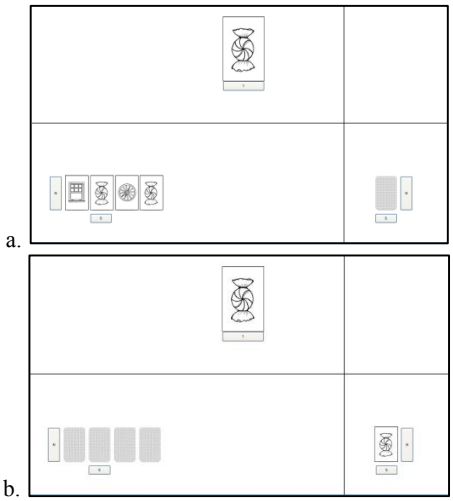

Figure 1: *Screenshots of the game with the match card (top center), the player's playing cards (bottom left) and block card (bottom right); the player (a) looks for possible matches when asked; and (b) looks to block the computer's stated move.*

2.1. Materials

The computer's whispered moves were pre-recorded: eight were target sentences (4 statements/ 4 questions) and nine were fillers (3 statements/ 6 questions). Fillers were used to introduce syntactic variation into the computer's speech (e.g., '*Do you have a candy?*'). Also recorded as computer 'utterances' were moves to keep track of turns during the game (e.g., '*It's your/my turn.*'), to respond to the player's

10.21437/SpeechProsody.2014-108

questions for a match (e.g., '*Yes.*' or '*No.*') and to block the player's match (e.g., '*I am blocking you.*'). The utterances representing the computer's moves were recorded in a sound-treated booth using an Audio-technica ATM75 head-worn microphone and a Marantz PMD 670 solid state recorder (mono, 32 kHz, 16 bits). The speaker was a 23-year-old female native speaker of American English.

Table 1 presents acoustic measurements taken over the target words' final syllables, that is the boundary tone landing sites. It shows mean intensity, duration, and the first through third formants, measured over the mid 50 ms of a vowel using the Burg method implemented in PRAAT [10]. The measurements show a comparable duration for statements and questions, a higher intensity in questions than statements, and in many cases higher formant values for questions.

Table 1. *Acoustic content of the final vowels, per target word, spoken as statement (S) or question (Q).*

	Speech act	Int. (dB)	Dur. (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)
candy	S	54.0	171	452	2720	3089
	Q	60.8	169	537	2925	3352
shoe	S	60.5	323	570	1679	2748
	Q	62.8	318	611	1832	2900
wheel	S	66.2	390	642	2551	2901
	Q	67.8	399	771	2653	3014
window	S	65.1	254	790	1766	2744
	Q	70.7	236	858	1683	2844

2.2. Participants and procedure

Fifteen American English participants were recruited at the University of Rochester, NY, USA (informed consent obtained). They were given both written and oral instructions. During testing the verbal interaction was recorded using a Realistic 33-984A Highball dynamic unidirectional table microphone placed between the computer speakers and the player, so that it would register both interlocutors. Eye movements were recorded using a head-worn ASL eye-tracker at 30 Hz, and a Sony DSR-30 digital video recorder, with Sony 184 DVCAM digital videotapes. Before the test started, the participant played a practice game that contained all possible game situations. Calibration was checked throughout the test game and always occurred before a new turn for the participant. The entire session lasted 30 to 45 minutes.

The order of events (question vs. statement) during the game was fully determined, but the order for the card items (candy, shoe, etc.) was rotated and balanced across four lists. For target utterances, the wave file started 1400 ms after the match card had changed. For filler items, the preceding silence was variable, but at least 1200 ms, thus introducing variation to increase the naturalness of the computer's utterances, as a real player also would not respond at regular intervals. The computer utterances were played to the participants at a comfortable listening level over computer speakers.

2.3. Results and discussion

The 33 ms video frames were coded manually from the onset of a target utterance until the participant's verbal response, i.e. a variant of '*I can('t) match/block*'. Five locations were coded: (1) the playing cards, (2) the match card, (3) the block card, (4) other on-screen locations, and (5) track loss. Saccade-initial frames were counted as fixations to the landing site.

