Coding pitch differences in voiceless fricatives: Whispered relative to normal speech

Willemijn F. L. Heeren[a)]
Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands

(Received 17 December 2013; revised 30 October 2015; accepted 30 October 2015; published online 4 December 2015)

Intonation can be perceived in whispered speech despite the absence of the fundamental frequency. In the past, acoustic correlates of pitch in whisper have been sought in vowel content, but, recently, studies of normal speech demonstrated correlates of intonation in consonants as well. This study examined how consonants may contribute to the coding of intonation in whispered relative to normal speech. The acoustic characteristics of whispered, voiceless fricatives /s/ and /f/, produced at different pitch targets (low, mid, high), were investigated and compared to corresponding normal speech productions to assess if whisper contained secondary or compensatory pitch correlates. Furthermore, listener sensitivity to fricative cues to pitch in whisper was established, also relative to normal speech. Consistent with recent studies, acoustic correlates of whispered and normal speech fricatives systematically varied with pitch target. Comparable findings across speech modes showed that acoustic correlates were secondary. Discrimination of vowel-fricative-vowel stimuli was less accurate and slower in whispered than normal speech, which is attributed to differences in acoustic cues available. Perception of fricatives presented without their vowel contexts, however, revealed comparable processing speeds and response accuracies between speech modes, supporting the finding that within fricatives, acoustic correlates of pitch are similar across speech modes.
 [http://dx.doi.org/10.1121/1.4936859]

[CHS] Pages: 3427–3438

I. INTRODUCTION

Prosody and, more specifically, lexical tone and sentence melody (intonation) can—to an extent—be perceived in whispered speech (e.g., Miller, 1961; Fónagy, 1969; Liu and Samuel, 2004; Kong and Zeng, 2006). Whisper, however, is produced without the quasi-periodic vibration of the vocal folds that generates the speaker's fundamental frequency (f0) (e.g., Monoson and Zemlin, 1984). This, together with its lower harmonics, is assumed to hold the main cue to intonation. In the few studies to date, acoustic correlates of intonation and pitch in whisper have been sought in vowel content. This work has shown that the lower formants, F1 and F2, can carry prosodic information in the absence of f0 (Higashikawa and Minifie, 1999; Heeren and Van Heuven, 2009), and that other contributing cues seem to be intensity (Meyer-Eppler, 1957; Denes, 1959), higher formants (Meyer-Eppler, 1957; Fónagy, 1969), and duration (Liu and Samuel, 2004). In recent years, studies of normal (i.e., phonated) speech have demonstrated acoustic correlates of intonation in voiceless consonants as well (Niebuhr, 2008, 2012), and listeners are sensitive to such information (Niebuhr, 2008; Kohler, 2011). In the present study, the ways in which consonants, rather than vowels, may contribute to the coding of pitch in whispered speech was examined.

The phonetic implementation of prosody affects characteristics of the segments, both vowels and consonants, over which the prosody is produced (cf. Kohler, 2012). Research on variation in consonant realization under the influence of prosody initially focused on articulation, which resulted in, for instance, the effect of articulatory strengthening: consonants are articulated more forcefully in initial than medial position, and when initiating stronger rather than weaker prosodic domains (e.g., Fougeron and Keating, 1997). Later on, acoustic consequences of these articulatory differences were studied, as well as their roles in speech processing. This work showed that not only vowel acoustics but also consonant properties vary with prosodic context (Cho and McQueen, 2005; Niebuhr, 2008; Kohler, 2011).

In normal, phonated speech the acoustic content of both stops and fricatives is influenced by prosodic factors. As for stop consonants, Niebuhr (2008) found that aspiration of utterance-final /t/ differed between two accent contours in its duration, intensity maximum, and in the location of the spectral peak above the lower spectral energy boundary, the "frequency… at which the first clear increase in energy is observed" (Niebuhr, 2008, p. 1254). In domain-initial position, the stops /t/ and /d/ showed longer closure durations with stronger prosodic boundaries (Cho and McQueen, 2005). Both stops also showed a difference in centre of gravity (CoG) with prosodic boundary depth, but the differences were not consistent across boundary depths. In that same study, the fricatives /s/ and /z/ also showed longer durations with stronger boundaries. Intensity was lower with stronger boundaries, which was most clearly found for /s/, but CoG

[a)]Also at: Leiden University Centre for Linguistics, Leiden University, Van Wijkplaats 4, 2311 BX Leiden, The Netherlands. Electronic mail: w.f.l.heeren@uu.nl

did not vary with prosodic boundary strength in domain-initial position. Recently, Niebuhr (2012) investigated the interaction of utterance-final changes in f0 and the realization of the voiceless fricatives /f/, /s/, /ʃ/ and /x/ in utterance-final position, placed directly after an f0 fall or f0 rise. In rising f0 contexts, intended as questions, voiceless fricatives had a higher CoG mean and range, and a higher intensity. However, when considering /s/ and /f/, the fricatives of interest in this study, intensity did not differ between high and low-f0 contexts for either fricative, mean CoG systematically differed between f0 contexts, and the CoG range only differed between contexts in /s/.

With respect to the perception of consonantal cues to intonation, Niebuhr (2008) found that the variation in utterance-final /t/ aspiration as a function of variation in intonation contours was sufficient to change listeners' ratings along semantic dimensions, such as questioning/non-questioning and certain/uncertain. In addition, Kohler (2011) showed that, in German, utterance-final /s/ frication alone was able to influence listeners' judgments, in a semantic differential framework comparable to that in Niebuhr (2008), by lowering the number of "question" interpretations when an /s/ produced in a low-f0 context was combined with an f0 contour associated with questions.

If consonant realizations are influenced by intonation in normal speech, they may also be in whisper. And, with such acoustic cues influencing perception in normal speech, they may also contribute to intonation perception in whisper. The first goal of the present investigation was to study the acoustic characteristics of whispered fricatives produced at different pitch targets (Sec. II). Assuming that the speaker's intention is to establish a certain pitch percept, the term pitch target here refers to the speaker's corresponding production goal, e.g., low or high. In normal speech, this would be mainly reached through changes in f0, but in whisper, other means are expected to be used. Perception of speech melody, or intonation, cannot simply be equated with perception of pitch, but the hypothesis that f0 is a main cue for either seems generally accepted. To incite whispering speakers to produce relatively large acoustic differences, stimuli were chosen that minimized semantic content through the use of nonsense words and that maximized pitch scaling, that is, the difference between utterances intended as low versus high. We assume that cues to pitch in such stimuli are likely to also function as cues to intonation in whispered running speech, but we do not claim that results obtained here directly translate to intonation in running speech and that the cues carry the same weight in either type of speech.

In addition, the acoustic correlates of pitch targets in whisper were compared to characteristics of fricatives produced in normal speech, and interpreted as secondary or compensatory cues to pitch. On the one hand, under the assumption that speech is a redundant signal, consonants may contain acoustic correlates secondary to f0 that code intonation in normal speech, and these might be present in comparable ways in whispered speech. On the other hand, given that speakers' possibilities for expressing intonation in whisper are restricted, it may be the case that the segments, including the consonants, are altered more or in different ways by whispering speakers to express intonation. The latter would be predicted under the hyper- and hypo-speech theory (Lindblom, 1990), which states that the speaker adapts the speech signal to the needs of the listener. Speakers, for instance, adapt their speech when addressing hearing impaired listeners or listeners in a noisy surrounding (cf. Uchanski, 2008). In the case of whispered speech, the listener may have higher needs as to how speakers convey intonation, prompting speakers to use more hyper-speech. The second goal of the present study was to investigate listeners' sensitivity to consonantal cues to pitch in whisper, also relative to the sensitivity to those cues in normal speech (Secs. III–V). To our knowledge, this is the first study to make this normal speech-whispered speech comparison for fricatives.

