INTERSPEECH 2009 BRIGHTON

Perception and Production of Boundary Tones in Whispered Dutch

W. Heeren, V.J. Van Heuven

Leiden University Center for Linguistics, Leiden University, The Netherlands
w.f.l.heeren@hum.leidenuniv.nl; v.j.j.p.van.heuven@hum.leidenuniv.nl

Abstract

The main cue to interrogativity in Dutch declarative questions is found in the final boundary tone. When whispering, a speaker does not produce the most important acoustic information conveying this: the fundamental frequency. In this paper listeners are shown to perceive the difference between whispered declarative questions and statements, though less clearly than in phonated speech. Moreover, possible acoustic correlates conveying whispered question intonation were investigated. The results show that the second formant may convey pitch in whispered speech, and also that first formant and intensity differences exist between high and low boundary tones in both phonated and whispered speech.

Index Terms: speech perception, speech production, whispered speech, acoustic cues, boundary tones

10.21437/Interspeech.2009-302

1. Introduction

The goal of this paper is to study question intonation in whispered speech. For this purpose we will focus on Dutch, a language for which question intonation in regular, phonated speech, has been systematically researched [1]. In phonated speech, the fundamental frequency, duration and intensity of speech sounds play a role in producing question-related prosody. In whispered speech, however, the vocal folds do not vibrate, causing the fundamental frequency --a crucial acoustic cue-- to be absent.

There are several views possible on the nature of whispered speech. One possibility is that whispered speech is simply like speech that is spoken aloud, but without the vibration of the vocal cords. In this case one could say that its nature is reduced. However, another possibility is that the lack of fundamental frequency is compensated for in some way in whispered speech. If that were the case, this could be interpreted as some kind of signal enhancement. In line with Lindblom´s Hypo & Hyperspeech theory [2], this study will approach whispered speech as a speech mode that needs extra effort in production in order to compensate its reduced intelligibility. Therefore, because of the absence of vibration by the vocal cords, whispered speech is expected to get some extra attention from the speaker, in order to be understood by the listener.

There have been a number of studies into the realisation of pitch correlates in whisper. These were mainly on the production and perception of whispered tones on individual syllables or words, e.g., [3][4]. These experiments have suggested a role for the first two formants, i.e. the two lowest resonance frequencies of the vocal tract, e.g., [3][5][6][7]. As earlier studies mainly focused on tone languages, and on relatively small units, we want to study how intonation is conveyed in whispered speech at the sentence level. Moreover, we are interested in the question of whether other acoustic correlates, such as intensity or duration, play a role.

Research on question intonation in Dutch has found that speakers use more prosodic means to mark questions when the sentence's syntax gives the listener little (or no) cues [1]. Declarative questions are questions that have no overt syntactic markers of their interrogative status, as their word order is the same as for statements, see example (1). The pitch cue that makes the main difference between the sentences in (1) is the presence/absence of a sentence-final pitch rise.

(1) a. John wants to sell his car.

b. John wants to sell his car?

As such questions in phonated speech are clearly marked by local and also by global changes in pitch, the absence of the fundamental frequency in whispered speech may challenge speakers in conveying the difference between statements and questions. Since there is some evidence, however, that speakers may compensate for this loss, we intend to investigate how it is done in the case of Dutch declarative questions versus statements. We will do this by assessing the perception and production of boundary tones.

2. Perception study

In this perception study we investigated how well Dutch listeners identify the sentence mode in whispered sentences, i.e. whether utterances are perceived as statements (L%) or questions (H%). As listeners can identify whispered tones in tone languages, we expect them to also be able to identify whispered boundary tones in an intonation language.

2.1. Materials

The test sentence is given in (2), where [ˈ] stands for an accented syllable. It can be pronounced as both a statement and as a question, depending on prosody only. The pitch movement directly associated with the boundary tone is located on the final syllable of the test sentence.

(2) Alumni willen miljoenen verdienen . / ?
[aˈlœmni ʊɪlə mɪlˈjunə vərˈdinə]
alumni want millions to earn
'alumni want to make millions'

The materials for this analysis were recorded as part of a larger study into the production and perception of intonation contours in whispered speech [8]. Orthogonal to a set of four accent patterns there were two boundary tones: a terminal low tone (L%) indicating a statement and a rising one (H%) indicating a question (see Figure 1). Combined, these formed eight intonation contours found in Dutch [9]. The four accent patterns were fitted onto the first three words of the sentence, the boundary tones onto the final word.