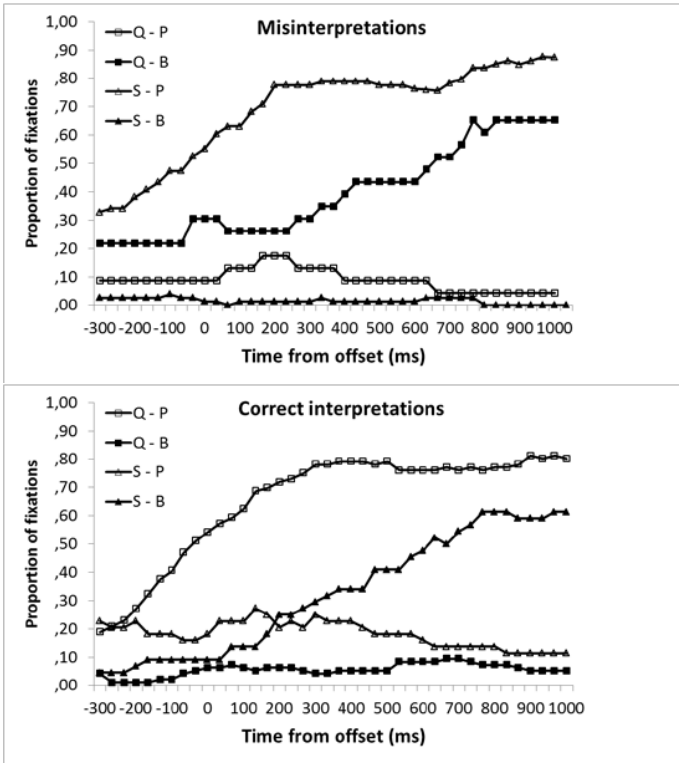

Figure 2: *Proportions of fixations to the playing cards (P) and block card (B), in the case of intended questions (Q) and statements (S). Misinterpretations and correct interpretations are plotted separately.*

Out of 240 trials, 140 were correctly interpreted (58%), 99 were misinterpreted (41%), and one trial was misunderstood. Misinterpretation means that intended questions were interpreted as statements, and the other way around. When looking at the division of correctly versus incorrectly interpreted trials per speech act, 76 out of 120 statements were misinterpreted, but only 23 out of 120 questions. On average, questions, but not statements, were correctly recognized above chance level.

Because of the large number of misinterpretations, eye movements to both correctly interpreted and misinterpreted trials were analyzed separately (see Fig. 2). For questions and statements the ratio of targets divided by the sum of targets and distractors, T/(T+D), was computed and compared in a two-tailed paired samples t-test, see [11]. This was done for two intervals, one before target offset (-300 ms to 0 ms) and one after target offset (0 ms to 300 ms). These intervals were chosen because they represent the regions before and after where the earliest effects of boundary tones are expected.

For correctly interpreted trials, no significant difference between the speech acts was found during the first analysis interval, $t(7)=0.8$, n.s, whereas during the second interval, a significant difference was found, $t(12)=2.7$, $p=0.020$. This pattern is in principle consistent with sensitivity to prosodic cues. When looking at the questions, however, there is an early preference for the playing cards that does not seem to be time-locked to the boundary tone (< 0 ms). In statements, the pattern of fixations is consistent with interpretation of the boundary tone, showing an increase in looks to the block card once the boundary tone becomes available (> 0 ms).

For misinterpreted trials, both analysis intervals showed a significant difference between the speech acts, $t(6)$=-5.2, p=0.002 and $t(8)$=-5.1, p=0.001, respectively. Eye-movements therefore do not seem to be time-locked to the prosodic events. In the case of misinterpreted questions (as statements), there is an initial bias to look at the blocking card early in the utterance, followed by a later increase in looks to the blocking card which is too delayed to reflect a time-locked response to the boundary tone. In the case of misinterpreted statements (as questions) there is an initial bias to look to the playing cards which begins to increase well before the final syllable.

