The interaction of accent and boundary tone in perception of whispered speech

Willemijn Heeren[1] and Vincent J. van Heuven[1,2]

[1]Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition
Leiden University, The Netherlands
[2]Department of Applied Linguistics, University of Pannonia, Veszprém, Hungary
{w.f.l.heeren,v.j.j.p.van.heuven}@hum.leidenuniv.nl

Abstract

We investigated how the perception of Dutch whispered boundary tones depends on the presence of an accent in the utterance-final word, i.e. the boundary tone landing site. Listeners performed near ceiling in normal speech, whereas the same listeners' performance dropped about 30% in whisper, while processing speed decreased in whisper compared to normal speech. Accent position furthermore influenced boundary tone perception. Initial-stress words showed a question bias that affected recognition of that speech act when accent and boundary tone did not coincide. On final-stress words, in which boundary tone and accent coincided, statements and questions were identified equally well.

Index Terms: boundary tone, nuclear accent, perception, whispered speech

1. Introduction

In whispered speech – where voicing and therefore a fundamental frequency (f0) are absent – listeners still perceive, albeit less reliably than in normal speech, prosodic differences that normally heavily depend on f0 presence. For instance, in whisper listeners recognize questions and statements expressed by different boundary tones (H% versus L%) in cases where prosody, rather than lexico-syntax, codes the crucial information [1-3]. Listeners also discriminate intended pitch height [4], differentiate emotional from neutral speech [5], and identify lexical tones [e.g. 6-8]. Many of these studies, however, assessed perception in single syllables, rather than in multi-syllabic or multi-word phrases, whereas the latter would be more ecologically valid. Multi-syllabic utterances will display some form of ranking as to the relative prominence of those syllables (e.g. imposed by lexical stress).

Though earlier work may indicate that intonation in whisper is perceptible, it does not provide much evidence on the perception of whispered intonation in more complicated linguistic structures. In the present study, we investigated how the perception of Dutch whispered boundary tones depends on characteristics of the tone-bearing word, by using disyllabic minimal stress pairs as boundary tone landing sites. In the case that lexical stress, realized as a nuclear accent, lands in final position, the two tonal events fall on the same syllable. In the case that lexical stress lands in initial position, the tonal events fall on adjacent syllables.

For Dutch, as found in studies on normal speech, the most reliable acoustic correlate of lexical stress in sentence context is relative syllable duration [9, 10]. Perceptually, duration also is a reliable cues to stress [11], but for the perception of prominence, f0 is taken to be the primary cue in Dutch [12], as well as in English [13]. In the absence of f0, expressing intonational contrasts in whispered speech seems to be necessarily more intertwined with segmental characteristics than in the case of normal speech. For instance, in whisper formants are not only used to code vowel identity, but also seem to contribute to expressing differences in height [e.g. 14, 15]. Moreover, if different intonational events land on the same syllable, the restricted resources in whisper may be burdened even further.

To our knowledge, one earlier study has addressed the interaction of accent and boundary tones in whispered speech [1], but in a descriptive manner only. In that investigation, Hungarian listeners classified disyllabic minimal stress pairs that were produced either as question or as statement into one of four categories: two lexical stress positions by two boundary tones. A confusion matrix of classification responses showed that boundary tones were identified above chance level, that declaratives were identified correctly more often than interrogatives, and that accent positions were confused less than boundary tones. In addition, there was confusion across accent positions and boundary tones. For instance, ten percent of final-stress declaratives were identified as initial-stress interrogatives, and such across-tonal event responses seem to support the claim that, in whisper, accents and boundary tones may interact in perception. The same type of utterances was classified without errors in normal speech.

To better understand prosody perception in whispered speech communication, the interaction of accent position and boundary tone perception in Dutch was studied in whispered compared to normal speech. A within-subjects design was used that also included reaction time measurements. We predict that listeners perform better when tonal events do not coincide on the same syllable.