Six native Dutch speakers participated in the production session (three male, three female). Five of them were staff members at the phonetics laboratory of Leiden University. They were phonetically trained. The sixth speaker was a PhD student in linguistics. Speakers received written instructions in

which the intonation contours were drawn and annotated with a conventional transcription [10]. Subjects were asked to produce the intonation contours both lexically, as in example (2), and in reiterant speech [11], as in example (3). In reiterant speech the same syllable, e.g., '*na*' [na], is substituted for target syllables in the original sentence, while inheriting the accentuation associated with the original syllable. By making a comparison between syllables consisting of the same speech sounds, but with different prosody, exactly those acoustic cues that convey intonation can be isolated.

(3) Nanana willen nanana verdienen.

Furthermore, speakers were asked to pronounce all sentences both whispered and phonated. In studying intonation in whispered speech, phonated speech was used as comparison. The speakers were given ample time to practise.

The sentences were digitally recorded (48 kHz) in a soundproofed room using a directional Sennheiser MKH-416 condensor microphone. The recordings were downsampled to 16 kHz and further processed on a Silicon Graphics computer. Speakers produced each sentence at least twice, but only one instance was chosen for further processing through auditory selection and visual inspection of their pitch contours in the phonated versions. This furthermore showed that one female speaker in general had some difficulty producing the contours.

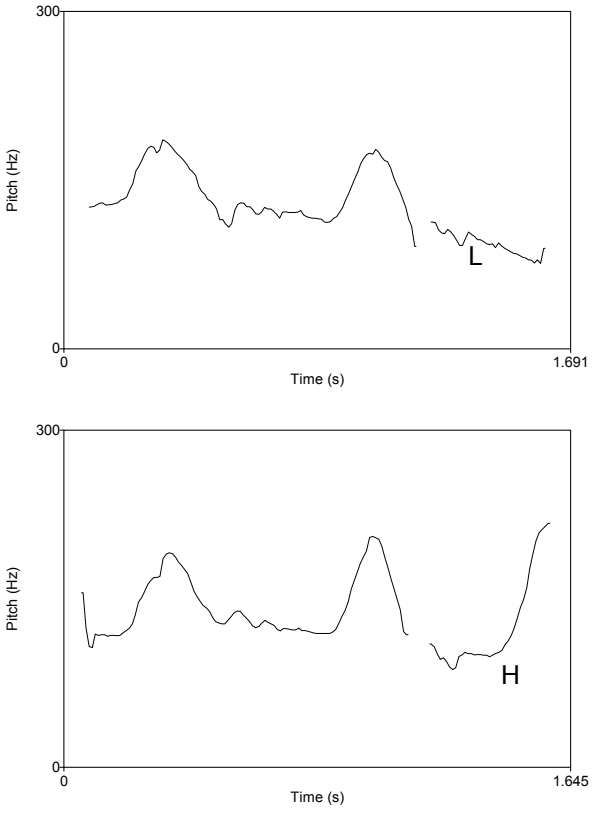

Figure 1: *Pitch contours for a phonated statement (upper panel) and question (lower panel) pair.*

2.2. Listeners

Twenty native speakers of Dutch participated in the perception experiment (14 female, 6 male). Their ages varied from 19 to 28 years with a mean of 21. All were students or had recently completed their master's degree.

2.3. Procedure

The 192 sentences were divided into four categories: lexical & phonated (LexPhon), reiterant & phonated (RePhon), lexical & whispered (LexWhis), and reiterant & whispered (ReWhis). The sentences within each category were ordered pseudo-randomly, and two consecutive stimuli never shared more than one of the properties Accent Pattern, Boundary Tone or Speaker. The four parts of the experiment were presented in two different orders. Half of the subjects heard the order LexPhon-RePhon-LexWhis-ReWhis, and the other half heard the order LexWhis-ReWhis-LexPhon-RePhon. Per speech mode, the lexical sentences were always presented before the reiterant ones, because it was thought to be unnatural if the meaningless reiterant sentences were presented without prior experience to their lexical versions. The speech modes were alternated to counterbalance habituation and learning effects.

The test was run in a quiet room. Subjects received written instructions. They were instructed to decide for each stimulus which intonation contour was most similar to the one they heard. This choice was made by marking both a) one out of four accent patterns, and b) one out of two boundary tones on the answer sheets, on which pictures of the patterns and boundary tones were drawn. Subjects were instructed to provide a response even when they were not sure or if they felt a choice could not be made. Before the test began, eight sentences were presented as an introduction. These were taken from the later test at random.

