The interaction of lexical and phrasal prosody in whispered speech

W. F. L. Heeren and V. J. van Heuven
Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition, Leiden University, Cleveringaplaats 1, 2311 BD Leiden, The Netherlands

(Received 2 May 2014; revised 17 October 2014; accepted 23 October 2014)

The production and perception of Dutch whispered boundary tones, i.e., phrasal prosody, was investigated as a function of characteristics of the tone-bearing word, i.e., lexical prosody. More specifically, the disyllabic tone-bearing word also carried a pitch accent, either on the same syllable as the boundary tone (clash condition), or on the directly adjacent syllable (no clash condition). In a statement/question classification task listeners showed moderate, but above-chance performance for both conditions in whisper, which, however, was much worse as well as slower than in normal speech. The syllabic rhymes of speakers’ productions were investigated for acoustic correlates of boundary tones. Results showed mainly secondary cues to intonation, that is, cues that are present in whisper as in normal speech, but minimal compensatory cues, which would reflect speakers’ efforts to enhance their whispered speech signal in some way. This suggests that multiple prosodic events in close proximity are challenging to perceive and produce in whispered speech. A moderate increase in classification performance was found when that acoustic cue was enhanced that whispering speakers seemed to employ in a compensatory way: changing the spectral tilt of the utterance-final syllable improved perception of especially the poorer speakers and of intonation on stressed syllables. © 2014 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4901705]

PACS number(s): 43.70.Fq, 43.71.Sy [SAF]

Pages: 3272–3289

I. 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 in cases where prosody, rather than lexico-syntax, codes the crucial information (Fónagy, 1969; Heeren and Van Heuven, 2009; Heeren and Lorenzi, 2014). Listeners also discriminate intended pitch height (Higashikawa et al., 1996), differentiate emotional from neutral speech (Tartter and Braun, 1994), and identify lexical tones (e.g., Miller, 1961; Abramson, 1972; Liu and Samuel, 2004). Many of these studies, however, assessed perception in single syllables, rather than in poly-syllabic or multi-word phrases. The latter will display some form of ranking as to the relative prominence of those syllables (e.g., imposed by lexical stress). Therefore, though earlier work may indicate that intonation in whisper is perceptible, it does not provide much evidence for perception of whispered intonation in the linguistic structures that are normally used during speech communication. In those structures, multiple layers of information may interact through prosody. We therefore investigated the perception and production of whispered utterances in which both lexical and phrasal information were expressed prosodically and in close proximity.

Whisper is a speech mode that may yield insight into natural ways of coding pitch in speech through other information than periodicity cues, that is cues related to the speaker’s F0 and harmonics (cf. Moore and Gockel, 2011). Such cues might for instance contribute to pitch coding for hearing-impaired listeners. In whisper, the vocal folds do not vibrate quasi-periodically as they do in normal speech, but remain partially opened allowing air to pass through thus creating a turbulent air flow (Eckert and Laver, 1994). Larynx movements (which often correspond with differences in vowel height, i.e., differences in F1) related to pitch changes may be comparable between whispered and normal speech (Coleman et al., 2004), and in whisper spectral cues associated with the vocal tract are assumed to be comparable to those in normal speech (e.g., Kong and Zeng, 2006; Freyman et al., 2013)—with the exception that formants tend to have higher center frequencies overall (e.g., Kallail and Emanuel, 1984a,b). It has repeatedly been demonstrated, and for a number of different languages, that intonation alone may code the distinction between statements and questions in whispered utterances (Hungarian: Fónagy, 1969; Dutch: Heeren and Van Heuven, 2009; French; Heeren and Lorenzi, 2014). In this study, the goal was to understand how this coding of phrasal intonation in whisper interacts with properties of its landing site, that is how intonation in whisper is produced and perceived in a case of “tonal clash.” For normal speech it has been found that tonal clash or “tonal crowding” environments result in changes of tonal alignment as well as tonal scaling. For instance, in Greek the phonetic implementation of intonational movements in polar questions is influenced by the position of the focused word in the utterance.
and the timing of boundary tones in Greek depends on the presence of stress in the final syllable (Katsika et al., 2014).

First, 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. Boundary tones align to the edges of intonational phrases, in this case to the trailing edge. A high boundary tone (annotated as H%) after a preceding H target is a sufficient (but not necessary) cue for a question (Van Heuven and Kirsnear, 2004), whereas a low boundary tone (annotated as L%) typically signals a statement (cf. Gussenhoven, 2005, for an autosegmental-metrical description of Dutch). When lexical stress, realized as a nuclear accent, lands in the final position, the two tonal events fall on the same syllable. In the case where lexical stress lands in the initial position, the tonal events fall on adjacent syllables. Although the amount of intonational events distributed over the lexical and phrasal levels does not differ between the two word conditions, we argue that the closeness of events differs between the two conditions, which we expect to yield an effect on the production (as in normal speech, e.g., Arvaniti et al., 2006) and perception of whisper. A difference in perception performance as a function of context, but not prosodic context per se, was reported by Gao (2002); listeners were asked to recognize whispered Mandarin lexical tones on CV syllables and they performed best when these were produced in isolation, and worst when they occurred at the beginning of sentences, with scores on sentence-final and sentence-medial positions in between.

Second, we assessed which acoustic correlates carry prosodic information in the case of whispered boundary tones that co-occur with nuclear accents. The question what acoustic correlates may carry this information in whispered speech, has received relatively little attention. Moreover, as some of these studies were done several decades ago, most evidence is qualitative rather than quantitative. Earlier work suggests that mainly the first and second formants (F1 and F2) carry prosodic information in the absence of F0. These formants were found to be higher when the pitch was intended as higher (Higashikawa et al., 1996; Konno et al., 2006; Heeren and van Heuven, 2009), listeners related differences in the lower two formants to changes in pitch (Higashikawa and Minifie, 1999), and musically trained listeners indicated F2s as corresponding to the intrinsic pitch of whispered vowels (Thomas, 1969). Other correlates that have been suggested to contribute are changes in intensity (Meyer-Eppler, 1957; Fónagy, 1969; Heeren and Van Heuven, 2009), and changes in higher formants, such as F3 (Meyer-Eppler, 1957; Denes, 1959; Fónagy, 1969). Meyer-Eppler (1957) analyzed spectrograms of German vowels that were sung without voice at different tones and in addition to formant changes, such as an increasing F3 for /a/ with higher tones, he observed a higher intensity with higher tones. The effect of F3 change with increasing height was replicated by Denes (1959), and also observed by Fónagy (1969) in a study of question versus statement intonation in whispered Hungarian. Expression of questions in that study was furthermore observed to co-occur with an intensity increase on the syllable carrying the intonational target movement. In the case of lexical tone identification, the secondary cue of duration was also found to be informative (Liu and Samuel, 2004). In the perception of whispered Mandarin tones, listeners on average performed well above chance level, and especially on tone 3, which is characterized by a longer duration than the other tones (Howie, 1976, p. 220), also in whisper (Chang and Yao, 2007). This perceptual advantage for tone 3, in different contexts, was also reported by Gao (2002).

More recently, Heeren and Lorenzi (2014) found evidence that intonation in whisper may be interpreted from spectral cues that are more distributed than single formants. None of these studies, however, have specifically looked at how the interaction of multiple intonational events in close proximity influences the realization and perception of prosody in whispered speech.

For Dutch, as found in studies on normal speech, the most reliable acoustic correlate of lexical stress in sentence context is relative syllable duration (Nooteeboom, 1972; Sluijter and van Heuven, 1996). Perceptually, duration also is a reliable cue to stress (Sluijter et al., 1997), but for the perception of prominence, F0 is taken to be the primary cue in Dutch (Van Katwijk, 1974, p. 87), as well as in English (Fry, 1958; Beckman, 1986, p. 195). As in a number of other languages, low versus high boundary tones may cue statements versus questions in Dutch when lexico-syntactic cues such as word order and question words are absent (Van Heuven and Haan, 2002). In addition, not only the boundary tone, but also sentence declination may differ between questions and statements (Thorsen, 1980; Van Heuven and Haan, 2000). The availability of prosodic cues to the sentence’s function of statement versus question, however, varies with the presence of other types of cues; the more lexico-syntactic indicators of sentence function, the less intonational cues are present (Van Heuven and Haan, 2000).