II. ACOUSTIC CORRELATES OF PITCH TARGETS IN FRICATIVES

Do the acoustic characteristics of the voiceless fricatives /s/ and /f/ systematically vary with the speaker's pitch target in whispered speech? If so, is that variation comparable to that observed in normal speech? The acoustics of stops and fricatives are influenced by prosodic factors in normal, phonated speech (Cho and McQueen, 2005; Niebuhr, 2008, 2012). Specific to fricatives, changes in duration, intensity, and CoG were found. In the present study, higher production targets were expected to show a higher CoG (e.g., Niebuhr, 2012). In addition, when the pitch target increases, speakers are expected to put in more effort, as in clear speech. This is associated with longer durations (e.g., Picheny *et al.*, 1986), an intensity increase in the 1–3 kHz range (Krause and Braida, 2004), and a less negative or even positive spectral slope (e.g., Glave and Rietveld, 1975; Sluijter and Van Heuven, 1996).

A. Materials

Nonsense vowel-consonant-vowel (VCV) structures were used in which both vowels were one of the three corner vowels (/i, a, u/) and the same within an item (V_iCV_i), and the consonant was a voiceless fricative, either an /s/ or an /f/. This resulted in the stimuli *isi*, *ifi*, *asa*, *afa*, *usu*, and *ufu*. To collect both whispered and normally spoken VCVs at different pitch targets (low, mid, high), speakers produced four tokens of a particular VCV in either a rising (low-mid-high-high) or a falling (high-mid-low-low) series. As an example of the task, two series of four sinusoidal signals corresponding to the musical notes C4-E4-G4-G4 and G4-E4-C4-C4, that is, with the first three notes three to four semitone (ST) steps apart, were each played once to the participants. Speakers were not specifically instructed to follow the example, but on average the range of their productions was 7.5 STs (see the Appendix). Measurements were taken from the first three VCVs only. The fourth was produced at the same pitch target as the third and had been added to prevent boundary effects (such as final lengthening) on the third token.

Twelve native speakers of Dutch took part in the recordings (six males and six females, with self-reported

normal-hearing), and each was paired up with a listener (three male speakers paired with male listeners, five male-female, four female-female pairs). The speaker-listener setup was intended to record listener-directed rather than read speech, given the higher ecological validity of the former. Speaker and listener could not see each other. The speaker was seated in a sound-treated booth and spoke into a Røde NTG-2 condenser microphone (Silverwater, Sydney, Australia) with "deadcat" windscreen placed about 8 cm from his/her mouth, and connected to an Edirol R-44 portable solid state recorder (Hamamatsu, Japan) (44.1 kHz, 24 bits, mono). The written production targets were presented to the speaker in pseudorandom order on a computer screen, and the intended pitch target change in the course of the series was indicated by visually shifting the subsequent written VCVs in either an upward or downward direction on screen. The listener was seated outside the recording booth wearing Sennheiser HD 414 SL headphones (Wedemark, Germany) listening to the speaker's productions with both ears. After each series, the listener indicated if it sounded as if it were rising or falling by hitting a key on a keyboard. The correctness of this judgment was immediately presented on the speaker's computer screen as listener feedback. After this, the next production series appeared on screen. The listener was not informed about the correctness of his/her responses, and feedback was not spoken so as not to compromise the recording quality.

Speaker and listener received written instructions, were given the opportunity to ask questions, and then filled out consent forms. Recording time was 3–5 min per speech mode, and the entire session including explanation and practice lasted about 30 min. Participants were paid a small amount for their efforts. Listeners labeled 98% of phonated 4-repetition series and 91% of whispered series correctly as either rising or falling (note that each speaker-listener pair was different). Using a six-point Likert scale (1 = very difficult, 6 = very easy) speakers rated the difficulty of their task directly after the recording of either speech mode. According to a Wilcoxon signed ranks test for paired samples, the task was judged more difficult in whisper (median = 4) than in normal speech (median = 6), $Z = -2.9$, $p = 0.004$. In spite of the two-point difference in medians between speech modes, both scores were in the upper half of the scale, suggesting that the task was perceived as doable in both speech modes.

In normal speech, two repetitions of each of the six series were recorded and in whisper, three repetitions of each were collected. The order in which speech modes were recorded was counterbalanced across subjects and four pseudorandom lists were used per speech mode. Data were transferred from the portable recorder onto a computer and each series was saved as a separate wave file (44.1 kHz, 24 bit). Each file was semi-automatically annotated at the phoneme, word, and series levels. Per item, e.g., an *asa* series, 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 (Boersma and Weenink, 2013). Next, automatic annotations were manually checked through visual inspection of each spectrogram, and if necessary by listening to the corresponding wave file. Mainly, the visual distinction between frication noise versus vowel formant structure in the spectrograms together with changes in intensity were used to place boundaries as precisely as possible, and in a consistent way. For acoustic analysis and perception tests, the second repetition of each recorded series was selected. If that was not of sufficient quality, due to the presence of voicing in a fricative or in a whispered vowel, or of non-speech sounds (tongue smack, etc.), one of the other repetitions was selected.

B. Acoustic measurements and analysis

All measurements were taken using Praat (Boersma and Weenink, 2013). The acoustic measures taken per fricative were (i) relative duration, that is, the percentage of C within the VCV, and over the full duration of the fricative (ii) mean intensity in dB, (iii) CoG in Hertz over the 0.05–8 kHz range, and (iv) spectral balance in decibels (Praat's averaging method: energy) comparing the difference in intensity between the 0.5–2 and 2–8 kHz frequency regions in each fricative's long term spectrum (LTAS in Praat): a larger difference was expected for high productions. The 8 kHz cutoff for the spectral measures defined a frequency region that contained the region of maximal energy for /s/. CoG was computed over a time-averaged spectrum over the full duration of each fricative, using eight 15-ms Hanning windows per fricative. Overlap between subsequent windows was over 50% for three out of 864 measurements with a maximum overlap of 56%. For acoustic analysis, 864 instances were available (12 speakers × 2 speech modes × 6 stimuli × 2 production series × 3 pitch targets), but as one /s/ was realized as a /z/, 863 instances remained.

Each acoustic correlate was modeled using linear mixed-effects models implemented in the *lme4* package (Bates *et al.*, 2012) in R (R Core Team, 2012). The fixed predictors were pitch target (low, mid, high), speech mode (normal, whisper), fricative (/s/, /f/), and vowel context (/a_a/, /i_i/, /u_u/). Random effects included in the model were for the individual speakers (12) and speaker-dependent differences in the fixed predictors. Of interest to the current research questions were simple effects of the predictor pitch target and interactions of pitch target by speech mode. A simple effect of pitch target would indicate a secondary cue, as the correlate would be available similarly in both normal speech and whisper. A speech mode by pitch target interaction could indicate a compensatory correlate in which the acoustic difference was only present in whisper, or was larger in whispered than in normal speech. The models were expected to also reflect intrinsic differences between the speech modes, and between the fricatives (in context), such as differences in intensity and CoG between /s/ and /f/ (e.g., Jongman *et al.*, 2000; Jesus and Shadle, 2002). These effects are presented in Table I and summarized in the discussion, but not extensively discussed in Sec. II C.

The base model contained the intercept as well as the stimulus properties represented by a fricative by vowel context interaction and the maximal random effects structure (Barr *et al.*, 2013). For each acoustic correlate, the optimal

TABLE I. Estimated fixed effects parameters, with standard errors in parentheses, for the base and extended mixed-effects models of relative duration, mean intensity, and CoG. Boldface indicates a significant effect[a] at $p < 0.05$.