2. Method

Perception of the speech act, i.e. interrogative versus declarative, as expressed through the boundary tone (H% versus L%, respectively) was determined in a classification task with reaction time measurements. Boundary tones were produced on disyllabic nouns with lexical stress, realized as a nuclear accent, in either initial or final position. In the latter case, boundary tone and lexical stress coincide on the same syllable (prosodic clash); in the former case lexical stress falls on the syllable preceding the one carrying the boundary tone (no clash). Minimal stress pairs were used, so that segmental structure would be comparable. To verify that the boundary tone does not alter its bearer's interpretation, perception of the items' stress positions was measured using the same task.

2.1. Materials

Four Dutch minimal stress pairs were used: (1) '*ca·non/ka*'*non*, /kanɔn/, 'canon/cannon', (2) '*Ser·visch/ser*'*vies*, /sɛrvis/, 'Serbian/crockery set', (3) '*Pla·to/pla*'*teau*, /plato/, 'Plato; plateau', and (4) '*voor·naam/voor*'*naam*, /vornam/, 'first name/dignified'. Target words were recorded in a neutral

10.21437/SpeechProsody.2014-66

carrier sentence *Hij zei…* 'He said…', which orthographically ended in either a full stop (to elicit L%) or a question mark (to elicit H%), and which forced the nuclear accent onto the target word, thus establishing the prosodic crowding contrast.

Twelve (self-reported) normal-hearing, Dutch native speakers (6 female) participated in 20-minute recording sessions (informed consent was obtained), and were paid a small amount for their efforts. For each speaker, a different listener was present to judge the recordings. This speaker-listener set-up was intended to prompt the speaker to use listener-directed rather than read speech.

Speakers received written instructions, and completed a short practice session, using different minimal pairs than during the actual recording, for both normal and whispered speech. The order of the speech modes was counterbalanced across speakers. Recordings were made using an Edirol R-44 portable recorder and Røde NTG-2 condenser microphone with 'dead cat' windscreen at 44.1 kHz, 24 bits in a sound-treated booth in the phonetics laboratory of Leiden University. Affirmative and interrogative targets were presented to the speaker one by one and in written form on a computer screen, in a pseudo-random order. The listener was seated outside the booth in a silent classroom, wearing Sennheiser HD 414 SL headphones, and used a keyboard to classify each of the speaker's utterances as affirmative or interrogative. Before the next target was presented, the speaker got feedback about the listener's understanding of the previous one. By keeping the listener outside the booth, and invisible to the speaker, the only cues the speaker could provide were auditory. Two repetitions per utterance were recorded and saved as separate wave files, resulting in 32 files per speaker.

The listeners who were present during the recordings labeled the boundary tones in normal speech correctly in 94% of the cases and in whisper, in 68% of the cases. All speaker-listener pairs were different. Using a 6-point Likert scale (1 = very difficult, 6 = very easy) speakers rated the difficulty of their task for both speech modes. According to a Wilcoxon signed ranks test for paired samples, the task was judged more difficult in whisper (median=3.0) than in normal speech (median = 4.5), $Z = -2.5$, $p = .013$. In neither speech mode was the task judged as particularly easy.

Per lexical item, one instance was annotated manually, and that annotation was used to automatically annotate all other instances of the same item using a dynamic time warping procedure in PRAAT [16]. These annotations were manually checked, and corrected if necessary. Target words were cut from the carrier sentences, and intensity was normalized by setting recordings within a speaker and speech mode to 60 dB (rms = 0.020), which corresponded to the minimum intensity of whispered items after scaling peaks to the maximum intensity range (using PRAAT's 'Scale peaks…'). There were 192 stimuli per speech mode: 8 items (4 initial, 4 final stress) × 2 speech acts (question, statement) × 12 speakers.

2.2. Participants and procedure

Twenty-four, right-handed Dutch native listeners (17 females), aged 19-57 (mean = 22 years), were hearing-screened to have normal hearing at octave frequencies between 0.125 and 8 kHz (informed consent obtained). Each of the 192 items per speech mode was presented once to each listener in a blocked design over tasks. Half of the subjects heard the first half of the materials in the Speech Act classification task, and the second half in the Lexical Stress Position classification task. The other half of the subjects listened to the complementary stimulus sets in each task. The set of materials was halved by including only one boundary tone realization, either H% or L%, per speaker and per target word in each half. Subjects received a small fee for participation in the 45-minute session.