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 not only code vowel identity, but also seem to contribute to expressing differences in height (e.g., Meyer-Eppler, 1957; Higashikawa and Minifie, 1999). Moreover, if different intonational events land in close proximity, the restricted resources in whisper may be burdened even further relative to normal speech. To our knowledge, one earlier study has addressed the interaction of accent and question intonation in whispered speech (Fónagy, 1969), but in a descriptive manner only. In that investigation, listeners classified Hungarian (quasi-)minimal stress pairs consisting of a disyllabic word (e.g., apad “ebb”) or two monosyllabic words (a pad “the bench”) that were produced either as question or as statement into one of four categories: two stress positions by two sentence functions. In Hungarian, a fixed initial stress language, the distinction between statements and yes/no questions is made through intonation alone, but not in the same way as in, e.g., Dutch and English. Acoustic characteristics of the pitch movement in penultimate position are of importance (Gósy and Terken, 1994); although both statement and question in Hungarian end with low pitch, the pitch remains
high until a later point in the utterance for a question than for a statement. Classification responses were above chance level, statements were identified correctly more often than questions, and accent positions were confused less than questions and statements. In addition, there was confusion across accent positions and sentence functions. For instance, 10% of final-stress statements were identified as initial-stress questions, and such across-tonal event responses seem to support the claim that, in whisper, accents and question intonation may interact in perception. The same type of utterances were classified without errors in normal speech.

In addition to studying what acoustic correlates may carry prosodic information in whisper, we are interested in the nature of these correlates: are they secondary or compensatory relative to the cues found in normal speech? Secondary cues are defined as those cues that are already present in normal, phonated speech, and that are maintained in whisper. Compensatory cues are defined as those cues that are available only or to a larger extent in whispered speech. The secondary cue hypothesis assumes that speech is a redundant signal. Early support for this hypothesis was provided in experiments on the perception of “vocabder whisper,” i.e., vocoded normal speech with the periodic excitation signal replaced by a noise source in the re-synthesis, in which lexical tone remained identifiable (Denes, 1959; Abramson, 1972). Acoustic studies have found some evidence that such cues may be intensity (Meyer-Eppler, 1957; Fonagy, 1969; Heeren and Van Heuven, 2009), F1 (Heeren and Van Heuven, 2009), and duration (Liu and Samuel, 2004). If pitch perception in whisper is coded through compensatory cues, this would be in line with the idea that speakers try to match their listeners’ needs, and put in more effort when needed, as in the case of hyperspeech (Lindblom, 1990) or clear speech (e.g., Picheny et al., 1986; Krause and Braida, 2004). For whispering speakers, the lack of F0 might be considered an incentive to put in more effort. Some evidence for this hypothesis was found in Heeren and Van Heuven (2009), who reported interactions between speech mode and intonation condition for F2; it was higher in whispered high boundary tones than in low boundary tones; the pattern reversed in phonated speech. In the present study, whispered speech acoustic measures were systematically compared with measurements on normal speech in order to assess this question of cue type: secondary or compensatory.

In Sec. II the question is addressed how the perception of Dutch whispered boundary tones depends on characteristics of the tone-bearing word. This is followed up in Sec. IV, where the perceptual effect of cue enhancement is investigated, informed by the results from the acoustic analysis (Sec. III).

II. EXPERIMENT I: PERCEPTION OF WHISPERED BOUNDARY TONES

Listeners’ perception of the sentence’s function, i.e., question or statement, 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 nuclear accent, in either initial or final position.

In the latter case, boundary tone and lexical stress coincide on the same syllable (“prosodic clash,” also referred to as tonal clash or crowding, e.g., Arvaniti et al., 2006); in the former case lexical stress falls on the syllable preceding the one carrying the boundary tone (no clash). The combination of a nuclear accent and boundary tone is also referred to as a nuclear contour (Gussenhoven, 2004, p. 296). Minimal stress pairs were used so that segmental structure would be comparable between the clash and no-clash versions. To verify that the boundary tone does not alter its bearer’s interpretation, perception of the lexical items’ stress positions was also measured, using the same task. We predict that in whisper—given the more restricted means for speakers to convey intonation—listeners will have more difficulty correctly identifying a boundary tone than in normal speech, and especially when the boundary tone falls on a lexically stressed syllable.

A. Method

1. Stimulus materials

Four Dutch lexical stress minimal pairs were selected that were disyllabic, with the same syllabification per member of a pair, and containing final-syllable vowels that together approached the corners of the vowel space: ‘ca-non’/’kan/on/,” ‘Ser-vi-shis’/’ser-vi’es’ “Serbian/clock-ery set” (/sérvis)/, ‘Pla-to/’plá-teau “Plato/plateau” (/pláto/), and ‘voor-naam’/’voor’naam “first name/respectable” (/vornam/). These target words were recorded in neutral carrier sentences of the type “He said…” (in Dutch: “Hij zei…”), that orthographically either ended in a full stop (to elicit a low boundary tone L%) or in a question mark (to elicit a high boundary tone H%), and that forced the sentence’s nuclear accent onto the target word in final position, thus establishing the prosodic crowding case in final stress words.

Twelve (self-reported) normal-hearing native speakers of Dutch (six male) participated in 20-min recordings sessions. Recordings were made using an Edirol R-44 portable recorder and Røde NTG-2 condenser microphone with a “deadcat” windscreen at 44.1 kHz, 24 bits in a sound-treated booth at the phonetics laboratory of Leiden University. Participants gave their informed consent 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, which was encouraged by having the listener provide live feedback to the speaker on whether an utterance was perceived as a statement or question. The listener was seated outside the booth in a silent classroom, wearing Sennheiser HD 414 SL head-phones, and used a keyboard to classify each of the speaker’s utterances as either statement or question by pressing one of two dedicated keys. The targets were presented to the speaker one by one and in written form on a computer screen, in a pseudo-random order. Visually, statements versus questions were distinguished by a full stop versus a question mark in the stimulus’ orthography, and initial versus final stress position was indicated by underlining the stressed syllable (e.g., “voornaam” versus “voornaam”). Before the next target was presented, the speaker got on-screen feedback about the
listener’s understanding of the previous one (“correct” versus “incorrect”). This was intended to incite the speaker to produce targets clearly enough to be understood correctly. By keeping the listener outside the booth, and invisible to the speaker, the only cues the speaker could provide were auditory. Participants received written instructions, and subsequently completed a short practice session, using different minimal pairs than during the actual recording. This was the case for both normal and whispered speech, and the order of the speech modes was counterbalanced across speakers. Two repetitions per utterance were recorded and saved as separate wave files (32 per speaker: 4 minimal pairs × 2 sentence functions × 2 speech modes × 2 repetitions).

The listeners, who were present during the recordings, labeled the boundary tone of phonated words correctly in 94% of the cases and of the whispered words, 68% were classified correctly. Remember that all speaker-listener pairs were different. Using a six-point Likert scale (1 = very difficult, 6 = very easy) speakers were asked to rate 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 = 0.013]. In neither speech mode was the task judged as particularly easy. Informal speakers’ responses suggest that this was due to the combined task of expressing lexical stress position and boundary tone in close proximity.

Per lexical item, one instance was annotated manually at the syllable and segment levels, and that annotation was used to automatically annotate all other instances of the same item using a dynamic time warping procedure in Praat (Boersma and Weenink, 2013). Automatic annotations were then manually checked, and corrected if necessary. In some cases, the correctness of pronunciation of either boundary tone realization or lexical stress position realization was questioned by the human annotator (the first author). Those cases (45 phonated and 15 whispered utterances) were subsequently all labeled by two other raters with respect to lexical stress position, which resulted in 24 phonated cases and 12 whispered cases being consistently perceived as “lexical stress produced on the incorrect syllable.” In four phonated cases and two whispered cases the additional raters judged lexical stress position differently. All these cases were excluded from further processing. For whisper, Cohen’s kappa showed an inter-rater agreement of 0.74; for normal speech kappa was 0.82. Suspected mispronunciations with respect to the boundary tone were consistently classified as “lexical stress produced on the correct syllable” by the raters, but inspection of their F0 tracks (all were phonated cases) showed that low boundary tones had indeed been produced as high ones, and the other way around.  

Target words were extracted from the carrier sentences, and their intensity was normalized by setting all recordings within a speaker and speech mode to 60 dB [root mean square (rms) = 0.020]. This corresponded to the minimum intensity of whispered items after scaling peaks to maximum intensity (using Praat: scale peaks...). There were 192 items per speech mode: 8 lexical items (4 initial, 4 final stress) × 2 sentence functions (question, statement) × 12 speakers (6 male, 6 female).

2. Participants and procedure

There were 24 Dutch native listeners (7 males) aged 19–57 (mean = 22 years). Subjects were hearing-screened using an Oscilla USB-300 audiometer to have normal-hearing at octave frequencies between 0.125 and 8 kHz; all were right-handed. 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 sentence function classification task, and the second half of the materials 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 L% (statement) or H% (question), per speaker and per target word in each half.

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. For sentence function classification listeners were instructed to indicate if the target sounded like a statement or a question. For lexical stress position classification listeners were instructed to indicate whether the initial or final syllable carried lexical stress. During practice and the listening tasks, the two response options were shown on screen, while listeners were asked to press one of two corresponding response buttons on a keyboard using their index fingers. Both tasks were once presented with normal speech materials, and once with whispered speech materials, 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 whispered speech), the corresponding whispered and normal speech items from the same speaker were presented to the same subject in the same task. Hearing screening, explanation, practice and four tests were completed within 45 min. Participants received a small fee for their voluntary participation.