	Relative duration				Mean intensity				CoG			
	Base		Extended		Base		Extended		Base		Extended	
	β	t	β	t	β	t	β	t	β	t	β	t
Intercept	**25.68 (1.02)**	**25.2**	**24.44 (1.04)**	**23.5**	**51.37 (1.02)**	**50.3**	**53.35 (1.31)**	**40.7**	**2012 (114)**	**17.6**	**2066 (260)**	**7.9**
Whisper			0.02 (1.10)	0.0			**−3.35 (1.18)**	**−2.8**			417 (237)	1.8
Pitch *Mid target*			0.16 (0.45)	0.4			−0.33 (0.45)	−0.7			**−292 (141)**	**−2.1**
Pitch *Low target*			**2.01 (0.55)**	**3.7**			**−1.32 (0.59)**	**−2.2**			−4.3 (228)	−1.8
Whisper: Mid target			0.10 (0.70)	0.2			−0.53 (0.63)	−0.8			13 (158)	0.1
Whisper: Low target			−0.97 (0.71)	−1.4			0.41 (0.75)	0.5			96 (216)	0.4
Fricative /s/	−0.48 (0.65)	−0.7	−0.43 (0.67)	−0.6	**10.51 (0.77)**	**13.7**	**8.76 (1.06)**	**8.2**	**3058 (239)**	**12.8**	**2258 (287)**	**7.9**
Vowel context /i_i/	0.58 (0.43)	1.4	0.39 (0.48)	0.8	**20.2 (0.39)**	**5.2**	**1.84 (0.42)**	**4.3**	**580 (109)**	**5.3**	**593 (111)**	**5.4**
Vowel context /u_u/	**2.72 (0.56)**	**4.9**	**3.08 (0.64)**	**4.8**	**4.06 (0.42)**	**9.6**	**3.24 (0.53)**	**6.1**	−153 (100)	−1.5	−151 (126)	−1.2
Fricative /s/: /i_i/	**3.03 (0.68)**	**4.4**	**1.91 (0.71)**	**2.7**	**−1.31 (0.57)**	**−2.3**	−0.62 (0.68)	−0.9	**−747 (182)**	**−4.1**	−344 (204)	−1.7
Fricative /s/: /u_u/	−1.50 (0.78)	−1.9	**−2.08 (0.84)**	**−2.5**	**−2.20 (0.58)**	**−3.8**	**−2.41 (0.63)**	**−3.9**	**−1249 (232)**	**−5.4**	**−892 (270)**	**−3.3**

[a]*p*-values were computed through a widely used method, which determines the degrees of freedom in the normal approximation procedure from the total number of data points. For comparison, if *p*-values were computed using the most conservative method (Hox, 2010), which uses the smallest number of second-level units (e.g., 12 speakers) to determine degrees of freedom, critical *t*-values would be 2.45 for the base models and 12.71 for the extended models.

model also included the speech mode by pitch target interaction, which assessed the main research question. Models were compared through likelihood ratio tests (Pinheiro and Bates, 2000).

C. Results

Pearson's r correlation coefficients were computed between the different measurements per speech mode and showed high correlations between CoG and spectral balance (whisper, $r = 0.84$; normal speech, $r = 0.75$). As visual exploration of the data confirmed this correspondence, only the CoG models are given below. In both speech modes, mean intensity and CoG showed a medium correlation (whisper, $r = 0.57$; normal speech, $r = 0.52$), and low correlations were found for relative duration and mean intensity, and relative duration and CoG ($r < 0.24$).

1. Relative duration of the fricative

Results are shown in Fig. 1 and Table I. The optimal model showed improvement over the base model [$\chi^2(5) = 7.6$, $p = 0.179$], which reflected that there was systematic variation in the fricatives' relative duration as a result of the experimental parameters pitch target and speech mode. The interaction of these predictors was not justified [$\chi^2(2) = 3.2$, $p = 0.20$], meaning that there was no difference between the speech modes. Across speech modes, low productions (26.9% of VCV duration) were longer than high ones (25.4%). As Table I shows, stimulus properties also influenced relative fricative duration, e.g., fricatives were longer in /u/-context (26.9%) than in /a/-context (24.8%).

2. Intensity

Results are given in Fig. 2 and Table I. The optimal model showed improvement over the base model [$\chi^2(5) = 12.0$, $p < 0.05$], reflecting that there was systematic variation in the fricatives' mean intensity with pitch target and speech mode. The interaction of these predictors was not justified [$\chi^2(2) = 3.6$, $p = 0.17$], showing that there was no difference in pitch target coding between the speech modes. Across speech modes, intensity was lower in low [56.1 dB, standard deviation (sd) = 6.8 dB] than high (57.2 dB, sd = 7.0 dB) productions and, in whisper, fricative intensity was lower (55.0 dB, sd = 7.4 dB) than in normal speech (58.4 dB, sd = 6.0 dB). Stimulus properties also led to intensity variation, for instance, reflecting that /s/ had a higher intensity than /f/.

3. CoG

Results are shown in Fig. 3 and Table I. The optimal model showed improvement over the base model [$\chi^2(5) = 15.3$, $p < 0.01$]; there was systematic variation in the fricatives' CoG as a function of the experimental parameters pitch target and speech mode. There was no difference in pitch target coding between the speech modes, as the interaction of these predictors was not justified [$\chi^2(2) = 0.3$, $p = 0.85$]. Across

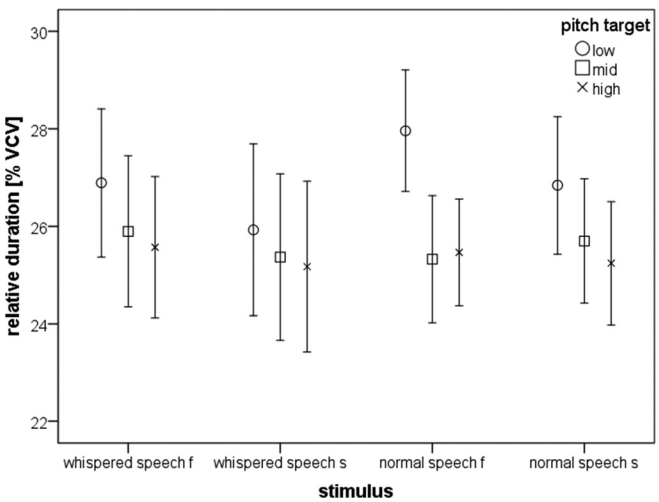

FIG. 1. Mean relative duration per fricative, speech mode, and pitch target. Error bars indicate the 95% confidence interval of the mean.

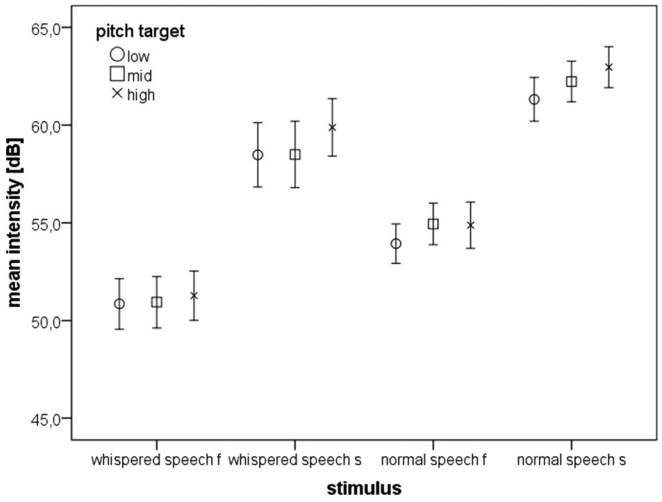

FIG. 2. Mean intensity (in dB) per fricative, speech mode, and pitch target. Error bars indicate the 95% confidence interval of the mean.

speech modes, CoG tended to be lower in low productions and higher in high productions (low, 2976 Hz; high, 3340 Hz), with mid productions in between (3055 Hz), but only the difference between high and mid became significant. Stimulus properties also influenced CoG, for instance, as that of /s/ was higher than that of /f/.