Subjects were seated in a sound-treated booth wearing Sennheiser HD 414 SL headphones. After general instructions in written form, more detailed instructions were presented on a computer screen. Response options were shown on screen, while listeners were asked to press one of two response buttons on a keyboard using their index fingers. During speech act classification listeners indicated whether the target sounded like a question or a statement. During lexical stress position classification listeners indicated whether the initial or the final syllable was more prominent. Both tasks were presented once with normal speech materials, and once with whispered speech, resulting in four subsequent tests. Speech modes, response keys and task orders were counterbalanced across subjects. To allow for within-subjects analyses including the factor speech mode (normal vs. whisper), corresponding whispered and normal speech items from the same speaker were presented to the same subject in the same task.

3. Analysis and results

Percent correct responses was computed for both subtasks, i.e. boundary tone (BT) and lexical stress (LS) classification, and transformed to rationalized arcsine units (RAU) [17]. Reaction times (RTs) were measured from target word onset. RTs under 500 ms and over two standard deviations beyond the mean, computed per listener-per speech mode, were excluded (BT: 2.4% of the data; LS: 4.2% of the data). RTs were transformed to their inverse (1/RT) for analysis. Both RAU scores and inverse RTs were subjected to repeated measures ANOVAs with within-subjects factors Speech Mode (normal, whisper), Speech Act (interrogative, declarative), Lexical Stress Position (initial, final) and Minimal Pair (4). If sphericity was violated, Huynh-Feldt correction was applied.

3.1. Boundary tone classification

Significant effects are presented in Table 1. Figure 1a shows that in whisper, boundary tone classification was significantly poorer than in normal speech (61 vs. 94%, respectively), but above chance level [binomial test: $N = 2304$, $p = ½$, $Z = 11.1$, $p < .001$]. Across speech modes, declaratives were classified correctly more often than interrogatives. The interaction of speech mode by speech act showed that in whisper, the difference in correct responses to declaratives versus interrogatives was larger than in normal speech (whisper: 71 vs. 52%; normal speech: 96 vs. 93%, respectively).

Across speech modes, stimuli with final stress yielded similar scores for the two speech acts, but stimuli with initial stress received more correct responses on declaratives than interrogatives. This interaction was found within both speech modes [normal speech: $F(1,23) = 53.2$, $p < .001$; whisper: $F(1,23) = 31.7$, $p < .001$]. Absolute differences were largest in whisper (see Fig. 1a), where for words with initial stress, declaratives were classified correctly well above chance level at 80%, whereas interrogatives were classified below chance level at 41% [$N = 576$, $p = ½$, $Z = -4.2$, $p < .001$]. On final-stress words, scores were 62.5 and 62.3%, respectively. Trends were comparable between minimal stress pairs, and followed the two main effects. Only for the *Plato/plateau* pair, did

speech mode and speech act interact, showing a larger performance difference between the interrogatives and declaratives across speech modes.

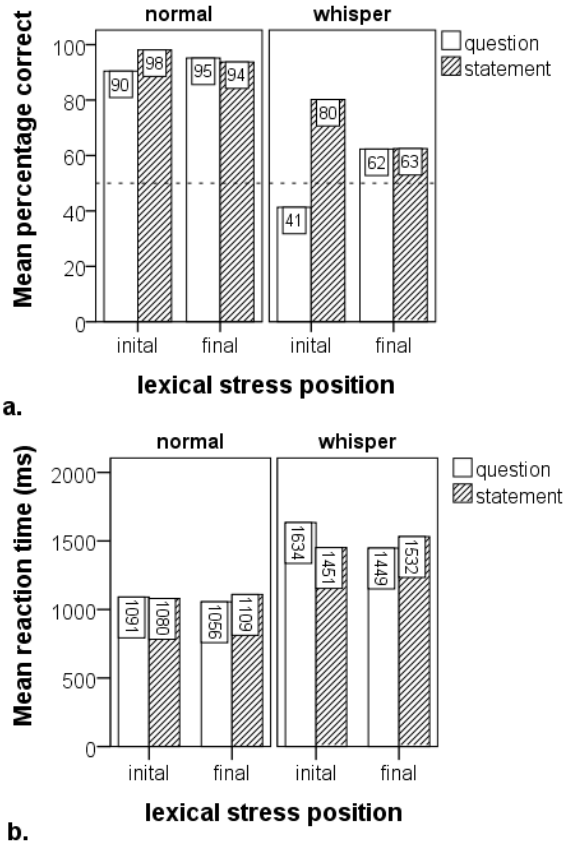

Figure 1: *BT task results, per speech mode, speech act and stress position. a) Mean percentage correct (chance level = 50%). b) Mean reaction time (ms).*