B. Analysis and results

Percent correct responses was calculated for each of the subtasks, sentence function (SF) classification and lexical stress position (LS) classification, and transformed to rationalized arcsine units (RAU) (Studebaker, 1985). For the SF and LS reaction times (RTs), which were measured from target word onset, responses under 500 ms and over two standard deviations beyond the mean, computed per listener and per speech mode, were excluded (SF: 2.4% of the data; LS: 4.2% of the data).

Word frequency may influence reaction times with high-frequency words being recognized faster than low-frequency words (e.g., Connine et al., 1993). We here expect minimal effects of word frequency because all words were low-frequency [< 10/million according to Subtlex (Keuleers et al., 2010), a database of 44 × 10^6 Dutch word tokens], and because the same words were presented repeatedly yielding
smaller frequency effects due to repetition priming (e.g., Balota and Spieler, 1999; Lowder et al., 2013). The exact frequencies varied somewhat within pairs: for /plato/ and /kanɔn/ the initial-stress item was more frequent, for the other two pairs it was the other way around.

RTs were transformed to their inverse (1/RT) for analysis. Five out of 48 SF listener-speech mode combinations and nine out of 48 LS listener-speech mode combinations resulted in non-normal distributions. Both the RAU scores and the inverse RTs were subjected to repeated measures (RM) analyses of variance (ANOVAs) (using IBM SPSS Statistics, version 21, generalized linear model > repeated measures, type III sum-of-squares) with within-subjects factors speech mode (normal, whisper), sentence function (question, statement), lexical stress position (initial, final), and minimal pair (4). If sphericity was violated, Huynh–Feldt corrections were applied (also in later analyses).

1. Sentence function classification and reaction times

Figure 1 shows the sentence function classification and RT results across items and listeners, and Table I gives the results of the statistical analyses for both measurements. In whisper, classification was significantly worse than in normal speech (60.8% versus 93.9%, respectively), but above chance level [binomial test: N = 2304, p = 1/2, Z = 11.1, p < 0.001]. Across speech modes, statements were classified correctly more often than questions. The interaction of speech mode by sentence function showed that in whisper, the difference in correct responses to statements versus questions was larger than in normal speech (whisper: 71.4 versus 51.8%; normal speech: 95.5 versus 92.8%).

Across speech modes stimuli with final stress yielded comparable scores for the two sentence functions, but stimuli with initial stress received more correct responses on statements than questions. This interaction was found within both speech modes [normal speech: F(1,23) = 31.7, p < 0.001, η²_par = 0.58; whisper: F(1,23) = 31.7, p < 0.001, η²_par = 0.58]. Absolute differences were largest in whisper (see Fig. 1), where on words with initial stress, statements were classified correctly well above chance level at 80.2%, whereas questions were classified below chance level at 41.3% [N = 576, p = 1/2, Z = −4.2, p < 0.001]. 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 [plato] did speech mode and sentence function interact, showing a larger performance difference between the questions and statements across speech modes.

RTs results showed that listeners were faster in normal speech than in whispered speech (1084 and 1502 ms, respectively), faster on words with final stress than 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 = 0.012, η²_par = 0.30], as reflected by a marginally significant speech mode by stress position interaction [F(1,18) = 4.4, p = 0.051, η²_par = 0.20].

Across speech modes, statements were responded to equally fast in words with initial and final stress, but questions were responded to faster in final-stress words. This sentence function by stress position interaction was found in comparable ways in both speech modes [whisper: F(1,18) = 17.6, p = 0.001, η²_par = 0.49, normal speech: F(1,23) = 16.5, p < 0.001, η²_par = 0.42]. 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 [vornam] pair the trend was in the opposite direction with faster responses to initial-stressed words. The four-way interaction indicated that the differences in response times by lexical stress position were mainly caused by differences measured in whisper.

FIG. 1. Sentence function classification results. Error bars indicate the standard error of the mean. (a) Mean percentage of correct responses to the different sentence functions, per lexical stress position and per speech mode. A reference line is drawn at chance level (= 50%). (b) Mean reaction times in milliseconds to the different sentence functions, per lexical stress position and per speech mode.
2. Lexical stress position classification and reaction times

Figure 2 shows the lexical stress position classification and RT results across items and listeners, and Table II gives the statistical results for both measurements.

The absence of a speech mode main effect shows that identification of lexical stress position went as well in whisper as in normal speech (89.2% and 90.9%, 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 statement received more correct responses when lexical stress fell on the initial syllable, whereas the correctness of responses to words pronounced as questions was comparable for both lexical stress positions. This trend was observed in both speech modes, but the interaction of sentence function by stress position was only significant within whispered speech [F(1,23) = 18.0, p < 0.001, \( \eta^2_{\text{par}} = 0.44 \)], not in normal speech.

In the RT results the absence of a speech mode main effect indicates that listeners were comparably fast at identifying lexical stress position in whispered and in normal speech (1304 and 1336 ms, respectively). Across speech modes, responses were faster to statements (1279 ms) than to questions (1344 ms), especially for the 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 become significant for normal speech; it was significant only for whisper [F(1,23) = 31.2, p < 0.001, \( \eta^2_{\text{par}} = 0.58 \)]. There was variation in the RTs to different words, with fastest responses to \( k\alpha n \nu n \) (1279 ms) and slowest responses to \( s\nu r \nu \nu s \nu c h \) (1399 ms). The trend for responses to statements to be faster than responses to questions was present for all minimal pairs, but the size of the difference varied between them. Finally, the sentence function by stress position interaction was followed by three out of four minimal pairs, but [plato], for which responses to either sentence function were comparably fast.

C. Discussion

Depending on the listening task, the same stimuli were responded to very differently. Whispered speech was as clear as normal speech with respect to the coding of lexical stress position, and on that task, processing was equally fast in both speech modes. For perception of the boundary tone, the situation was very different. In normal speech, listeners performed near ceiling, whereas the same listeners’ performance dropped about 30 percentage points in whisper, and at the same time processing speed decreased significantly. It was expected that in whisper—given the more restricted means of conveying intonation—listeners would have more difficulty correctly identifying boundary tones, but especially when coinciding with lexical stress position. Though the general performance decrease was obtained as expected, the main effect of stress position within whispered speech 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 within responses to whispered materials only.

Listener performance showed comparable means around 60% correct on whispered words with initial and final stress; across stress positions, performance was better on statements than on questions. This in general suggests that cues to questions were less clear in whisper, which was also found in Fónagy (1969), and for Dutch whispered boundary tones more generally (Heeren and Van Heuven, 2009). 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 statements than on questions. Effectively, at only 40% correct, questions were not recognized as such on words with initial stress. For whispered words with final stress, performance was comparable between the sentence functions. These results suggest that only on whispered stimuli in which accent and boundary tone coincided on the same (i.e., final) syllable (“clash”
As performance on questions produced 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 “statement” responses to words with initial stress, whereas equal numbers of either response category were expected [normal speech: 620 out of 1152 “statement” responses, $\chi^2(1) = 6.7$, $p = 0.010$; whisper: 800 out of 1152 “statement” responses, $\chi^2(1) = 174.2$, $p < 0.001$]. In normal Dutch conversational speech, statements occur (much) more often than questions, which is for instance reflected by an analysis of utterance-final orthography in a corpus of spontaneous conversational speech, which is part of the Spoken Dutch Corpus (Oostdijk, 2000); 81.7% of the utterances ($N > 467k$) were transcribed as ending in a full stop (.) versus 8.5% ending in a question mark (?). 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, which is what listeners did. For words with lexical stress on the same (i.e., final) syllable as the boundary tone, whispering speakers were able to convey the sentence’s function. Listeners here did not show the bias, possibly because the syllable’s prominence better supports perception of the boundary tone.

In comparison with other studies on boundary tone perception in whisper (Heeren and van Heuven, 2009; Heeren and Lorenzi, 2014), 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 international events in close proximity, which would be in line with expectations. On the one hand, this is taken to reflect that more complicated linguistic structures, as in the present study, may moderate earlier results on how well prosody is processed in whisper (see also Fónyagy, 1969). 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 sentence function difference, also during the boundary tone task, which in turn may have influenced performance in a negative way.

There was a large difference in the mean reaction times for classification of whispered versus phonated boundary tones. This cannot be explained by the difference in stimulus duration between the speech modes, as this was only on the order of 100 ms (693 vs 583 ms 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. In the next section, it is investigated what acoustic cues to boundary tone realization can be found in whisper, as well as how they are influenced by lexical stress position and speech mode.