D. Discussion

The acoustic analysis was, in the first place, aimed at revealing acoustic correlates of pitch target in the fricatives /s/ and /f/ in whispered speech. For comparison, normally spoken fricatives from the same speakers were included to determine whether the acoustic correlates in whisper should be interpreted as secondary or compensatory. The analyses showed that acoustic correlates of pitch target were found for all measures and in highly comparable ways in the two speech modes. Therefore, correlates were secondary; none of the speech mode by pitch target interactions, which could have signaled compensatory behavior in whispered fricatives, was significant (see Table I).

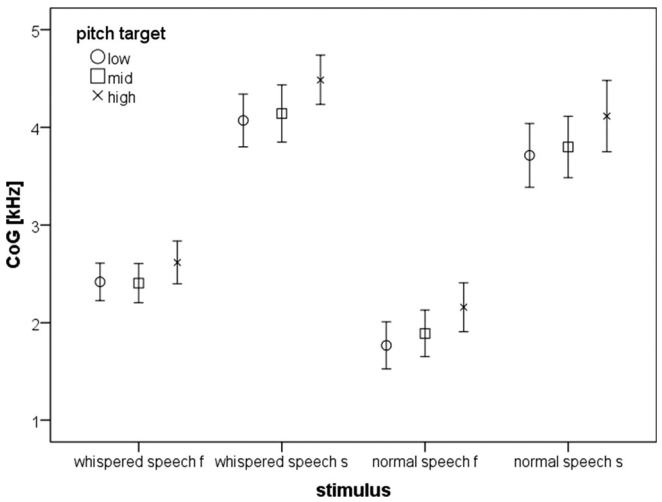

FIG. 3. Mean CoG (in kHz) per fricative, speech mode, and pitch target. Error bars indicate the 95% confidence interval of the mean.

Across speech modes, relative duration of the fricative was influenced by the pitch target; longer durations were found for lower productions. This seems to contradict the clear speech prediction of longer durations with higher, more effortful productions (e.g., Picheny *et al.*, 1986), but as VCV durations did lengthen with higher productions, vowel duration increased relative to fricative duration. As the absence of a speech mode simple effect shows, the fricatives were not systematically longer relative to the vowel in whispered than in normal speech. Across speech modes, relative durations were larger in /u/-contexts than /a/-contexts, which may be explained by less time being available for fricatives between intrinsically long relative to short vowels (e.g., Lehiste, 1970, pp. 18–19).

Intensity decreased with decreasing pitch target, and differences were on the order of 1 dB between low and high or high and mid productions. The size and direction of this effect are in line with results reported by Niebuhr (2012) on normal speech, even though differences in that study were not significant for /s/ and /f/. A 1 dB difference in intensity, as found in both studies, may not be a very informative cue, though, as it, for instance, is at the same level as fluctuations in intensity caused by the speaker changing his/her head direction and it, furthermore, fell well within the range of standard deviations obtained. Across speech modes, fricatives in the context of the vowel /a/ were produced with less intensity than in the context of /i/ or /u/, which may be due to a trade-off in the amount of energy available for producing the utterance as a whole. The CoG also showed an influence of pitch target, in both speech modes, with productions intended as higher showing a higher CoG than lower productions. This effect is comparable to that reported in Niebuhr (2012) for normal speech and was, in the current study, observed for both speech modes.

Acoustic correlates not only varied with pitch target, but also reflected intrinsic differences between the fricatives, consistent with reports in the literature. The CoG of /s/ was higher than that of /f/ (Jongman *et al.*, 2000; Jesus and Shadle, 2002), and intensity of /s/ was higher than that of /f/ (e.g., Behrens and Blumstein, 1988; Jongman *et al.*, 2000). CoG was influenced by vowel context, which may be due to co-articulation with the following vowel (Soli, 1981). Relative duration also varied with vowel context, in line with the finding of Jongman *et al.* (2000) that fricatives are shorter in the context of low than high vowels. When comparing fricatives between speech modes, the intensity effect was comparable to that in Jovičić and Šarić (2006).

Whispered, as well as normally spoken fricatives /s/ and /f/, were influenced by pitch target and acoustic correlates seemed to be of a secondary nature, that is, comparable between the speech modes. In the following sections, the contribution of fricative content to pitch target discrimination in whisper relative to that in normal speech was investigated. This is assessed in Sec. IV by (i) presenting VCV stimulus pairs in which only fricatives varied in pitch target between members of a pair, whereas vowel contexts' pitch targets were kept constant (the fricatives-in-context task), and by (ii) presenting pairs of fricatives in isolation, that is, excised from the vowel context (the fricatives-in-isolation

task). Before sensitivity to cues in fricatives was determined, perception of the stimulus items as recorded was assessed to establish baseline performance in whispered versus normal speech for our VCV stimuli (Sec. III).

III. BASELINE SENSITIVITY IN WHISPERED AND PHONATED SPEECH

A. Method

Baseline sensitivity to differences in pitch targets between VCV stimuli was established in both speech modes using the discrimination task that was also used in the rest of this study. One-step V_iCV_i-V_iCV_i pairs (low-mid/mid-high, in both directions) from each of the 12 speakers were presented to listeners in a two-interval forced choice (2IFC) discrimination task with a 500 ms inter-stimulus interval, and without feedback. Reaction times (RTs) were measured as well. All stimulus pairs were taken from within a production series and were, therefore, within speakers, resulting in 144 trials per speech mode [12 speakers × 2 fricatives × 3 vowel contexts × 2 comparisons (up/down)]. Five-ms fade in and fade out was applied to all stimuli in this and subsequent experiments to prevent audible clicks at the beginning or end of the stimuli.

Ten native Dutch listeners (three males, seven females) aged 19–27 yr (mean = 23 yr), from whom informed consent was obtained, participated in the experiment. Subjects were hearing-screened using an Oscilla USB-330 audiometer (Aarhus, Denmark) to ensure they were normal hearing at octave frequencies between 0.125 and 8 kHz and all were right-handed. Listeners received written instructions and six practice trials with feedback before the main test, which was intended to acquaint them with the procedure. The listeners' task was to indicate if the second member of the V_iCV_i-V_iCV_i pair was higher or lower than the first, by pressing one of two keys on the keyboard using their index fingers, and to do so as fast as possible. On screen, the two response options ("higher"/"lower") were shown on the same side as the corresponding response key on the keyboard. Four pseudorandom lists were used for each speech mode and listeners heard different lists for each speech mode. Speech modes and response keys were counterbalanced across subjects. Each task lasted 12–15 min and the entire session, including hearing screening, instruction and pauses, remained under 45 min. Subjects received a small fee for their efforts.

B. Analysis, results, and discussion

To investigate if discrimination performance was above chance level and if it varied with speech mode, response accuracy was modeled using mixed-effects logistic regression and RTs with linear mixed-effects models, as implemented in the *lmer*( ) function from the *lme4* package (Bates *et al.*, 2012) in R (R Core Team, 2012). The fixed categorical predictors were speech mode (2), fricative (2), and vowel context (3). Random effects included in the model were for the individual listeners (10), individual speakers (12), and speaker- and/or listener-dependent differences in the fixed predictors. The base model contained the intercept, as well as the stimulus properties represented by a fricative by vowel context interaction, and the maximal random effects structure justified by the model (Barr *et al.*, 2013). The optimal model also included the predictor speech mode. Only RTs on correct trials were included, and outliers were removed below a minimum response time of 300 ms and above a maximum response time of two standard deviations beyond the mean (computed per listener, per speech mode). Thus, 3.9% of the RT data were excluded. RTs were transformed to meet the normality requirement (as assessed by a Kolmogorov–Smirnov test), by raising them to the power of −0.5. The accuracy model was based on 2880 responses and the RT model on 2419 responses.