The results show only a weak reliance on prosodic cues in whisper. Firstly, many trials were misinterpreted which indicates that prosodic cues to speech act were not correctly used or not used at all. Secondly, in correctly interpreted questions the proportion of fixations to the target increased well before boundary tone information became available. Only in correctly interpreted statements, does the time course of fixation proportions suggest that prosodic cues were used. Because of the predetermined order for moves in the game, the majority of statements occurred during the second half of the game. We used a mixed-effects logistic regression to explore the effect of trial number and speech act type on participants' answers. The model with the fixed and random effects structure most justified by the data (as accessed by model comparison) did not contain a significant intercept (B=0.19, z=0.80, p=0.42) nor a main effect of trial number (B=0.057, z=0.60, p=0.55), but it did show a significant main effect of speech act type (B=-1.21, z=-6.54, p<0.001); questions were more likely to be answered correctly than statements. The interaction between type and trial number was also significant (B=0.22, z=5.25, p<0.001). Separate logistic regressions by speech act type revealed that participants improved in performance across trials only on statements, B=0.28, z=3.89, p<0.001. On questions listeners actually became worse across trials (B=-0.21, z=-2.99, p=0.003), but remember that overall participants performed better on questions than on statements.

Earlier research [e.g. 6,7] and Table 1 suggest that acoustic cues to speech act type were present. There are several explanations as to why this information was not fully used. Listeners may either be relatively insensitive to the prosodic information that the speaker attempted to convey, perhaps because whisper as a speech mode is not used very often and the cue that normally carries boundary tones, f0, is absent. Listeners might also find it difficult to extract prosodic information because speakers may not provide consistent cues to intonation in whisper. These possibilities were addressed further in section 3.0.

3. Variation in boundary tone realization

Gotta-utterances from three additional speakers were recorded and speaker strategies for signalling prosody were compared to explore if the listeners' difficulty with prosodic cues in whisper may be explained by varying speaker strategies.

3.1. Method

Two male and one female native speaker of American English were recorded [Shure SM57 microphone, mono 44.1 kHz, 16 bit] using a script that took speakers through a game scenario intended to elicit each target utterance twice. Participants were told the outline of the game and then asked to imagine themselves as best as possible in the game scenario, saying the sentences how they would say them if they were really playing. They were not asked to make the questions and statements as acoustically distinct as possible; they were just told to be as clear as possible given that they would be whispering. They were also told to take their time and read the scenario descriptions carefully so that they would be in the right mindset for producing the utterances.

Recordings were annotated at the segment level, and for all final-syllable vowels we measured the relative syllable duration, mean intensity, and the first through third formants over the vowel's mid 50 ms using the Burg method implemented in PRAAT. Formant measurements were visually verified in the spectrogram (some F1s, mainly of [u], could not be determined). These acoustic characteristics were selected as they have been put forward as cues to whispered tones and/or intonation [12-15]. Comparisons between the speech acts were done per speaker, using paired samples Wilcoxon signed ranks tests.

3.2. Results and discussion

Relative vowel duration was longer in questions than statements for Speaker 1 only, Z=-2.5, p=0.012. The difference was marginal for Speaker 2 (p=0.058), and non-significant for Speaker 3 (p=0.16). Mean vowel intensity showed no significant differences between statements and questions for any of the speakers. Due to data sparsity, effects in F1 were hard to measure, but no significant differences were found, and no consistent trends were observed. For F2, Speaker 3 produced a significantly higher second formant (F2) in questions compared to statements (Z=-2.1, p=0.036). The mean difference was 82 Hz. Speaker 2 showed a tendency in the same direction with a 119 Hz mean difference (p=0.077). As for F3, speakers 1 and 3 had a higher value in questions than statements, with mean differences of 120 Hz and 96 Hz, respectively (both Z=-2.5, p=0.012). Speaker 2's data showed a comparable trend with a 92 Hz mean difference.

The results show that speakers vary in the acoustic dimensions that they use to signal different speech acts, along three dimensions that have been proposed as potentially contributing to the expression of intonation in whisper. Speakers provided both durational and spectral cues, but not in exactly the same way. Spectral differences, which are assumed to provide the most direct cue, were also present in the productions of Experiment 1's speaker. Duration may be taken as a secondary cue, through lengthening of the speech act that requires most production effort, the question. Two out of three speakers made a durational difference, but this had not been the case for the speaker of Experiment 1. Whereas that speaker seemed to vary vowel intensity with speech act, a comparable intensity difference was not found for the other three speakers.

4. Tuning in to whispered prosody

Taken together, results of Experiment 1 and the analysis of multiple speakers suggest that listeners' difficulty with the extraction of prosodic information from whispered utterances may arise because different speakers provide different cues to prosody in whisper. Therefore listeners require exposure to a particular speaker to learn the relevant cues that remain when f0 is absent. The fact that 'tuning in' took place is supported by the increased number of correctly interpreted trials in the second half of Experiment 1. To further investigate if participants are in fact tuning in to prosodic information or

were just making lucky guesses, listeners were exposed to different patterns of acoustic cues to test the hypothesis that they adapt to and use the cues in each case.