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 in Dutch is most reliably realized by durational differences (Sluijter and Van Heuven, 1996; Van Heuven and Sluijter, 1996), which also form a main cue for listeners (Sluijter et al., 1997; Van Heuven and De Jonge, 2011). Moreover, we take the high listener scores to indicate that lexical meaning was generally not influenced by boundary tone realization, in either speech mode. Whispering speakers also do not seem to compromise lexical meaning to enhance the boundary tone contrast. Words with initial stress pronounced as statements 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 questions. A possible explanation is that the nuclear accent is more often confused with a question reading when it falls on the final syllable. Mainly in whisper, responses to questions 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.

III. ACOUSTIC CORRELATES OF BOUNDARY TONES IN WHISPER

The goal of this analysis was to identify acoustic correlates that systematically vary with boundary tone realization, and that may therefore cue boundary tones to listeners in whispered speech. Moreover, correlates can be either secondary or compensatory. By secondary correlates those correlates are meant that exist in normal speech and that are maintained in whispered speech. By “compensatory correlates,” those correlates are meant that systematically vary with boundary tone in whispered speech, but not in normal speech, or that vary to a greater extent in whispered than normal speech. In contrast with earlier studies, a comparison with phonated speech was made here, as a result of which the nature of the correlates could be investigated.

As boundary tones are tied to phrase beginnings and ends (see Gussenhoven, 2005, for Dutch), acoustic correlates of boundary tones in the context studied here are expected to mainly appear over the final syllable of the utterance. We therefore focus on final syllable rhymes. In the literature, acoustic correlates of intonation in whisper have been sought in vowel content, and mainly in vowel formants, intensity and duration. More recently, acoustic correlates of intonation have been demonstrated in voiceless stop and fricative consonants in normal speech as well (Cho and McQueen, 2005; Niebuhr, 2008). Niebuhr (2008) showed that aspiration of German utterance-final /l/ differed systematically between two accent contours in duration, maximum intensity and the location of the spectral peak above the lower spectral energy boundary (Niebuhr, 2008). Cho and McQueen (2005) found that the stops /t/ and /d/ showed longer closure durations with stronger prosodic boundaries in domain-initial position. With varying prosodic boundary depth these stops also differed in their center of gravity (CoG), but not in a consistent manner. In that same study, the fricatives /s/ and /z/ showed longer durations with stronger boundaries, whereas intensity was lower with stronger boundaries, and most clearly for /s/. CoG in fricatives did not vary with prosodic boundary strength in domain-initial position. Niebuhr (2012) studied the interaction of utterance-final changes in F0 and the realization of the voiceless fricatives /f/, /ʃ/, /f/, and /s/ in utterance-final position, placed directly after an F0 fall or F0 rise. In rising F0 contexts that were intended as questions, voiceless fricatives had a higher CoG and intensity. In whispered speech, comparable results have been obtained: the voiceless fricatives /f/ and /s/ vary acoustically with the speaker’s intended vocal pitch (Heeren, 2014): nonsense VCVs produced at different intended heights (low, medium, high) showed systematic increases in CoG and intensity with increasing height.

Listeners can be sensitive to acoustic correlates of intonation in consonants, as shown in two studies on German that used listeners’ ratings along semantic dimensions, such as questioning/non-questioning and certain/uncertain (Niebuhr, 2008; Kohler, 2011). Listeners’ responses were also supported by increases in intensity and duration for the higher utterances. In comparison with earlier studies, however, the pressure put on the final syllable in the case that lexical stress position and boundary tone co-occur may result in less and smaller acoustic differences between the boundary tones in that condition.

A. Analysis

As a demonstration of the speakers’ consistent manner of signaling sentence function in normal speech, Fig. 3 reflects the time-normalized mean F0 trajectories for statements and questions, for male and female speakers at 20%, 40%, 60%, and 80% into the utterances’ final vowels. Individual trajectories were generally consistent with this pattern, showing a clear separation between statements and questions for all speakers, but the timing of the final rise onset in questions showed some variability between speakers: most were relatively early (20%–40%), but some were later (60%). We assume that in whispered speech, speakers were aiming for comparable targets.

For the vowels we measured duration in milliseconds (ms), mean intensity in decibels (dB), and the first through third formants (F1, F2, F3) in Hertz (Hz), along with their bandwidths (B1, B2, B3 in Hz) and levels (A1, A2, A3, in dB). In addition to the formant measurements a more general measure of spectral composition was included in the form of the spectral center of gravity (CoG, in Hz). If CoG would...
were main effects of sentence function and interactions of sentence function by speech mode. A main effect of sentence function is an indication of a secondary correlate to intonation, as the correlate would be available similarly in normal speech and in whisper. A speech mode by sentence function interaction could indicate a compensatory correlate, namely, in the case that the acoustic difference is present only or larger in whispered than in normal speech. Second, interactions involving the factors sentence function and stress position might be indicative of a difference in correlates to boundary tones for the prosodic clash versus no-clash conditions. In the upcoming results subsections, the effects directly relevant to the research questions are presented first, followed by other effects present in the data.

B. Results for the final syllable vowels

1. Duration

There was a main effect of sentence function with questions (191 ± 74 ms) being longer than statements (170 ± 57 ms) \([F(1,11) = 7.1, \, p = 0.022, \, \eta^2_{par} = 0.39]\). This main effect was also present within normal speech with means of 164 ms (±44 ms) and 151 ms (±46 ms), \([F(1,11) = 10.1, \, p = 0.009, \, \eta^2_{par} = 0.48]\), but did not become significant in whisper, \([F(1,11) = 3.5, \, p = 0.11, \, \eta^2_{par} = 0.26]\) having more intensity than statements \([F(1,11) = 2.4, \, p = 0.14, \, \eta^2_{par} = 0.19]\). Vowel duration was longer in whisper (203 ± 77 ms) than in normal speech (158 ± 45 ms). The presence of lexical stress on the final syllable lengthened that syllable (190 ms vs 171 ms for the unstressed case), and within speech modes, the effect was also found. In normal speech, mean vowel durations were 165 and 151 ms, respectively \([F(1,11) = 19.4, \, p = 0.001, \, \eta^2_{par} = 0.64]\), and in whisper, mean vowel durations were 215 and 190 ms, respectively \([F(1,11) = 14.1, \, p = 0.003, \, \eta^2_{par} = 0.56]\). There were two-way interactions of speech mode by minimal stress pair \([F(3,33) = 8.6, \, p < 0.001, \, \eta^2_{par} = 0.44]\), and of stress position by minimal stress pair \([F(2,4,26.7) = 9.6, \, p < 0.001, \, \eta^2_{par} = 0.47]\). The former interaction indicates that the duration difference between the speech modes was not fully comparable across minimal stress pairs; for /s/ vs /s/, the difference was relatively smaller than for the other three pairs. The latter interaction indicates that across speech modes, only final vowel durations in /pla/ and /vom/ varied with the presence of lexical stress on that syllable.

2. Mean intensity

There was a main effect of sentence function with questions (64.3 ± 12.0 dB) having more intensity than statements (61.5 ± 11.8 dB), \([F(1,11) = 44.2, \, p < 0.001, \, \eta^2_{par} = 0.80]\), and the speech mode by sentence function interaction was marginally significant \([F(1,11) = 4.2, \, p = 0.065, \, \eta^2_{par} = 0.28]\). The effect of sentence function was significant within both
speech modes [normal speech: F(1,11) = 40.1, p < 0.001, \( \eta^2_{\text{par}} = 0.79 \); whisper: F(1,11) = 19.0, p < 0.001, \( \eta^2_{\text{par}} = 0.63 \)], but absolute differences in intensity between the sentence functions were larger in normal speech (75.4 vs 72.0 dB) than in whisper (53.2 vs 51.1 dB). There were two higher-order interactions involving the factor sentence function: sentence function by stress position [F(1,11) = 5.7, \( p = 0.036, \eta^2_{\text{par}} = 0.34 \)], and speech mode by sentence function by stress position [F(1,11) = 41.8, \( p < 0.001, \eta^2_{\text{par}} = 0.79 \)]. Across speech modes, final vowels in words with final stress had more intensity, and final vowels in questions had more intensity than in statements. In whisper, the intensity difference between the sentence functions was comparable between words with initial and final stress, whereas in normal speech, the difference was especially large when stress fell on the word’s initial syllable.

There were main effects of speech mode, showing that final vowels in whisper had much less intensity than in normal speech, [F(1,11) = 553.1, \( p < 0.001, \eta^2_{\text{par}} = 0.98 \)], of stress position [F(1,11) = 225.1, \( p < 0.001, \eta^2_{\text{par}} = 0.95 \)], and of minimal stress pair [F(3,33) = 19.9, \( p < 0.001, \eta^2_{\text{par}} = 0.64 \)]. The latter effect represents intrinsic vowel differences: across speech modes, final vowel realizations of /vornam/ had most intensity (mean = 65.2 dB) and realizations of /s/vorn/ had the least (mean = 61.2 dB). The target vowel had more intensity, across speech modes, in words with final than initial stress. There were two-way interactions of speech mode by minimal stress pair [F(3,33) = 35.2, \( p < 0.001, \eta^2_{\text{par}} = 0.76 \)], and of stress position by minimal stress pair [F(3,33) = 6.4, \( p = 0.002, \eta^2_{\text{par}} = 0.37 \)]. These interactions indicate that the speech mode and stress position main effects were maintained per minimal stress pair, but that absolute intensity differences between the conditions varied per pair.