The optimal model for response accuracy had a significantly positive intercept, confirming that discrimination performance was above chance level [= 50%, $\beta = 3.55$, standard error (SE) = 0.35, $Z = 10.1$, $p < 0.001$]. Listeners were more accurate in normal (94.7%) than in whispered speech (80.2%, $\beta = -1.86$, SE = 0.32, $Z = -5.9$, $p < 0.001$). The inclusion of the predictor speech mode improved the model, $\chi^2(1) = 11.1$, $p < 0.001$. Across speech modes, performance was 1% more accurate on stimulus pairs containing /f/ than /s/ ($\beta = -0.43$, SE = 0.21, $Z = -2.1$, $p < 0.05$).

The optimal model for transformed RTs performed better than the base model, $\chi^2(1) = 12.5$, $p < 0.001$. It had a significant intercept ($\beta = 3.6 \times 10^{-2}$, SE = 1.13×10^{-3}, $t = 31.7$, $p < 0.001$) and, consistent with the discrimination results, showed that listeners were faster to respond in normal speech (860 ms, sd = 341 ms) than in whisper (1090 ms, sd = 434 ms), ($\beta = -4.1 \times 10^{-3}$, SE = 8.3×10^{-4}, $t = -4.9$, $p < 0.001$). RTs did not vary as a function of the stimulus properties fricative, vowel context, or their interactions [$|t| \leq 1.4$, n.s. (= not significant)]. With a longer duration for whispered than normally spoken items, the information needed to decide on pitch target may be expected to arrive somewhat later in whisper, but as the actual durational difference between speech modes was only 30 ms on average (normal speech, 546 ms; whisper, 575 ms), this cannot fully explain the 230-ms processing delay. This suggests that pitch processing in whisper might be slower than in normal speech due to differences in acoustic cues available to listeners.

Pitch target differences were perceived rather well in whispered speech, but not as well as in normal speech. This is in line with results reported in earlier work (e.g., Higashikawa *et al.*, 1996; Heeren and Van Heuven, 2009; Heeren and Lorenzi, 2014). Responses were only minimally affected by the nature of the fricative, /s/ versus /f/, and the 1%-difference was in the opposite direction than predicted: stimuli containing /f/ did slightly better than those with /s/. This may mean that the fricatives carry comparable amounts of information as to pitch target, or that the vowels are the main carriers of information, overriding cues that might reside in the fricatives. The literature, together with the acoustic analyses presented above, suggests that there may be perceptual cues to pitch target in the fricatives. The following perception experiment investigated if those cues can be used when fricatives are heard in isolation and when fricatives are heard in the context of vowels.

IV. PITCH PERCEPTION FROM CUES IN WHISPERED FRICATIVES

Given results on normal speech (Niebuhr, 2008; Kohler, 2011) and systematic differences found in the acoustic analysis, it was expected that listeners would be sensitive to these subtle acoustic differences conveyed by consonants produced at different pitch targets. Based on the similar correlates found in the acoustic analysis, comparable performance was expected in whispered and normal speech. Should our acoustic analysis have missed important listener cues, performance on whisper may be different from that on normal speech.

The spectrum of /s/ contains a region of high prominence in the 4–5 kHz range, whereas the spectrum of /f/ does not show localized high energy, meaning that the spectrum is flatter as a whole (e.g., Behrens and Blumstein, 1988). If a shift in spectral peak location is more easily perceived as a cue to pitch change than a non-localized spectral change, there may be a listener advantage for /s/ relative to /f/. A localized spectral peak might be a better cue to spectral pitch, as it maximally excites a specific region of the basilar membrane.

A. Method and materials

The same 2IFC discrimination task was used as in the baseline experiment (see Sec. III A for details) and RTs were measured. Two versions were designed: (i) the fricatives-in-context task, and (ii) the fricatives-in-isolation task. For (i), stimulus pairs were made in which only the pitch target of the fricative changed, but not that of its vowel context. The latter remained constant across the two members of a stimulus pair. Each speaker had produced a rising and a falling series. Cross-splicing was used to generate all VCV stimuli for this task, which was done by taking the vowel context produced in one series at a particular pitch target, e.g., low /a_a/ from the rising series, and splicing in a fricative taken from the other series, in this example, /f/ or /s/ from the falling one. In this way, the fricative was always substituted to prevent listeners from basing their discrimination responses on non-manipulated versus manipulated stimuli. The fricative in one member of the pair had been produced at the same pitch target as the vowels used (e.g., both fricative and vowel context came from a low production), whereas the fricative in the other member of the pair had been produced at a different target (e.g., a high fricative in between low vowels), thus, introducing the pitch target difference between members of the stimulus pair through fricative content only. Splicing was always done within the same speech mode.

In the fricatives-in-context task, two-step differences (high-low, low-high) between fricatives in a constant and, therefore, non-informative, vowel context were used.[1] There were 144 stimuli per speech mode (12 speakers × 2 fricatives × 3 vowel contexts × 2 production orders). With two possible orders of stimuli within a discrimination pair (AB, BA), the total of 288 stimulus pairs was divided over four presentation lists of 72 pairs in which each pitch target difference (high-low, low-high) occurred only once for each speaker, per fricative, and per vowel context. As the acoustic analysis has shown, mean consonant intensity varied with pitch target, even though mean differences were only in the range of 1 dB within a fricative. Intensity was normalized to minimize between-trial and between-speech mode differences by setting the recordings from each speaker plus speech mode combination to 65 dB [root-mean-square (rms) = 0.036], which corresponded to the minimum intensity of whispered items after scaling peaks to the maximum intensity range (using Praat's "Scale peaks…" function).

In the fricatives-in-isolation task, the vowel context was omitted. Stimuli were made by excising the fricatives from their vowel context and smoothing the edges of the cut-out intervals over 5 ms. Eliminating the vowel context yielded 144 stimuli per speech mode (12 speakers × 2 fricatives × 3 pitch targets × 2 directions). Four lists were created, each containing 48 two-step pairs (high-low, low-high) and 48 one-step pairs (high-mid, mid-low, and vice versa). In each list, speakers, fricatives, and directions were evenly distributed, whereas there was one pitch target pair, e.g., high-low, per speaker by fricative by direction combination. Intensity was normalized as before.

B. Subjects and procedure

Twenty native Dutch listeners (6 males, 14 females) aged 18–28 yr (mean = 21 yr) participated in the experiment (informed consent was obtained). Subjects were hearing-screened (see Sec. III A for details) and all were right-handed. In a blocked design, subjects participated in both tasks: fricatives-in-context, and fricatives-in-isolation. Test order and response buttons were counterbalanced across listeners. Each subtask lasted 6–8 min and started with six practice trials with feedback. Total test duration, including instruction and pauses, remained under 45 min. Subjects received a small fee for their participation.

C. Analysis, results, and discussion

To investigate if discrimination performance was above chance level, and if accuracy and RTs depended on speech mode and step size, responses were modeled using mixed-effects logistic regression (for accuracy) and linear mixed-effects models (for RT), implemented in the *lmer*( ) function from the *lme4* package (Bates *et al.*, 2012) in R. The fixed predictors were speech mode (2), fricative (2), and, for the in-isolation task, step size (one-step, two-step) or, for the in-context task, vowel context (3). Random effects included in the model were for the individual listeners (20), individual speakers (12), and speaker- and/or listener-dependent differences in the fixed predictors. The base model contained the intercept as well as the maximal random effects structure justified by the model (Barr *et al.*, 2013). Each extended model also included the fixed predictors.

RTs were processed to remove outliers, see Sec. III B for details (4.8% of the fricative-in-isolation RT data were excluded, 6% of the fricative-in-context data), and transformed to meet the normality requirement (as assessed by a Kolmogorov–Smirnov test), which was done by taking inverse RT values (1/RT). The accuracy model was based on 2880 responses, and the RT model was based on 2251

responses for the in-isolation task and on 1292 responses for the in-context task.