Reaction time results (see Fig. 1b) showed that listeners were faster in normal speech than in whispered speech (1084 and 1502 ms, respectively), faster on words with final stress than with initial stress (1244 and 1255 ms, respectively), and RTs also varied with minimal pair. The effect of stress position, following the main effect, was only significant within whispered speech [$F(1,18) = 7.9$, $p = .012$], as reflected by a marginally significant speech mode by stress position interaction [$F(1,18) = 4.4$, $p = .051$].

Across speech modes, declaratives were responded to equally fast in words with initial and final stress, but interrogatives were responded to faster in final-stress words. This speech act by stress position interaction was found in similar ways in both speech modes [whisper: $F(1,18) = 17.6$, $p = .001$, normal speech: $F(1,23) = 16.5$, $p < .001$]. Across minimal pairs, RTs were longer in whispered than in normal speech, but not exactly to the same extent. Across speech modes, responses to the different minimal pairs followed the main effect of faster responses to words with final stress, but for the *voornaam* pair the trend was in the opposite direction with faster responses to initial-stressed words. The four-way interaction showed that the differences in response times by lexical stress position were mainly caused by differences measured in whisper.

Table 1. *Significant effects and interactions of the RM ANOVAs on BT classification and reaction time data.*

Classification		
Speech mode	$F(1,23) = 403.5$	$p < .001$
Speech act	$F(1,23) = 26.3$	$p < .001$
Minimal pair	$F(3,69) = 3.7$	$p = .016$
Speech mode × speech act	$F(1,23) = 20.2$	$p < .001$
Speech act × stress position	$F(1,23) = 47.8$	$p < .001$
Sp. mode × sp. act × stress pos.	$F(1,23) = 14.8$	$p = .001$
Sp. mode × sp. act × min. pair	$F(3,69) = 5.4$	$p = .002$
Reaction times		
Speech mode	$F(1,18) = 129.1$	$p < .001$
Stress position	$F(1,18) = 6.9$	$p = .017$
Minimal pair	$F(3,54) = 8.9$	$p < .001$
Speech act × stress position	$F(1,18) = 29.6$	$p < .001$
Speech mode × minimal pair	$F(3,54) = 3.6$	$p = .020$
Stress position × minimal pair	$F(3,54) = 3.1$	$p = .033$
Sp. mode × sp. act × min. pair	$F(3,54) = 3.0$	$p = .039$
Sp. mode × sp. act × stress pos.× min. pair	$F(3,54) = 3.0$	$p = .041$

3.2. Lexical stress position classification

Table 2 lists the significant effects for both lexical stress position classification and reaction times. The absence of a speech mode main effect shows that identification of lexical stress position in whisper went as well as in normal speech (89 and 91%, respectively). Across speech modes, there was some variation in mean classification scores per minimal pair, but this difference remained under 4% between the lowest and highest mean scores per pair. Words pronounced as declaratives received more correct responses for initial stress, whereas the correctness of responses to words pronounced as interrogatives was comparable for both stress positions. This trend was observed in both speech modes, but the speech act by stress position interaction was only significant in whisper [$F(1,23) = 18.0$, $p < .001$], not normal speech ($p = .069$).