3. First–third formants (F1–F3)

a. F1. There was no main effect of sentence function [F(1,11) = 1.5, n.s.], nor a sentence function by speech mode interaction [F < 1].

There were main effects of speech mode [F(1,11) = 211.4, \( p < 0.001, \eta^2_{\text{par}} = 0.95 \)], and of minimal stress pair [F(1,4,15,2) = 119.1, \( p < 0.001, \eta^2_{\text{par}} = 0.92 \)]. F1 was higher in whisper than in normal speech, showing a 130–350 Hz mean difference per vowel, variation that is reflected by the interaction of speech mode by minimal stress pair [F(1,8,19,4) = 11.1, \( p = 0.001, \eta^2_{\text{par}} = 0.50 \)]. There was an interaction of speech mode by stress position, [F(1,11) = 8.5, \( p = 0.014, \eta^2_{\text{par}} = 0.44 \)], showing that differences in final vowel F1 between words with initial vs final stress were small in both speech modes but in opposite directions, and there was an interaction of stress position by minimal stress pair, [F(3,33) = 3.0, \( p = 0.045, \eta^2_{\text{par}} = 0.21 \)], which reflected that only the final vowel’s F1 in /vornam/ was higher when the final syllable was stressed, whereas there was no effect of stress for the other three minimal pairs.

b. F2. There was a speech mode by sentence function interaction [F(1,9) = 7.6, \( p = 0.022, \eta^2_{\text{par}} = 0.46 \)], but no sentence function main effect [F(1,9) < 1], see Fig. 4. Across speech modes, F2 did not depend on the sentence function, and the interaction shows that in whisper, F2 was somewhat higher in questions than statements (1522 and 1505 Hz, respectively), whereas it was the other way around in normal speech (1382 and 1395 Hz, respectively). In neither speech mode, however, was there a main effect of sentence function, showing that F2 did not vary with boundary tone.

There were main effects of speech mode [F(1,9) = 26.4, \( p = 0.001, \eta^2_{\text{par}} = 0.75 \)], and of lexical stress pair [F(1,7,15,6) = 830.0, \( p < 0.001, \eta^2_{\text{par}} = 0.99 \)]. In whisper, F2 was 125 Hz higher across vowels than in normal speech, and an interaction of speech mode by lexical stress pair reflected that the absolute size of the difference between the speech modes varied between final vowels, [F(3,27) = 24.2, \( p < 0.001, \eta^2_{\text{par}} = 0.73 \)]. Finally, there was an interaction of stress position by minimal stress pair [F(3,27) = 16.7, \( p < 0.001, \eta^2_{\text{par}} = 0.65 \)], which showed that the presence/absence of lexical stress also influenced the final vowel’s F2, but not similarly across minimal stress pairs. Across speech modes, the /s/vorn/ pair showed a higher F2 on a stressed than unstressed final vowel, the /plato/ pair showed a trend in the opposite direction, and the other two pairs showed no differences.

c. F3. There was no sentence function main effect, [F(1,11) < 1], and no speech mode by sentence function interaction, [F(1,11) < 1]. There were interactions of sentence function by minimal stress pair showing that, across speech modes, the difference in F3 between statements and questions varied with minimal stress pair, where in addition to the size of the difference, the direction could change [F(1,7.7, 18.8) = 3.8, \( p = 0.047, \eta^2_{\text{par}} = 0.26 \)]. The four-way interaction was also significant, [F(2.5,27,5) = 3.6, \( p = 0.033, \eta^2_{\text{par}} = 0.25 \)].

There were main effects of stress position showing a higher F3 when the syllable is stressed, [F(1,11) = 14.4,
There was a main effect of sentence function with questions having higher power than statements, [F(1,11) = 7.3, p < 0.001, η^2_{par} = 0.96], of stress position showing a higher level when the syllable was stressed, [F(1,11) = 89.8, p < 0.001, η^2_{par} = 0.89], and of minimal stress pair, [F(3,33) = 4.2, p = 0.013, η^2_{par} = 0.28]. There were interactions of speech mode by minimal stress pair, [F(3,33) = 29.8, p < 0.001 η^2_{par} = 0.73], which reflected that the level difference between the speech modes varied per pair, and of stress position by minimal stress pair [F(3,33) = 4.0, p = 0.016, η^2_{par} = 0.27], showing that the level difference with the presence/absence of stress varied per pair. In the two latter interactions, main effects were preserved for each of the minimal stress pairs.

b. A2. There was a main effect of sentence function with a higher level in questions than statements, [F(1,11) = 8.9, p = 0.012, η^2_{par} = 0.45]. The speech mode by sentence function interaction was not significant, [F(1,11) < 1]. There was an interaction of sentence function by minimal stress position, [F(1,11) = 10.1, p = 0.009, η^2_{par} = 0.48], showing that only when the final syllable was unstressed, was there a power difference between the sentence functions.

There were main effects of speech mode, showing a higher level in normal than whispered speech, [F(1,11) = 30.1, p < 0.001, η^2_{par} = 0.73], of stress position, reflecting that the presence of stress increased A2, [F(1,11) = 58.9, p < 0.001, η^2_{par} = 0.84], and of minimal stress pair, [F(3,33) = 43.8, p < 0.001, η^2_{par} = 0.80]. There was a two-way interaction of speech mode by minimal stress pair, [F(3,33) = 10.4, p < 0.001, η^2_{par} = 0.49]. The absolute difference between the mean power per minimal stress pair in normal vs whispered speech differed between pairs, whereas the main effect of speech mode was reflected in each minimal stress pair’s results.

c. A3. There was no main effect of sentence function, [F < 1], nor a speech mode by sentence function interaction, [F(1,11) = 3.5, n.s.], but there was an interaction of sentence function by stress position by minimal stress pair [F(3,33) = 6.7, p = 0.001, η^2_{par} = 0.38]. This reflected that across speech modes, A3 was higher when the vowel was stressed than unstressed, but that the power difference between questions and statements was larger in unstressed vowels, and that the direction depended on the minimal stress pair.

There were main effects of speech mode, with power being higher in normal than whispered speech, [F(1,11) = 14.6, p = 0.003, η^2_{par} = 0.57], of stress position with power being higher in stressed than unstressed final syllables, [F(1,11) = 84.3, p < 0.001, η^2_{par} = 0.89], and of minimal stress pair, [F(3,33) = 29.6, p < 0.001, η^2_{par} = 0.73].

5. F1–F3 bandwidth (B1–B3)

None of the formant bandwidths showed main effects or interactions involving the factor sentence function, meaning that they did not serve as acoustic correlates of boundary tones in normal and whispered speech in the word types studied here. They will therefore not be discussed further.

6. Center of gravity (CoG)

The main effect of sentence function did not reach significance, [F(1,11) = 4.0, p = 0.070], but there was an interaction of speech mode by sentence function [F(1,11) = 5.2, p = 0.043, η^2_{par} = 0.32]. In whispered final vowels, CoG in questions (1535 Hz) was higher than in statements (1442 Hz) [F(1,11) = 4.8, p = 0.050, η^2_{par} = 0.31], whereas this was not the case in normal speech (418 vs 411 Hz, respectively), [F(1,11) < 1]. Furthermore, a speech mode by sentence function by minimal stress pair interaction, [F(3,33) = 12.8, p < 0.001, η^2_{par} = 0.54], and one of sentence function by minimal stress pair, [F(3,33) = 14.3, p < 0.001, η^2_{par} = 0.57], reflected that in normal speech there was no difference in CoG as a function of the sentence function, and that the effect in whisper was mainly explained by an effect of sentence function for the /serves/ pair (see Fig. 5), and reflected by a significant difference between the sentence functions for that pair only [F(1,11) = 19.1, p = 0.001].

There were main effects of speech mode [F(1,11) = 182.5, p < 0.001, η^2_{par} = 0.94], of stress position [F(1,11) = 20.0, p = 0.001, η^2_{par} = 0.65], and of minimal stress pair [F(1,21.2) = 78.4, p < 0.001, η^2_{par} = 0.88]. The final vowel’s CoG was higher in whisper (1489 Hz) than in normal speech (414 Hz), higher when the final syllable was stressed rather than unstressed, and varied between final vowels. There furthermore were interactions of speech mode by minimal stress pair, showing that the largest CoG effect between the speech modes was found for /serves/, whereas the difference between the speech modes was comparable for the other three minimal pairs, [F(1,5,16.7) = 98.7, p < 0.001, η^2_{par} = 0.90], and of stress position by minimal
stress pair showing that, across speech modes, the effect of stress position was largest in /serves/’ final vowel, \(F(1.7,19.1) = 8.9, p = 0.003, \eta^2_{\text{par}} = 0.45\). Finally, there was a three-way interaction of speech mode by stress position by minimal stress pair \(F(3.33) = 13.2, p < 0.001, \eta^2_{\text{par}} = 0.55\).