The stimuli were presented binaurally over Sennheiser MD-424 headphones. The stimuli (16 bit, 16 kHz) were presented on-line from a Silicon Graphics computer. Since two choices had to be made, one stimulus consisted of a single utterance presented twice with a 500 ms interval. Between stimuli a 3000 ms response time was given. After every ten stimuli a 1000 Hz tone was presented with a duration of 100 ms as a reference point for the subjects. Each part of the test comprised 50 pairs of sentences. The first two stimuli were the same as the last two and served as an introduction. These were not included in the analysis. Between the four parts of the test there were short breaks. Each part lasted about seven minutes.

2.4. Analysis and results

To examine the performance of the listeners, the percentages of correct identifications for the tasks of choosing an accent pattern and boundary tone were examined individually across speech modes. Identifying an accent pattern was much more difficult than identifying a boundary tone. To exclude the subjects who performed poorly, a performance limit was fixed at chance level of the accent pattern identification task, i.e. 25% correct overall. The performance limit was reached by 18 out of 20 subjects. Moreover, averaged recognition scores showed that listeners scored lowest on the sentences from the speaker, whose *phonated* F0 contour had least resembled the intended intonation movements. Therefore, one speaker and two listeners were excluded from further analysis. Only the boundary tone identification results are presented here.

The mean scores per speech mode and per boundary tone were computed for each of the listeners. A two-way ANOVA with Boundary Tone and Speech Mode as fixed factors was run. A Speech Mode × Boundary Tone interaction was found, $F(1,71)=9.3$; $p<0.05$, showing that question identification was more difficult in whispered than in phonated speech (79% v 95%, respectively), whereas the performance difference was much smaller for statement identification (94% v 99%, respectively). Note, however, that 79% correct identification

in whispered speech is still well above chance level. Moreover, statements were correctly identified more often than questions, $F(1,71)=28.8$; $p<0.001$, and recognition of boundary tones was better in phonated than in whispered speech, $F(1,71)=27.2$; $p<0.001$.

3. Production study

As listeners hear the difference between whispered declarative questions and statements, though less clearly than in phonated speech, the question central to this section is whether one or more acoustic correlates of the boundary tone, other than F0, are present in whispered speech. On the basis of earlier research we expect to find possible correlates of pitch in the formants of whispered speech, e.g., [5][7][12], that have in many cases also been the only correlates investigated. In this analysis vowel duration, peak intensity and formant bandwidths were also included.

3.1. Analysis

Only the reiterant versions of the sentences were used for the acoustic analysis. The vowel in each final syllable, /ə/, is expected to show the signs of a boundary tone and therefore it was segmented and labelled in each of the reiterant sentences. For the boundary tone analysis 80 items were used (5 speakers × 4 accent patterns × 2 boundary tones × 2 speech modes).

Before analysis the sizes of the pitch ranges were compared between high and low boundary tones in phonated speech. A paired samples t-test showed an effect of Boundary Tone on the pitch range on the final syllable, $t(19)=-7.3$, $p<0.001$, showing that our speakers made a clear pitch difference between the L% and H% tones in phonated speech.

For both the phonated and the whispered sentence-final vowels duration, peak intensity, the first two formants and their bandwidths were measured using Praat [13]. The formants were determined at the moment of maximum intensity (analysis window = 25 ms, window shift = 10 ms). The Burg algorithm was used to find spectral envelopes. Finally, a discriminant analysis was run to evaluate the acoustic variables as predictors for the boundary tone.

3.2. Results

To compare the realizations of the boundary tones between phonated and whispered speech two-way ANOVAs with Speech Mode and Boundary Tone as fixed factors were run for each of the dependent variables (i.e. vowel duration, F1, F2, formant bandwidths, and peak intensity).

The duration of the sentence-final vowel showed a main effect of Speech Mode, $F(1,79)=19.3$; $p<0.001$, but not of boundary tone (see Table 1). The whispered phonemes were longer than the phonated ones, but H% tones did not systematically differ in length from L% tones.