4.1. Method

A Web survey was conducted in which participants provided responses to whispered audio samples of statements and questions produced by one of the three speakers from Experiment 2. Participants were instructed that they would hear a whispering speaker play a simple card game, and that they would be asked to indicate whether they heard the player make a match (=statement response) or ask for a card (=question response), by clicking one of two buttons. Feedback on the correctness of their answers was provided, both during practice and testing. During practice, sentences from Experiment 1's fillers were used (i.e. different speaker, no elliptical structure). During testing, there were 16 trials: 2 repetitions of each target word (4) in each speech act (2).

Using the online crowd-sourcing service Amazon's Mechanical Turk, 258 Human Intelligence Tasks (HITs) were posted for (self-reported) American English participants (253 unique participants). Per speaker, four lists were used, each with a different order for the response options. In a pretest it was established that listeners' home equipment was set up to perceive whispered speech well. Ten participants were eliminated because sound files did not play properly during the pretest. Three additional participants were eliminated due to experimenter error, leaving 240 participants (80/speaker, 20/list). Participants were paid $2 for their efforts.

4.2. Analysis, results and discussion

The data were analysed with a hierarchical mixed-effects logistic regression with Trial Number, Speech Act Type, and Speaker as predictors using the maximal random effects structure as justified by the model, which included random intercepts by subject and by item, as well as random slopes by item over trial. The main effects of trial, type, and speaker were entered first, followed by the two-way interactions, followed by the three-way interaction.

We used model comparison to select the model most justified by the data. The final model included the intercept, and the main effects of trial, type, and speaker. This model accounted for a significant portion of the variance above and beyond the model with just the intercept $\chi^2(4)=22.62$, $p<0.001$. The intercept itself was significant ($B=0.22$, $z=2.53$, $p<0.05$), indicating that on average listeners performed above chance. The main effect of trial was significant ($B=0.056$, $z=2.82$, $p<0.01$); on average, as trial number increased, performance improved. The main effect of type was also significant ($B=-0.22$, $z=-2.67$, $p<0.01$); on average questions were more often answered correctly as compared to statements. The intercepts for Speakers 1 and 3 were significantly different from each other ($B=0.10$, $z=3.27$, $p<0.01$). The intercepts for Speakers 1 and 2 were not significantly different ($B=0.050$, $z=1.57$, $p=0.12$). The intercepts for Speakers 2 and 3 were marginally different ($B=0.052$, $z=1.71$, $p=0.088$). Separate logistic regressions for each speaker revealed a significant intercept for speaker 1 ($B=0.33$, $z=2.67$, $p=0.008$), a marginal intercept for speaker 2 ($B=0.199$, $z=1.66$, $p=0.097$), and a non-significant intercept for speaker 3 ($B=0.043$, $z=0.31$, $p=0.76$).

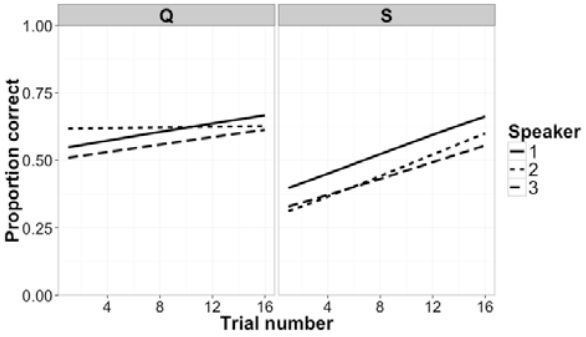

Figure 3: *Proportions of correct answers across trials plotted for the three speakers separated by questions (left panel) and statements (right panel).*

First, as Figure 3 shows, there was an overall trend across all speakers for listeners to improve performance as trial number increased. Thus, while it is true that different speakers signal prosody differently in whisper, listeners are able to adapt to those differences. This result contrasts with Experiment 1 where a main effect of trial was not found. Recall, however, that only one ordering for question and statement moves was used in Experiment 1, whereas four lists were used here. Second, there was a trend for listeners to perform better on questions than on statements (see Fig. 3), just as in Experiment 1. This difference seems to indicate a bias to interpret the "got a" construction as a question as opposed to a statement, see [16]. Third, we have some evidence that productions of different speakers have different effects on how listeners are able to interpret their intentions in whispered speech. The listeners who received Speaker 1's stimuli performed above chance overall (61% correct), in comparison to the listeners receiving Speaker 2, who were marginally above chance (58%), and Speaker 3 who were not statistically different from chance (54%). Possibly, listeners tuned in more quickly to Speaker 1 because that speaker significantly changed two acoustic dimensions, spectral and durational, between the speech acts rather than just one, spectral or durational.