C. Results for the coda consonants (/s, m, n/)

1. Duration

Coda duration was not influenced by sentence function. There were main effects of speech mode \(F(1,11) = 7.9, p = 0.017, \eta^2_{\text{par}} = 0.42\), of stress position \(F(1,11) = 23.0, p < 0.001, \eta^2_{\text{par}} = 0.68\), and of minimal stress pair \(F(1.5,16.2) = 33.3, p < 0.001, \eta^2_{\text{par}} = 0.75\). There also was a speech mode by stress position by minimal stress pair interaction \(F(2,22) = 8.0, p = 0.002, \eta^2_{\text{par}} = 0.42\). Coda durations were longer in whisper (147 ms) than in normal speech (126 ms), and longer when the syllable was stressed than unstressed (145 vs 128 ms, respectively). The interaction showed that the effect of stress position was comparable across speech modes and minimal stress pairs, but that the difference between the speech modes was not constant over stress pairs.

2. Intensity

Mean intensity showed a main effect of sentence function, with coda consonants in questions (58.5 dB) having a higher intensity than in statements (55.9 dB) \(F(1,11) = 25.8, p < 0.001, \eta^2_{\text{par}} = 0.70\), and a marginally significant speech mode by sentence function interaction \(F(1,11) = 4.8, p = 0.051, \eta^2_{\text{par}} = 0.30\). The main effect of sentence function was found within each of the speech modes [normal: \(F(1,11) = 20.1, p = 0.001, \eta^2_{\text{par}} = 0.65\]; whispered: \(F(1,11) = 10.5, p = 0.008, \eta^2_{\text{par}} = 0.49\). There was an interaction of speech mode by sentence function by minimal stress position \(F(1,11) = 5.3, p = 0.042, \eta^2_{\text{par}} = 0.32\), showing that when the final syllable was stressed, intensity differences between coda consonants in questions and statements were comparable in whispered and normal speech, but when the final syllable was unstressed, the difference between the sentence functions was no longer found in whisper.

There were further main effects of speech mode \(F(1,11) = 563.9, p < 0.001, \eta^2_{\text{par}} = 0.98\), and of stress position \(F(1,11) = 41.9, p < 0.001, \eta^2_{\text{par}} = 0.79\). Coda consonants had more intensity in normal speech than in whisper, and when the syllable was stressed than unstressed. An interaction of speech mode by minimal stress pair showed that both /m/ and /n/ had less intensity in whisper than in normal speech, but for /s/, intensities were comparable between the speech modes, \(F(2,22) = 34.2, p < 0.001, \eta^2_{\text{par}} = 0.97\).

3. Center of gravity

The CoG showed a main effect of sentence function with a higher CoG in questions (1876 Hz) than statements (1740 Hz) \(F(1,11) = 26.7, p < 0.001, \eta^2_{\text{par}} = 0.71\), but no speech mode by sentence function interaction \(F(1,11) < 1\). The main effect was present within each of the speech modes [normal: \(F(1,11) = 8.4, p = 0.015, \eta^2_{\text{par}} = 0.43\); whispered: \(F(1,11) = 9.8, p = 0.010, \eta^2_{\text{par}} = 0.47\)]. There also was a sentence function by minimal stress pair interaction, indicating that across speech modes, CoG of /s/, but not of /m/ and /n/, varied with sentence function \(F(2,22) = 7.2, p = 0.004, \eta^2_{\text{par}} = 0.39\).

There were further main effects of speech mode, with CoG being higher in whisper than in normal speech \(F(1,11) = 27.8, p < 0.001, \eta^2_{\text{par}} = 0.72\), of stress position, with CoG being higher in stressed than unstressed syllables \(F(1,11) = 21.5, p = 0.001, \eta^2_{\text{par}} = 0.66\), and of minimal stress pair reflecting that the CoG of /s/ was much higher (4327 Hz) than that of /m/ or /n/ (with means between 524 and 573 Hz), \(F(2,22) = 446.5, p < 0.001, \eta^2_{\text{par}} = 0.98\). There was an interaction of speech mode by minimal stress pair, showing that CoG of /m/ and /n/ changed with speech mode (due to presence/absence of voicing), but not that of /s/, \(F(2,22) = 31.8, p < 0.001, \eta^2_{\text{par}} = 0.74\).

D. Discussion

Over the final syllables’ rhymes we looked for acoustic correlates of boundary tones in whisper, which may be cues to boundary tone perception for listeners. Moreover, we were interested in the nature of the correlates, either secondary or compensatory. A number of acoustic correlates systematically varied with boundary tone, and most of these were of a secondary nature, that is present in comparable ways in normal and whispered speech. These correlates are vowel duration, vowel intensity, F1 and F2 power, consonant intensity, and...
consonant CoG; correlates were longer or higher in the case of a high boundary tone. Compensatory information could be present in the form of the final vowel’s CoG.

Earlier work suggested that mainly the lower formants, F1 and F2, could carry prosodic information in the absence of F0 (Higashikawa and Minifie, 1999; Heeren and Van Heuven, 2009; Konno et al., 2006), but changes in higher formants, such as F3, have been reported as well (Meyer-Eppler, 1957; Fónagy, 1969). As listener cues to intonation in the absence of F0, spectral correlates are thought to be the most direct ones, and of such cues, the formants are suitable candidates (cf. De Cheveigné, 2005). But here it was not the case that formants were consistently higher when a high boundary tone was intended, and one reason for the absence of this effect may be that formants were also influenced by the presence of stress, as demonstrated by stress position main effects for F1 and F3. Only whispered F2s showed a tendency to be higher in questions than statements, but this trend was not confirmed by a main effect of sentence function.

Instead of shifts in spectral peaks, i.e., formants, a more general measure of spectral composition, the final vowel’s center of gravity, seemed to be a compensatory acoustic correlate of boundary tone. When looking more closely, this effect was only found for the /sɛrvis/ minimal stress pair: CoG in questions was higher than in statements for this pair when whispered, but not when spoken normally. For the coda consonants, CoG was a secondary correlate as it varied with sentence function in both speech modes (see also Heeren, 2014 and Niebuhr, 2008 for comparable results on normal speech). On both segment types the presence of stress affected CoG, making it higher in stressed than unstressed syllables (see also Sluijter and Van Heuven, 1996). When looking for the cause of the CoG shift related to signaling the boundary tone contrast, formant frequencies cannot provide an explanation as they were not found to change with sentence function, not even for whispered [i] only. Intensity, and with it formant levels, showed a change with sentence function, but only when this increase would not be distributed uniformly across the spectrum is a shift in CoG expected. When inspecting each of the whispered minimal stress pairs individually, a less negative spectral tilt was seen for questions over statements in the /sɛrvis/ pair. This was confirmed by RM ANOVAs on the spectral tilt per minimal stress pair, computed through a comparison of energy in the 0.5–2 kHz and 2–8 kHz bands, with within-subject factors sentence function and stress position. This yielded a main effect of sentence function for /sɛrvis/ only, [F(1,11) = 5.6, p = 0.038].

In addition to some spectral correlates, other acoustic dimensions systematically varied with intonation. Final vowels in whispered questions tended to be longer than in statements. In Heeren and Van Heuven (2009) a comparable result was obtained on whispered Dutch boundary tones that always fell on unstressed syllables. Coda consonant durations did not vary with the sentence function, in either speech mode. Duration may thus function as a cue to intonation in whisper, which was also reported by Fónagy (1969). Moreover, in the case of lexical tone identification, Liu and Samuel (2004) found that the secondary cue of duration was informative for listeners. The presence of lexical stress, realized as an accent, in some cases also lengthened segments in the final syllable: across speech modes, the final vowel durations in /plato/ and /vornam/ varied with the presence of lexical stress, and coda consonants were longer in stressed than unstressed syllables. This finding seems in contrast with earlier research on the interaction of accent and final lengthening in Dutch, which showed that the two effects are not additive, which they are in English (Cambier-Langeveld and Turk, 1999). Our results suggest that when under pressure, also the Dutch may add the two durational effects.