1. Discrimination of fricatives in a constant, non-informative vowel context

Results are shown in Fig. 4. The extended response accuracy model showed improvement over the base model, $\chi^2(4) = 12.0$, $p = 0.017$. There was a simple effect of speech mode, with more correct responses in whispered (51.8%) than in normal speech (43.8%, $\beta = 0.33$, SE = 0.09, $Z = 3.7$, $p < 0.001$), but the non-significant intercept showed that performance was not above chance-level ($\beta = -0.18$, SE = 0.01, $Z = -1.8$, n.s.). As opposed to when original vowel contexts surrounded the fricatives, there seemed to be no usable information in the stimulus pairs containing constant vowel contexts. Across speech modes, performance did not vary with fricative ($\beta = -0.02$, SE = 0.08, $Z = -0.3$, n.s.) or with vowel context (/i_i/: $\beta = -0.08$, SE = 0.09, $Z = -0.9$, n.s.; /u_u/: $\beta = -0.13$, SE = 0.10, $Z = -1.3$, n.s.). Adding an interaction of fricative by vowel context did not improve the extended model, $\chi^2(2) = 1.91$, $p = 0.38$.

Responses to phonated stimuli, in fact, were below chance level, which may suggest a response bias. Chi square analyses were run to explore this hypothesis, but in both speech modes, responses were equally distributed over response categories [normal speech, χ^2 (2, $N = 1440$) = 0.011, n.s.; whisper, χ^2 (2, $N = 1440$) = 0.278, n.s.]. An alternative explanation is that in the normal speech stimuli there was information in the vowels or vowel-fricative transitions that guided responses in the opposite direction. This hypothesis is addressed in the control perception experiment presented in Sec. V.

For RTs, the extended model did not show improvement over the base model, $\chi^2(4) = 3.77$, $p = 0.44$. RTs did not differ between the speech modes ($\beta = 4.50 \times 10^{-5}$, SE = 3.31×10^{-5}, $t = 1.4$, n.s.) with means of 1336 ms (sd = 543 ms) on whispered trials, and of 1481 ms (sd = 752 ms) on normally spoken ones. RTs also did not depend on the nature of the fricative ($\beta = 5.95 \times 10^{-6}$, SE = 1.42×10^{-5}, $t = 0.4$, n.s.) or the vowel context (/i_i/: $\beta = -2.73 \times 10^{-5}$, SE = 2.05×10^{-5}, $t = 1.3$, n.s.; /u_u/: $\beta = -1.04 \times 10^{-5}$, SE = 1.95×10^{-5}, $t = 0.5$, n.s.). Only the intercept was significant ($\beta = 7.90 \times 10^{-4}$, SE = 4.94×10^{-5}, $t = 16.0$, $p < 0.001$). Adding an interaction of fricative by vowel context did not improve the extended model, $\chi^2(4) = 3.77$, $p = 0.44$.

2. Discrimination of fricatives in isolation

Results are shown in Fig. 5. The extended model for the fricatives-in-isolation task did not show improvement over the base model, $\chi^2(3) = 0.91$, $p = 0.82$. None of the predictors showed a simple effect as evidenced by comparable listener scores for the different factor levels, with mean percentages of 60.8% correct for whispered fricatives and 61.4% correct for normal speech ones ($\beta = -0.05$, SE = 0.13, $Z = 0.1$, n.s.), means of 60.6% correct for one-step stimulus pairs, and 61.6% for two-step pairs ($\beta = 0.06$, SE = 0.08, $Z = 0.9$, n.s.), and means of 59.6% for /s/ and 62.6% for /f/ ($\beta = -0.14$, SE = 0.19, $Z = -0.7$, n.s.). The significantly positive intercept showed that responses were above chance level (= 50%, $\beta = 0.54$, SE = 0.16, $Z = 3.4$, $p < 0.001$).

Above-chance performance indicated that speakers conveyed pitch target information in their fricatives, but with means around 60% correct; this information was not very helpful for listeners. The drop in performance from nearly 80% correct on the whispered VCV utterances (see Sec. III) to around 60% on fricatives only (Mann-Whitney U test for independent samples, $Z = -6.3$, $p < 0.001$) supports the earlier suggestion that vowels seem to be the main carriers of information for pitch perception in whisper as they are in normal speech. Speakers did not convey systematically clearer pitch target information in fricative pairs that were taken from utterances at two-step differences than one-step differences. There was no evidence that changes in a fricative with a localized spectral peak, /s/, were easier to perceive than changes in a fricative with a relatively flat spectrum, /f/.

As for the RTs, the extended model showed no improvement over the base model, $\chi^2(3) = 6.05$, $p = 0.11$. The

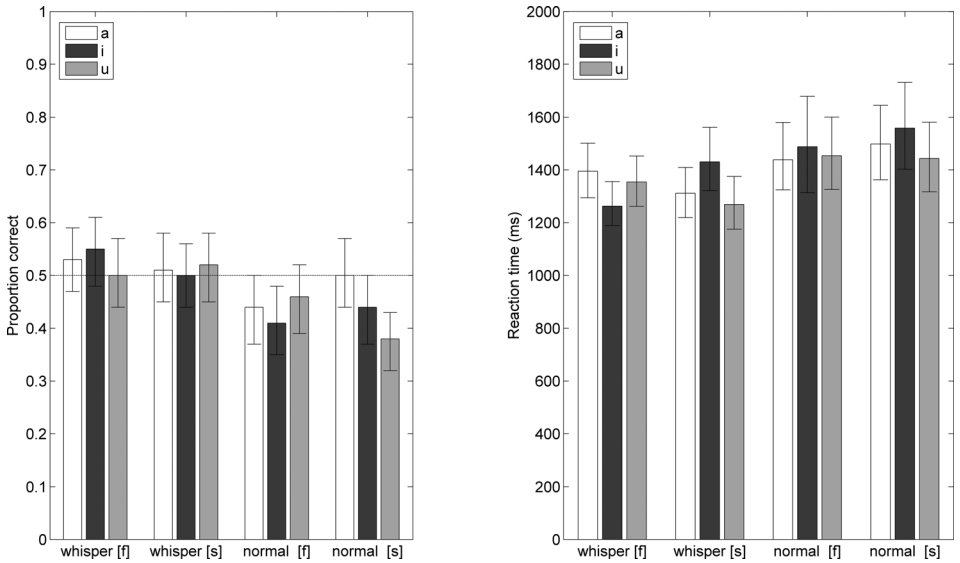

FIG. 4. Mean proportion of correct responses with a reference line at chance level (left), and RTs in milliseconds (right) for the fricative-in-context discrimination task. Error bars indicate the 95% confidence interval of the mean. Clustered bars represent the different vowel contexts.