5. Conclusion

We have replicated the effect that listeners can distinguish boundary tones in whispered speech. In addition we have provided provisional evidence that speakers can incorporate information about the boundary tones in real time processing of whispered sentences. This effect is complicated by the fact that listeners also appear to take into account cues from the lexical content of the utterance (a question bias for "got a"), and by the fact that cues to boundary tones are more difficult to distinguish in whispered speech. In addition, speakers have different strategies for how they signal boundary tones in whisper. Therefore, compared to the more systematic f0 cue in regular speech, listeners may need more exposure before they can utilize (speaker-dependent) systematic cues to prosody in whispered utterances. We note that there is emerging evidence that listeners adapt to the reliability of different prosodic cues for individual speakers even in normal speech [17]. Therefore, the mechanisms used by listeners to process prosody in whispered speech might partially be similar to those used when processing normal speech.

6. References

[1] Dahan D., Tanenhaus M. K. and Chambers C. G., "Accent and reference resolution in spoken-language comprehension", J. Mem. Lang., 47:292-314, 2002.

[2] Watson, D. G., Gunlogson, C. A. and Tanenhaus, M. K., "On-line methods for the investigation of prosody", in I. Mleinek [Ed.] Methods in Empirical Prosody Research. Berlin: Mouton de Gruyter, 2006.

[3] Weber, A., Braun, B. and Crocker, M. W., "Finding referents in time: Eye-tracking evidence for the role of contrastive accents", Lang. Speech, 49:367-392, 2006.

[4] Ito, K. and Speer, S. R., "Anticipatory effects of intonation: eye movements during instructed visual search", J. Mem. Lang., 58:542-573, 2008.

[5] Watson D. G., Tanenhaus M. K. and Gunlogson C. A., "Interpreting pitch accents in online comprehension: H* vs. L+H*", Cogn. Science, 32:1232-1244, 2008.

[6] Fónagy, J. (1969). "Accent et intonation dans la parole chuchotée," Phonetica, 20:177-192, 1969.

[7] Heeren, W. F. L. and Van Heuven, V. J., "Perception and production of boundary tones in whispered Dutch", in Proc. Interspeech 2009, Brighton2411-2414, 2009.

[8] Brown-Schmidt, S. and Tanenhaus, M. K., "Real-time investigation of referential domains in unscripted conversation: A targeted language game approach", Cogn. Science, 32: 643-684, 2008.

[9] Bibyk, S., Heeren, W., Gunlogson, C. and Tanenhaus, M. K., "Asking or telling? Real-time processing of boundary tones", LSA annual meeting, Boston, January 3-6 2013, 2013.

[10] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]," retrieved from http://www.praat.org/, 2013.

[11] Dahan, D. and Tanenhaus, M. K., "Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition", Psychon. B. Rev., 12:453-459, 2005.

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

[13] Denes, P., "A preliminary investigation of certain aspects of intonation," Lang. Speech, 2:106-122, 1959.

[14] Liu, S. and Samuel, A. G., "Perception of Mandarin lexical tones when F0 is neutralized," Lang. Speech, 47:109-138, 2004.

[15] Meyer–Eppler, W., "Realization of prosodic features in whispered speech," J. Acoust. Soc. Am., 19:104-106, 1957.

[16] Bibyk, S., Heeren, W., Gunlogson, C. and Tanenhaus, M. K., "Asking or Telling - Real-time processing of boundary tones", submitted.

[17] Kurumada, C., Brown, M. and Tanenhaus, M. K., "Pragmatic interpretation of contrastive prosody; It looks like speech adaptation", in N. Miyake, D. Peebles and R. P. Cooper [Eds.], Proc. 34th Annual Conference of the Cognitive Science Society, 647-652, 2012.