Both final vowels and coda consonants showed more intensity in questions than statements, but absolute differences in intensity between the sentence functions were larger in normal speech than in whisper. Comparable results were obtained for whispered versus normal speech boundary tones in Dutch (Heeren and Van Heuven, 2009) and American English (Heeren and Van Heuven, 2011). Moreover, whispered vowels intended as higher have more generally been found to have more intensity (Meyer-Eppler, 1957; Denes, 1959; Fónagy, 1969, Heeren and Lorenzi, 2014). Intensity therefore seems to be a rather consistent correlate of boundary tones in whisper. It also varied with the presence of stress in both segment types with more intensity in stressed than unstressed final syllables, which is consistent with the literature on Dutch (Sluijter and Van Heuven, 1996). As a correlate of the boundary tone, intensity was furthermore influenced by the presence of stress on the final syllable. For coda consonants it was the case that intensity differences between questions and statements were comparable in whispered and normal speech for stressed final syllables, but the intensity difference between the sentence functions was no longer found in whisper when the final syllable was unstressed. For whispered final vowels, the intensity difference between the sentence functions was comparable between words with initial and final stress, whereas in normal speech, the difference was larger when stress fell on the word’s initial syllable. This provides additional evidence for the possibility of an additive effect in Dutch, as in English (Cambier-Langeveld and Turk, 1999), in the case that speakers’ means for expressing contrasts are restricted. In accordance with the results on intensity, the lower two formants had more power in questions than statements. But interactions with stress position moderated this finding to final-syllable power differences in the case of words with initial stress only, that is when lexical stress does not coincide with the boundary tone carrier. For F3, power did not depend on the sentence function, and for all three formants, power was influenced by the presence of stress.

Many of the acoustic correlates included in this investigation were influenced by the presence versus absence of stress on the final (i.e., target) syllable. This finding confirms our assumption that cues to boundary tones in whisper are “under pressure” as they at the same time contribute to expressing other linguistic contrasts, in this case stress position and vowel identity. More specifically, did our data provide evidence for the hypothesis that boundary tones are harder to express in whisper when coinciding with lexical
stress, that is, in words with final stress? To answer this question we translated “harder to express” to “produced with fewer and smaller acoustic differences,” where the latter would be captured by sentence function by stress position interactions in the data. Such interactions were only found for F1 power and F2 power (i.e., A1 and A2), and showed that power differences between the sentence functions were only found in unstressed final syllables, i.e., in the no clash condition. This finding is in line with the prediction, but the scope of the differences, being restricted to formant power, is smaller than expected. An explanation may be that in general, acoustic correlates to boundary tones occurring on words that also carry lexical stress, realized by a nuclear accent, were restricted. We hereby mean that speakers hardly compensated for the loss of F0 in whispered speech, and that secondary correlates did not include changes in formant values, which are considered the best cues in the literature. Therefore, results seem to be in line with our general hypothesis that more complicated linguistic structures, such as the one used in this study, may be more difficult to produce, and therefore to perceive, in whisper. This is also reflected in the results of the perception experiment, which yielded classification scores that were lower than those reported before on Dutch boundary tones in a no-clash situation (79% correct: Heeren and Van Heuven, 2009).

Even though our speakers were prompted to produce listener-directed speech, having been recorded in the presence of a listener from whom they got direct feedback, they had difficulty providing boundary tone cues in whisper. It is expected, however, that if spectral changes would be relevant for listeners to hear out the boundary tone in whisper, and we here found CoG to change in whispered /s\text{ervis}/ only, that boundary tone classification on that particular minimal stress pair would be better than on the other pairs. This pattern of results was indeed reflected by the listener results of Sec. II that yielded the highest mean score for the /s\text{ervis}/ minimal pair in whisper (64% vs 59%–60% for the other pairs). Higher listener scores on one whispered minimal stress pair provide only little evidence for the role of this spectral cue in boundary tone perception. In the next section, we report our investigation of whether enhancement of this cue would aid listeners.

IV. EXPERIMENT II: ENHANCING THE BOUNDARY TONE CONTRAST

Classification of the boundary tone, when it occurred close to a nuclear accent associated with lexical stress, was difficult in whisper (see Sec. II); averaged over conditions, listeners classified the sentence function correctly in only about 60% of the cases. In this experiment we investigated if listener performance could be improved by enhancing that acoustic cue to sentence function that seemed to underlie the compensatory information provided by our speakers in whispered as opposed to normal speech, that is, a change in the center of gravity. This change was seemingly due to a change in spectral tilt rather than a change in formant locations. The expectation was that listener performance would improve as a result of enhancement of this particular cue.

A. Method

The natural speaker tendencies to produce higher CoGs in questions and lower CoGs in statements were exaggerated by tilting the spectra of all target, that is final, syllables while preserving syllable intensity; spectra of questions were positively tilted (+3 dB/octave), and spectra of statements were negatively tilted (−3 dB/octave), which was accomplished by multiplying the spectra with a linear function of the desired steepness in logarithmic space. The value of ±3 dB/octave was chosen as it was the maximum tilt that did not lead to audible artifacts according to informal piloting. On average, the CoG (measured over the 0.05–8 kHz range) of final syllables in manipulated and original questions were 1792 vs 1476 Hz, respectively, and of final syllables in manipulated and original statements were 1117 vs 1418 Hz, respectively. For manipulated stimuli the unaltered initial syllable of the target word was concatenated with the tilted final syllable of the target word with 5-ms overlap, using the concatenation procedure in PRAAT. After manipulation, stimuli were intensity-normalized by setting all recordings within a speaker and condition to 60 dB (rms = 0.020). Figure 6 shows example spectral envelopes for the final vowel [a] produced by one of the male speakers for both the original speech stimuli and the manipulated ones.

There were 24 (10 male) Dutch native listeners between 18 and 53 years of age (mean = 23 years). Subjects were hearing-screened using an Oscilla USB-300 audiometer to
have normal-hearing at octave frequencies between 0.125 and 8 kHz, and all were right-handed. Listeners participated in classification tasks for both the original whispered stimuli and the manipulated whispered stimuli (see Sec. II A 2, for details). Each condition included 192 trials: 8 words (4 initial, 4 final stress) × 2 sentence functions (question, statement) × 12 speakers (6 male). Task order, presentation lists and response buttons were counterbalanced across subjects.

B. Analysis, results and discussion

Data were processed and analyzed as in Sec. II (6.7% of the RT results excluded). Two listeners’ RTs stood out, being three to four times larger than those of the other listeners. Assuming that these listeners did not attempt to respond as quickly as possible, their RTs were left out of the analysis. Exclusion of these two subjects did not affect the identification results, which is why those results are presented including all listeners. Across sentence functions, both RAU scores and inverse RTs were subjected to RM ANOVAs with within-subjects factors speech condition (original, manipulated), stress position (initial, final), and minimal stress pair (4).

Identification responses [see Fig. 7(a)] showed main effects of speech condition [F(1,23) = 8.1, p = 0.009, $\eta^2_{\text{par}} = 0.26$]. In line with expectations, more correct responses were given to the manipulated (62.9%) than the original whispered speech items (60.1%). The absolute improvement was small, however. There furthermore were two-way interactions of speech condition by stress position [F(1,23) = 6.2, p = 0.020, $\eta^2_{\text{par}} = 0.21$], and of speech condition by minimal stress pair [F(3, 69) = 4.3, p = 0.007, $\eta^2_{\text{par}} = 0.16$]. The former interaction showed that only in the original whispered speech was there an effect of stress position on performance with higher scores on words with initial rather than final stress [original: F(1,23) = 6.9, p = 0.015, $\eta^2_{\text{par}} = 0.23$; manipulated: F(1,23) < 1]. Figure 7(a) shows that the absence of the difference in manipulated stimuli was due to improvement on mainly the words with final syllable stress. This suggests that the manipulation only aided boundary tone perception when the syllable it was applied to was also stressed. Absence of a manipulation effect on boundary tone perception in words with initial stress may be due to the relatively low prominence of the manipulated syllable, rendering the manipulation less to not effective. The latter interaction showed that performance tended to be better on the manipulated than the original whispered stimuli but not to exactly the same degree across stress pairs.

Reaction times did not show an effect of tilt manipulation [F(1,21) = 1.3, n.s.], meaning that processing speed was not, positively or negatively, affected by the manipulation [see Fig. 7(b)]. There were main effects of stress position [F(1,21) = 30.1, p < 0.001, $\eta^2_{\text{par}} = 0.59$], with faster responses to words with final stress (1450 ms) than initial stress (1511 ms) [manipulated: F(1,21) = 16.4, p = 0.001, $\eta^2_{\text{par}} = 0.44$; original: F(1,21) = 25.6, p < 0.001, $\eta^2_{\text{par}} = 0.55$], and of minimal stress pair [F(3,63) = 22.2, p < 0.001, $\eta^2_{\text{par}} = 0.51$]. There was a two-way interaction of stress position by minimal stress pair [F(3,63) = 7.1, p < 0.001, $\eta^2_{\text{par}} = 0.25$]. Three out of four minimal pairs followed the stress position main effect, but /vornam/, which did not show faster responses to its final-stressed form. This was explained by a trend in the opposite direction in the manipulated speech condition only, as revealed by a three-way interaction of speech condition by stress position by minimal stress pair [F(3,63) = 2.8, p = 0.046, $\eta^2_{\text{par}} = 0.12$].