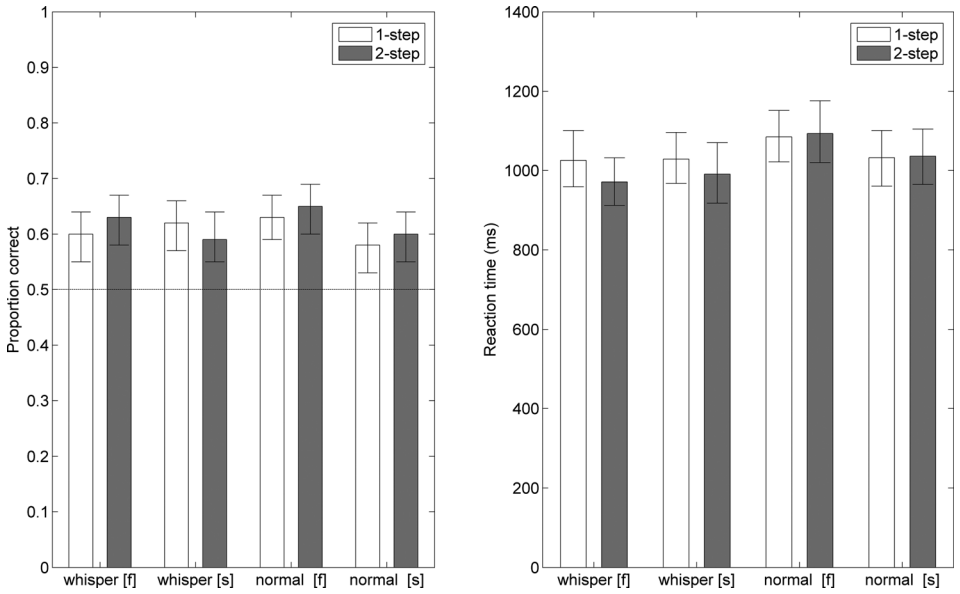

FIG. 5. Mean proportion of correct responses with a reference line at chance level (left), and RTs in milliseconds (right) for the fricative-in-isolation discrimination task. Error bars indicate the 95% confidence interval of the mean. Clustered bars represent the different step sizes.

intercept was significant ($\beta = 0.001$, SE $= 7.7 \times 10^{-5}$, $t = 14.2$, $p < 0.001$), but there were no effects as a function of speech mode ($\beta = 3.24 \times 10^{-5}$, SE $= 2.65 \times 10^{-5}$, $t = 1.2$, n.s.), step size ($\beta = 3.49 \times 10^{-5}$, SE $= 1.87 \times 10^{-5}$, $t = 1.9$, n.s.) or fricative ($\beta = 2.38 \times 10^{-5}$, SE $= 1.42 \times 10^{-5}$, $t = 1.7$, n.s.). The mean RT to whispered trials was 1004 ms (sd = 598 ms) and to normally spoken ones, it was 1063 ms (sd = 616 ms).

The fact that listener performance was comparable across speech modes provided no evidence for the existence of compensatory information to pitch in the whispered fricatives. This is in line with the findings of the acoustic analysis, in which only secondary cues to pitch targets were found. When the fricatives were presented in isolation, their cues to pitch could be used by listeners, but when the same fricatives were presented in a non-informative vowel context, listeners were unable to reliably discriminate pitch target differences.

V. CONTROL EXPERIMENT: INFORMATION IN THE VOWEL CONTEXT

In the fricatives-in-context task, performance on normal speech stimuli was below chance level at about 44%. This result may have been obtained by chance, but it may also reflect that the phonated vowel context interfered with pitch target cues in normally spoken fricatives in a way that listeners were given misleading information. An example of such information would be small differences in the f0 of the first relative to the second vowel in a stimulus, which led listeners to perceive a periodicity pitch difference that went counter to the shift expressed mainly spectrally in the fricatives of the stimulus pair. This was evaluated by presenting fricatives that had been produced in normal speech in the context of whispered vowels. It was predicted that performance would no longer be below chance level and also that it would be comparable to that of the whispered stimuli, as the cues available to listeners were being made more comparable between speech modes.

A. Method

To select materials with comparable information in the fricatives in both speech modes, fricative stimulus pairs were selected from those six speakers (three males, three females) who received above-chance responses in the fricative-in-isolation task (67.8%) and who, averaged across speakers, got comparable listener scores in normal and whispered speech on that task. For this speaker subset, performance on the fricative-in-context task had been above chance level at 55.2% for the whispered items ($N = 720$, $p = [1/2]$, $Z = 2.7$, $p < 0.01$), whereas it had been below chance level at 43.4% for the phonated ones ($N = 720$, $p = [1/2]$, $Z = -3.5$, $p < 0.01$). All normally spoken, voiceless fricatives were presented in a whispered vowel context: this hybrid condition consisted of whispered vowel contexts within which a normally spoken fricative from the same vowel context was spliced. Whispered stimuli served as a reference condition. The mean intensity of each normally spoken, voiceless consonant was equalized to the intensity of the corresponding whispered consonant. Two pseudorandom lists were used containing all 144 stimuli per speech mode (6 speakers $\times$ 2 fricatives $\times$ 3 vowel contexts $\times$ 2 production orders $\times$ 2 stimulus exchange orders). Stimulus intensity was normalized as before.

The same discrimination task was used, with within-speaker VCV pairs at two-step (high-low, low-high) differences. RTs were not measured. The two types of stimuli, whispered and hybrid, were presented in separate tasks, the order of which was counterbalanced across subjects. There were ten native Dutch subjects (four males, six females, aged 20–54 yr, mean = 26 yr) who were hearing-screened (see Sec. III A for details), and from whom informed consent was obtained. Each task lasted 12–15 min and the entire session including hearing screening, instructions, and practice lasted about 45 min. Subjects received a small fee for their efforts.

B. Analysis, results, and discussion

To investigate if discrimination performance was at or above chance level rather than below, response accuracy

was modeled using mixed-effects logistic regression (see Sec. IV C). The fixed predictors were speech condition (hybrid, whispered), fricative (2), and vowel context (3). Random effects included in the model were for the individual listeners (10), individual speakers (6), and speaker- and/or listener-dependent differences in the fixed predictors. The base model contained the intercept as well as the maximal random effects structure justified by the model (Barr *et al.*, 2013). The extended model also included the fixed predictors ($N = 2880$).

The extended model reflected that performance was comparable across conditions ($\beta = 0.12$, SE $= 0.08$, $Z = 1.6$, n.s.) and at chance level ($\beta = 0.11$, SE $= 0.10$, $Z = 1.1$, n.s.). None of the predictors reflecting stimulus properties was significant ($|Z| \leq 1.5$, $p \geq 0.13$), and the extended model did not improve relative to the base model, $\chi^2(6) = 6.84$, $p = 0.34$. This shows that with scores of 52.3% for the hybrid and 55.2% correct for the whispered stimuli, both predictions were met. The whisper score replicates the 56% correct on the same stimulus subset by 20 listeners in the main experiment. The absence of a speech mode effect supports the explanation that secondary rather than compensatory cues are present in whispered fricatives, and that the voiced vowel context somehow influenced responses in the main experiment. Similar effects were found in early studies on the role of non-f0 cues in the perception of lexical tones and intonation in speech (Denes, 1959; Abramson, 1972). These studies showed that in "vocoder whisper," i.e., vocoded normal speech in which the periodic excitation signal was replaced by a noise source, tonal information remained identifiable, whereas this was not the case when f0 was set to a monotone.

VI. GENERAL DISCUSSION

Cues to pitch in whisper seem to be mainly carried by vowels, not consonants, given the drop in performance from 80% correct discrimination on whispered VCVs to around 60% on fricatives in isolation. Still, the voiceless fricative consonants contain cues to pitch in whisper as well as in normal speech, which is in line with studies demonstrating that prosodic contexts may change consonants (Cho and McQueen, 2005; Niebuhr, 2008; Kohler, 2011).

In the context of constant, non-informative vowels, mean discrimination performance did not differ from chance level. The information present in the fricatives, as demonstrated by the fricatives-in-isolation task, was not or could not be used by the listeners. This negative effect of vowel context may have had different reasons. First, when vowels are present, listeners may expect them to contribute information for pitch estimation. This could invite listeners to divide their attention between the consonants and vowels or even mainly focus on the vowels, thus, reducing performance in the in-context task. Evidence for listeners' preference for cues in specific parts of the signal can, for instance, be found in Repp (1977) and Fujimura *et al.* (1978), who found that perception of medial consonants in VCVs is dominated by the CV transition rather than the VC transition. Second, a difference in duration between the stimulus types, fricatives-in-isolation versus fricatives-in-context, resulted in a longer interval between the to-be-discriminated fricatives in the in-context case. As a result, representations could not be compared equally fast in the two tasks, and it is known that performance decreases with an increasing inter-stimulus interval (Pisoni, 1973). Part of the performance difference may be attributable to this. Third, to avoid an unfair comparison between stimulus conditions, all stimuli were generated through cross-splicing. As this manipulation is likely to have reduced the naturalness of the transitions between segments, it may also have had an influence on performance.