Table 2. *Significant effects and interactions of the RM ANOVAs on LS classification and reaction time data.*

Classification		
Minimal pair	$F(3,69) = 3.6$	$p = .018$
Speech act × stress position	$F(1,23) = 12.7$	$p = .002$
Reaction times		
Speech act	$F(1,23) = 31.1$	$p < .001$
Stress position	$F(1,23) = 14.4$	$p = .001$
Minimal pair	$F(3,69) = 4.7$	$p = .005$
Speech mode × stress position	$F(1,23) = 19.0$	$p < .001$
Speech act × stress position	$F(1,23) = 11.1$	$p = .003$
Speech act × minimal pair	$F(3,69) = 4.8$	$p = .004$
Sp. act × stress pos.× min. pair	$F(3,69) = 4.0$	$p = .011$

The absence of a speech mode main effect in the RTs indicated that listeners were as fast at identifying lexical stress position in whispered as in normal speech (1304 and 1336 ms, respectively). Across speech modes, responses were faster to declaratives (1279 ms) than to interrogatives (1344 ms), especially for words with initial stress. Responses were faster to final-stress (1289 ms) than to initial-stress (1352 ms) words, but the latter effect did not assume significance for normal speech ($p = .318$), only for whisper [$F(1,23) = 31.2$, $p < .001$].

There was variation in the response times to different words, with fastest responses to *ka'non* (1279 ms) and slowest responses to '*servisch* (1399 ms). The trend for responses to declaratives to be faster than to interrogatives was present in all minimal pairs, but the size of the difference varied between them. Finally, the speech act by stress position interaction was observed in three out of four minimal pairs; for *Plato/plateau*, however, responses to either speech act were equally fast.

4. Discussion

It was expected that in whisper – given the more restricted means of conveying intonation – listeners would have more difficulty correctly identifying boundary tones, and especially when coinciding with lexical stress position. Though the general performance decrease was obtained as expected, the main effect of stress position was not found. Reaction times reflected that processing of final-stress words was in fact somewhat faster, which may hint at easier processing; this effect became significant for whispered stimuli only.

Listener performance showed comparable means around 60% correct on whispered words with initial and final stress; across stress positions, performance was better on declaratives than on interrogatives. This in general suggests that cues to interrogativity were less clear in whisper, which was also found in [1]. But as Fig. 1 shows, performance varied with the stimulus' stress position, especially in whisper. On whispered words with initial stress, where boundary tone and accent do not coincide, performance was much better on declaratives than on interrogatives. Effectively, at 40% correct, questions were not recognized as such on words with initial stress. Moreover, RTs were generally longer for this type of stimulus. For whispered words with final stress, performance was comparable between the speech acts. These results suggest that only on whispered stimuli in which accent and boundary tone coincided on the same (i.e. final) syllable (clash condition), were listeners able to reliably identify the boundary tone.

As performance on interrogatives pronounced on initial-stress words was, in fact, below chance level, we looked for potential response biases in the data. Chi square analyses per speech mode per lexical stress position revealed that in both speech modes, listeners gave a majority of 'interrogative' responses to words with initial stress, whereas equal numbers of either response category were expected (normal speech: 543 out of 1008 'interrogative' responses, $\chi^2(1) = 6.0$, $p = .014$; whisper: 711 out of 1008 'interrogative' responses, $\chi^2(1) = 170$, $p < .001$). In normal conversational speech, statements occur (much) more often than questions [e.g. 18]. As speech perception generally reflects differences in the token frequencies of categories, we expect listeners to respond with the statement category unless there is clear evidence to the contrary. This was not what listeners did. Possibly, listeners interpreted the accent in initial position as prominence in a more general sense that was then associated with an interrogative reading of the utterance as a whole. Alternatively, first-syllable prominence may have been interpreted as a direct cue to a potentially upcoming question, which was not overruled by evidence provided in the relatively weak final syllable. [19] showed that the size of an object accent influenced listener expectations about whether an utterance was a statement or a question: larger object accents triggered stronger question expectation. In the present study, the accent on the first syllable may have similarly signaled a potentially upcoming question, especially when f0 was absent.

For words with lexical stress on the same (i.e. final) syllable as the boundary tone, whispering speakers were able to convey the speech act. In comparison with other studies on boundary tone identification in whisper [2, 3], the task seems to have been relatively more difficult in the present study. The difference may be due to higher demands placed on processing by the two intonational events in close proximity. On the one hand, this is taken to reflect that more complicated linguistic structures, as in the present study, may moderate earlier results on the processing of prosody in whisper (see also [1]). On the other hand, the exclusive use of minimal pairs in the present study may have made listeners aware of the lexical contrast in addition to the speech act difference, also during the boundary tone task, which in turn may have influenced performance.