As a result of the manipulation, all individual speakers’ stimuli received responses that were above chance level (N = 352, p = 1/2, Z ≥ 2.45, p ≤ 0.016), whereas the original stimuli did not yield above-chance performance on speakers 1–3. In a post hoc analysis we assessed the question of whether the tilt manipulation was equally effective on all speakers’ productions, or differentially affected listener scores by speaker quality. Data from those four speakers who had received the highest listener scores and those four speakers who had received the lowest listener scores in experiment I were selected in the current experiment’s results, and aggregated per listener and per condition. An RM ANOVA with within-subjects factors speaker quality (2) and speech condition (2) showed a main effect of speaker quality, [F(1,23) = 36.5, $p < 0.001, \eta^2_{\text{par}} = 0.61$], as well as a speaker quality by speech condition interaction, [F(1,23) = 7.7, $p = 0.011, \eta^2_{\text{par}} = 0.25$]. As Fig. 8 shows, the largest performance difference as a result of tilt manipulation was found for the poorest speakers. This was confirmed by a significant correlation between performance in the original speech condition and amount of improvement by speaker, [Pearson’s $r = -0.77, p = 0.003$].

In sum, changing the spectral tilt by plus or minus 3 dB/octave over the final syllable, which in turn caused an upward or downward shift in CoG of about 300 Hz, resulted in a significant, but small improvement of listeners’ boundary tone classification results. This was the case especially in final-stress words, and especially for those speakers whose utterances had received the poorest scores in the original speech condition.
V. GENERAL DISCUSSION AND CONCLUSION

The perception and production of Dutch whispered boundary tones were investigated as a function of characteristics of the tone-bearing word to better understand how intonation is coded in whispered speech. In general, perception of whispered boundary tones depends on the tone-bearer, as the close proximity of intonational events seemed to affect overall perception performance, which we estimate to be lower than in earlier studies. We thus provided additional evidence for context-dependencies in the coding of whispered intonation: in addition to the target syllable’s vowel quality (Meyer-Eppler, 1957; Heeren and Lorenzi, 2014), and the position of the target syllable in the utterance (Gao, 2002), we here showed that the phonological structure of the tone-bearing word matters. Within the tone-bearing word, it made a difference whether the boundary tone occurred on an unstressed or on a stressed final syllable, as only in the latter case were both sentence functions perceived above chance level. This suggests that only in a setting in which the target syllable bears sufficient prominence, is the coding of a whispered boundary tone perceived. This latter observation was contrary to the prediction that the condition in which no tonal clash occurred would allow for better perception of boundary tones, as there would be a greater separation of the tonal events in that case.

We expected mainly spectral cues to contribute to pitch coding in whisper, and that speakers would use acoustic correlates in a compensatory way. There were a number of acoustic correlates that systematically varied with the boundary tone, but the majority of those correlates were of a secondary nature. This included correlates in final consonants, which fits in with results for normal speech (Cho and McQueen, 2005; Niebuhr, 2008), with earlier results on whispered fricatives produced at different heights (Heeren, 2014), and more generally with recent findings on the interaction between segmental and suprasegmental information in speech (cf. Kohler, 2012). The center of gravity in final vowels provided some compensatory information in whispered speech, be it only in one out of four minimal stress pairs. This finding is in line with the observation that melodic realization in whisper may vary with the carrier, as first proposed by Meyer-Eppler (1957). He observed that for some vowels formant positions changed whereas for others there was only an increase in intensity in “higher spectral regions” (Meyer-Eppler, 1957, p. 181).

Assuming that speakers try to meet their listeners’ needs, and put in more effort when this is required by the communicative situation (e.g., Lindblom, 1990; Picheny et al., 1986), it seems surprising that speakers did not provide much evidence for it here. This does not mean, however, that they did not try. Speakers’ ratings of the ease of their task were not particularly high; these, for instance, were low in comparison with ratings obtained after producing different tone heights in whisper on series of VCV stimuli that contained neither prosodic clashes nor lexical meaning (Heeren, 2014). We take this to reflect that speakers were aware of the difficulty of their task, and we see three factors contributing to the relatively weak coding of boundary tones. First, as was mentioned before, the linguistic setting within which the sentence function contrast was expressed made it challenging for speakers, asking them to produce both a lexical and a phrasal contrast through prosody and in close proximity. Evidence for this is found in the acoustic results that not only reflect the acoustic dimensions’ use for differentiating boundary tones, but also stress positions and, in particular, cases, vowel categories. Second, speakers vary in how well they convey intonation in whisper, which suggests that they may have different strategies for coding speech melody (see, e.g., Heeren et al., 2014). Considerable variation in realizations has also been seen in clear speech (Picheny et al., 1986; Ferguson and Kewley-Port, 2007), and it may complicate the task for listeners when they encounter multiple speakers in the same experiment. Moreover, the fact that we used twelve different speakers may in itself have made performance poorer than in earlier studies using less speaker variation. Multi-speaker versus single-speaker sets, for instance, negatively influenced listener performance in lexical tone (Lee et al., 2009) and spoken word recognition (Mullenix et al., 1989). Third, the good listener results found for stress classification suggest that speakers did not enhance the realization of boundary tones at the risk of changing lexical meaning. This in turn suggests that phrasal prosodic information was deemed subordinate here to lexical prosodic information. The high scores on the stress classification task furthermore showed that perception of whispered speech, as far as we were able to measure here, is not harder per se; linguistic contrasts that do not heavily rely on F0 are perceived as well as in normal speech (e.g., Lehiste, 1982), and they are also processed equally fast.

Contrary to earlier accounts that attributed pitch perception in whisper to formant change, and especially to F2 (e.g., Higashikawa and Minifie, 1999; Thomas, 1969), listeners
were able here to at least partly trace the sentence melody without changes in formant locations. This finding is in line with listeners’ ability to perceive boundary tones in whispered French when individual formants are left out of the speech spectrum, as they seem able to also exploit other spectral and non-spectral cues (Heeren and Lorenzi, 2014). Absence of formant shifts, however, seemed to negatively affect performance in comparison with a classification study on the same boundary tone contrast in the same language (Heeren and Van Heuven, 2009). The present results show that different types of (spectral) cues support intonation perception in whisper, but still allow for the explanation that spectral peaks are in fact the best listener cues.

Enhancement of precisely that cue that imparted compensatory information in whispered natural productions, led to improved perception, but the advantage remained small. It was mainly found for those speakers who received the lowest mean listener scores in experiment I, and for syllables that were stressed. We exaggerated here what our speakers did naturally, but there are alternative ways of effecting a change in CoG other than by tilting the spectrum as a whole, such as local shifts in formant locations, or changes in the relative prominence of formants. Based on the relatively poor performance in the present perception tests, the significance given to formant changes in the literature on whispered pitch together with the importance of spectral peaks in spectral pitch perception, we expect one of these other types of CoG manipulation to hold more effective cues for listeners, even if these are not naturally used by speakers in our setting. We plan to address this question in future research. Finally, support provided by non-spectral cues seems to be of importance as well, as the enhancement of spectral cues was most effective in stressed syllables. As we suggested before, this may be a consequence of the target syllable being sufficiently prominent for the boundary tone to be reliably audible only in the case that it is stressed. Alternatively, or perhaps in addition, the effort associated with the stress may act as a direct cue to the boundary tone, which would follow Gussenhoven’s (2002) Effort Code. According to that theory perceived effort may be associated with a higher pitch excursion. We therefore expect enhancement of non-spectral cues to further improve perception of the boundary tone contrast in whisper, as long as the extent of the change is not such that it alters lexical stress position, i.e., word meaning.

ACKNOWLEDGMENTS

This work was supported by Veni Grant No. 275-75-008 from the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Jos Pacilly for developing the scripts for acoustic analysis and Judith Varkevisser and Myrthe Wildeboer for testing subjects.

1As argued in Van Heuven (2014), none of these studies, however, provided conclusive evidence for the superiority of F0 over temporal cues.

2There is an imbalance in the numbers of mispronunciations between speech and whispered mispronunciations remaining undetected more often.

3Obvious boundary tone errors were not noticed in whisper, which—however—cannot exclude that boundary tone errors were missed by the first rater, because—as the live listeners confirm—it is harder to correctly perceive boundary tone realization in whisper.

4Inverse RT gave results closest to the normality requirement. Without transformation, for SF classification 35 out of 48 and for LS classification 41 out of 48 listener-speech mode combinations were not normally distributed. For the log transform the numbers were 17 out of 48 for SF and 22 out of 48 cases for LS.