Table 1. *Duration and peak intensity for the sentence-final vowels in phonated and whispered speech*

	Phonated		Whispered	
	L%	H%	L%	H%
Duration	86 ms	90 ms	108 ms	123 ms
Intensity	68 dB	79 dB	61 dB	67 dB

For peak intensity (see Table 1), an interaction of Speech Mode × Boundary Tone was present, $F(1,79)=5.9$; $p<0.05$. Moreover, main effects of both Speech Mode and Boundary Tone were found, $F(1,79)=108.2$; $p<0.001$ and $F(1,79)=88.7$; $p<0.001$, respectively. In both speech modes, the intensity of the rising boundary tone was higher than that of the falling one, but this effect was larger in phonated speech. Not surprisingly, the intensity in phonated speech was higher than in whispered speech. In whispered speech, the mean intensity of the falling boundary tone was significantly lower than that of the rising boundary tone, $t(38)=-4.7$, $p<0.05$.

The first formant showed main effects of both Speech Mode and Boundary Tone, $F(1,79)=344.9$; $p<0.001$ and $F(1,79)=4.9$; $p<0.05$ respectively. The mean frequency in phonated speech (339 Hz) was lower than in whispered speech (784 Hz). In both speech modes and across speakers, the mean F1 of the H% was higher than that of L%. This pattern was found for three out of five speakers, see Table 2.

The bandwidths of F1 showed main effects of both Speech Mode and Boundary Tone, $F(1,79)=14.2$; $p<0.001$ and $F(1,79)=4.2$; $p<0.05$, respectively. The bandwidths in phonated speech were smaller than in whispered speech. Furthermore, rising boundary tones were more sharply tuned than falling boundary tones.

Table 2. *Mean first and second formant frequencies in Hertz for each of the speakers.*

Speaker & formant	Phonated		Whispered	
	L%	H%	L%	H%
1 (m): F1	391	336	801	803
F2	1713	1624	1749	1750
2 (m): F1	301	400	629	852
F2	1831	1728	1790	2065
3 (m): F1	254	367	735	850
F2	1643	1504	1870	1771
4 (f): F1	254	293	864	913
F2	1901	1954	2047	1925
5 (f): F1	400	398	723	666
F2	1940	1748	1784	1956

As for the second formant, a Speech Mode × Boundary Tone interaction was found, $F(1,79)=4.3$; $p<0.05$. In whispered speech, the frequency of the rising boundary tone was higher than that of the falling boundary tone. In phonated speech however, it was just the other way around. Moreover, a main effect of Speech Mode was found, $F(1,79)=11.1$; $p=0.001$. Across speakers, the mean frequencies of both boundary tones were higher in whispered speech than in phonated speech.

Examination of the F2 bandwidth also showed a Speech Mode × Boundary Tone interaction, $F(1,79)=8.9$; $p<0.05$. In whispered speech, the bandwidth of F2 was smaller for the rising than for the falling boundary tone. In phonated speech however, it was the other way around. Furthermore, an effect of Speech Mode was found, $F(1,79)=5.4$; $p<0.05$. The mean of 283 Hz in whispered speech was smaller than the mean of 406 Hz in phonated speech. This difference was mainly due to the relatively large distance between the rising boundary tones. The F2 bandwidth of rising boundary tones in whispered speech was much smaller than the others. Within the whispered stimuli, a main effect of Boundary Tone was found, $t(38)=2.4$; $p<0.005$.

3.3. Linear discriminant analysis

Linear discriminant analysis (LDA) was run to build predictive models of tone categorisation based on the acoustic variables. When all acoustic variables were included, tone category was correctly predicted in 80% of the whispered cases, and in 97.5% of the phonates cases. These numbers roughly

correspond to the listeners' performance (see section 2.4). In both speech modes L% tones were predicted correctly more often than H% tones.

The mean percentages of correct predictions for each of the individual parameters are given in Table 3. In phonated speech, intensity was a good predictor of boundary tone, whereas the formants were weak predictors and duration was not useful. In whispered speech, each parameter resulted in classifications that were above chance level. Both F2 and especially duration led to better results on whispered than on phonated boundary tones. The acoustic analysis (see section 3.2) had not shown a significant difference in duration between whispered H% and L% tones, but the data showed that H% was longer than L% in 14 out of 20 cases. Intensity remained the best predictor, though less successful than in phonated speech.

Table 3. *Percentages of correctly predicted tones based on the individual acoustic parameters.*

	Correctly predicted tones	
Variable	Phonated	Whispered
Vowel duration	47.5%	60.0%
Peak Intensity	90.0%	70.0%
F1	62.5%	60.0%
F2	57.5%	62.5%

4. General discussion

The perception and production of boundary tones in whispered Dutch were examined. We asked whether listeners were able to perceive the difference between statements and questions as signalled by low and high boundary tones, respectively. We furthermore investigated which acoustic characteristics in whispered speech may convey the nature of the boundary tone in the absence of F0, through both acoustic analysis and discriminant analysis.