There was a large difference in the mean reaction times to whispered versus phonated boundary tones. This cannot be explained by the difference in stimulus duration between the speech modes, as this difference was only on the order of 100 ms (693 vs. 583 ms means for whispered and normal speech items, respectively), whereas the reaction time difference was on the order of 400 ms. The slower responses in whisper therefore seem to mainly reflect an increase in processing time due to a difference in cues to boundary tones between the speech modes, including the absence/presence of f0.

Over the same set of stimuli, listeners classified lexical stress position with high accuracy and with similar reaction times in the two speech modes. These results are consistent with the finding that lexical stress position is most reliably realized by durational differences [10], which also form a main cue for listeners [11]. A planned acoustic analysis of the data is expected to reflect the presence of durational information. Moreover, we take the high listener scores to indicate that lexical meaning was generally not influenced by boundary tone realization, in either speech mode.

Words with initial stress pronounced as declaratives were more often identified correctly with respect to stress position than their counterparts with final stress, whereas no difference was found when the same words had been pronounced as interrogatives. Mainly in whisper, responses to interrogatives were around 60 ms faster on words with final stress than on words with initial stress, whereas the average durational difference between the words types was very small (~10 ms). This hints at a small processing benefit for the former type of words, which may be explained by a smaller demand on short term memory for items with final stress.

In sum, depending on the listening task, the same stimuli were responded to very differently. Whispered speech was as clear as normal speech with respect to lexical stress position. For boundary tone perception, however, listeners performed near ceiling in normal speech, whereas the same listeners' performance dropped about 30% in whisper, while processing speed decreased significantly. Accent position furthermore influenced boundary tone perception. Initial-stress words showed a question bias that affected recognition of that speech act. On final-stress words, in which boundary tone and accent coincided, the speech acts were identified comparably.

5. Acknowledgements

This work was supported by a VENI grant made available to the first author by the Netherlands Organisation for Scientific Research (NWO).

6. References

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

[2] Heeren, W. F. L. and Van Heuven, V. J., "Perception and production of boundary tones in whispered Dutch", in Proc. Interspeech 2009, Brighton, 2411-2414, 2009.

[3] Heeren, W. F. L. and Lorenzi, C., "Perception of prosody in whispered French", J. Acoust. Soc. Am., to appear.

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

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

[6] Abramson, A. S., "Tonal experiments with whispered Thai", in A. Valdman [ed.], Papers on linguistics and phonetics to the memory of Pierre Delattre, 29–44, The Hague: Mouton, 1972.

[7] Miller, J. D., "Word tone recognition in Vietnamese whispered speech", Word, 17:11-15, 1961.

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

[9] Nooteboom, S.G., "Production and Perception of Vowel Duration. A Study of durational Properties of Vowels in Dutch", Unpublished Doctor's Thesis, Utrecht: University of Utrecht, 1972.

[10] Sluijter, A. and Van Heuven, V. J., "Spectral balance as an acoustic correlate of linguistic stress", J. Acoust. Soc. Am., 100:2471-2485, 1996.

[11] Sluijter, A. M., van Heuven, V. J. and Pacilly, J. J., "Spectral balance as a cue in the perception of linguistic stress", J. Acoust. Soc. Am., 101:503-513, 1997.

[12] Van Katwijk, A., "Accentuation in Dutch", Amsterdam/Assen: Van Gorcum, 1974.

[13] Fry, D. B., "Experiments in the perception of stress", Lang. Speech, 1:126-152, 1958.

[14] 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.

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

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

[17] Studebaker, G. A., "A "Rationalized" Arcsine Transform", J. Speech Hear. Res. 28:455-462, 1985.

[18] Van Heuven, V. J., Haan, J. and Pacilly, J. J., "Global and local characteristics of Dutch questions in play-acted and spontaneous speech", in Proc. ESCA workshop on sound patterns of spontaneous speech, La Baume-les-Aix, 139-142, 1998.

[19] Van Heuven, V. J. and Haan, J., "Temporal development of interrogativity cues in Dutch", in C. Gussenhoven and N. Warner [Eds.], Laboratory Phonology 7, 61-86, Berlin: Mouton de Gruyter, 2002.