The results of the perception test showed that the identification of questions was more difficult in whispered than in phonated speech, but still well above chance level. This divergence did probably not result from intensity differences between the two speech modes, as similar changes for both boundary tones would have been expected. However, when listeners were not entirely sure whether a question boundary was heard, they may have made more "statement" choices, which were in accordance with the form of the sentence. Still, almost 79% of the questions was correctly identified as such, which we take as evidence that the rising boundary tone can be conveyed prosodically in whispered speech. This confirms our expectation based on earlier research into whispered tone perception, e.g., [3].

In section 3 the question was examined which correlates of the boundary tone are present in the acoustic signal of whispered speech. We examined vowel duration, the first two formants, and peak intensity. The F2 and its bandwidth showed interactions between speech mode and boundary tone, indicating that F2 behaves differently in whispered speech than in phonated speech. Hence, it may be used by speakers to compensate for the loss of pitch, as has been suggested before, see e.g., [12]. Moreover, LDA showed that the predictive capacity of some acoustic cues was higher in whispered than in phonated speech. However, as we have investigated boundary tones on one vowel only, we were not able to assess if the speaker's manner of compensation would change with the vowel, as is reported by [7]. A study using more variation in the sentence-final syllables might shed more light on this issue.

In addition to the possibility of compensatory cues in whispered speech, listeners may also use secondary cues to accents that remain available when phonation does not. Both intensity and F1 showed differences between high and low boundary tones in both speech modes. For Dutch listeners peak intensity is expected to be a weak cue, as they hardly use it for pitch accent perception [14], despite its suitability as a predictor according to LDA. Instead of peak intensity the distribution of energy across the spectrum, spectral tilt [14], is used by Dutch listeners as a secondary cue to pitch accents in phonated speech [15]. It has not been included in our current analyses, but some evidence for a role for this cue in whispered intonation can be found in Fig. 4.5 in [16].

This investigation has shown that listeners hear the difference between whispered statements and questions, and it has also identified a number of candidate acoustic cues that might be available to listeners for perception of boundary tones. The question of which of these cues is or are most important for listeners remains an open question.

5. References

[1] Van Heuven, V.J. and Haan, J. (2000). Phonetic correlates of statement versus question intonation in Dutch. In A. Botinis (ed.): *Intonation: Analysis, Modelling and Technology*. Dordrecht/Boston/London, Kluwer: 119-144.

[2] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory. In W.J. Hardcastle & A. Marchal (eds.): *Speech Production and Speech Modelling*, Dordrecht: Kluwer, 403-439.

[3] Higashikawa, M. Nakai, K., Sakakura, A. and Takahashi, H. (1996). Perceived pitch of whispered vowels – relationship with formant frequencies: a preliminary study. *Journal of Voice,* 10, 155-158

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

[5] Li, X-L. and Xu, B-L. (2005). Formant comparison between whispered and voiced vowels in Mandarin. *Acta Acustica*, 91, 1079-1085.

[6] Kong, Y-Y., and Zeng, F-G. (2006). Temporal and spectral cues in Mandarin tone recognition. *Journal of the Acoustical Society of America*, 120, 2830-2840.

[7] Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America,* 29, 180-182.

[8] Heeren, W. (2001). *Intonation in whispered Dutch: correlates of production and perception.* MA thesis, Leiden University.

[9] Gussenhoven, C. (2006). Transcription of Dutch intonation. In S-A. Jun (ed.): *Prosodic typology and transcription: a unified approach*, Oxford University Press, 118-145.

[10] 't Hart, J. Collier, R. and Cohen A. (1990). *A perceptual study of intonation.* Cambridge: Cambridge University Press

[11] Liberman, M.Y. and Streeter, L.A. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena. *Journal of the Acoustical Society of America*, 63, 231-233.

[12] Thomas, I.B. (1969). Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America*, 46: 468-470.

[13] Boersma, P. and Weenink, D. (2005). *Praat: doing phonetics by computer (Version 5.1)* [Computer program].

[14] Sluijter, A.M.C. and Van Heuven, V.J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471-2485.

[15] Sluijter, A.M.C., Van Heuven, V.J. and Pacilly, J.J.A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America* 101, 503-513.

[16] Van Rossum, M. (2005). *Prosody in alaryngeal speech*. PhD thesis, Utrecht University, LOT dissertation series 108, Utrecht: LOT.