In line with the results from the acoustic analyses, the perception experiments showed no evidence for compensatory cues in whisper. Pitch discrimination of whispered and normal speech fricatives was comparable and, furthermore, suggested that the consonantal cues were not very strong. It seems fair to state that also in earlier work (Kohler, 2011; Niebuhr, 2012), the influence of consonantal cues on the perception of intonation in normal speech was not very large. In Kohler (2011), for instance, an utterance-final /s/ that had been produced at a low pitch target, as part of an assertion, made an utterance sound less questioning than the original high-target /s/ that had been produced as part of a question. The manipulated utterance, however, was still perceived as a question. On the in-context task, we found that fricative information was not used by listeners, whereas the earlier studies had found effects of fricative content using linguistic phrases. This difference in results may be explained by the fact that in both earlier studies the consonants were in final rather than medial position and, therefore, were not followed by a vowel that might have interfered with the evaluation of pitch information present in the fricative. This interference hypothesis is supported by the fact that our listeners were able to use such information from the same fricatives in the in-isolation task, in which there was no vowel following the target segment, as had been the case in the earlier studies. As for the usability of fricative cues to pitch perception in running speech, this would suggest that only in utterance-final position, as coda consonant in a sentence's last syllable, or maybe in phrase-final position, a voiceless fricative may make a moderate contribution to pitch perception. Even though we found no evidence for a consonantal contribution to pitch perception in medial position, results from Mixdorff and Niebuhr (2013) suggest that these fricatives may still contribute to perceived f0 continuity and, thus, to perception of prominence.

The absence of a difference in performance between one-step and two-step pairs of isolated fricatives suggests that speakers were unable to systematically convey larger versus smaller differences in pitch targets. This may mean that speakers were unsuccessful at their attempts to do so, or that they did not try to express a pitch target difference through their fricatives (but only through their vowels). In the former case, the acoustic changes with pitch target that we observed were insufficiently differentiating the different targets for listeners, for instance, through large standard deviations, as in the intensity results, and relatively small differences between the means, as may be said of the CoG

results. In the latter case, assuming that speakers put in more effort when producing high-target VCVs, the added effort per se may explain (part of) the production and perception results. Rather than expressing pitch targets through voiceless fricatives, speakers may in fact be putting additional effort into those segments because of the prosodic structures they appear in, as in articulatory strengthening (e.g., Fougeron and Keating, 1997). Speech that is produced with additional effort, but that is not necessarily intended as higher, shows a number of changes in acoustics that are comparable to the ones found here, including changes in duration (Picheny *et al.*, 1986), a more positive spectral balance (Jesus and Shadle, 2002; Krause and Braida, 2004), and a higher CoG (Maniwa *et al.*, 2009). As for perception, effort may be associated with larger movements in intonation as captured by the Effort Code, which "concerns inferred pitch excursion size, not height of pitch *per se*" (Gussenhoven, 2002, p. 67). This would make added effort a sensible strategy for speakers when trying to convey a change in pitch target without the use of f0.

The speech stimuli used in the present investigation are different from natural running speech. A main motivation for using nonsense CVCs at different pitch targets was to get speakers to produce relatively large acoustic differences by not asking them to attend to other layers of linguistic information at the same time. Acoustic results obtained in Heeren and Van Heuven (2014) suggest that whispering speakers whose production task is made challenging from a linguistic point of view may only show small acoustic differences between experimental conditions. In that study, speakers were asked to produce an intonation contour consisting of both a nuclear pitch accent and a high or low boundary tone within the same disyllabic word or even within the same syllable of that word. Speakers did not produce systematic differences in target vowels' formant frequencies, whereas formant shifts are considered the main cue for whispered pitch perception, as was found in studies that used linguistically simpler stimuli (e.g., Higashikawa and Minifie, 1999; Heeren and Van Heuven, 2009). As mentioned in the introduction, cues to pitch in whispered VCVs may also function as cues to intonation in whispered running speech, though possibly not in precisely the same way. To understand how exactly the acoustic correlates found here may function in running speech, further research is needed.

In comparison with pitch accent sizes in running speech, the average distance of 7.5 ST between speakers' low and high pitch targets (as measured in normal speech) is close to the pitch span found in Dutch (cf. Kraayeveld, 1997, pp. 164–166). This may be taken as evidence that the setup aimed at eliciting maximal acoustic differences was successful, and we assume that our speakers aimed for comparable pitch differences while whispering. The facts that the high-low differences produced by speakers were relatively large, and perceived well by the listeners who attended the recordings (see Sec. II A), suggest that speakers succeeded in getting their message across. In doing so, the role for the fricative seemed restricted, and in spite of the absence of f0, speakers did not enhance fricative cues in whisper to help their listeners.

VII. CONCLUSION

Acoustic characteristics of the whispered, voiceless fricatives /s/ and /f/ produced at different pitch targets were studied and compared to characteristics of the same fricatives produced in normal speech to assess if the former contained secondary or compensatory correlates of pitch. In line with recent studies, we found changes in the acoustic characteristics of these consonants with differences in pitch target intended by the speaker. These acoustic correlates of pitch targets were of a secondary nature, as they are comparable in normal and whispered speech.

Furthermore, listener sensitivity to consonantal cues to pitch in whisper was investigated, also relative to that in normal speech. Results showed that in VCV stimuli, discrimination was less accurate and processing was slower in whispered than in normal speech, which is presumably due to the differences in acoustic cues available to listeners. When looking at fricatives in isolation, however, processing speed and response accuracy for pitch information were comparable between speech modes, suggesting that fricative cues are secondary. This was furthermore consistent with the comparability of acoustic correlates across the speech modes.

ACKNOWLEDGMENTS

The author would like to thank Jos Pacilly for writing the scripts for acoustic analysis, Hugo Quené for help in the statistical analyses, Vincent van Heuven for helpful comments on an earlier version of this manuscript, and Judith Varkevisser and Myrthe Wildeboer for testing subjects. This work was supported by a Veni grant from the Netherlands Organisation for Scientific Research (NWO) awarded to the author.

APPENDIX

Table II summarizes the speakers' use of f0 in the normal speech recordings.

TABLE II. Mean f0 and its standard deviation (in Hertz) for the speakers' high, mid, and low target vowels, along with the difference between high and mid, and mid and low targets (in semitones). (Note: Gender is denoted by "m" for male and "f" for female.)

Speaker	Gender	High	Mid	Low	High-mid	Mid-low
1	m	108.4 (11.0)	96.1 (4.0)	92.5 (10.3)	2.1	0.7
2	f	343.3 (17.7)	279.2 (35.9)	210.3 (17.8)	3.6	4.9
3	m	132.3 (7.6)	111.0 (24.6)	89.3 (9.8)	3.0	3.8
4	m	171.8 (19.2)	142.7 (3.6)	113.5 (3.0)	3.2	5.6
5	m	183.0 (9.8)	140.4 (10.5)	103.3 (5.9)	4.6	5.3
6	m	176.5 (3.7)	147.3 (9.7)	117.0 (8.7)	3.1	4.0
7	f	283.4 (35.9)	243.8 (27.0)	207.0 (8.2)	2.6	2.8
8	m	143.7 (5.0)	118.5 (3.9)	101.2 (7.1)	3.3	2.7
9	f	390.0 (18.8)	280.4 (30.3)	197.3 (9.2)	5.7	6.1
10	f	228.6 (25.4)	197.0 (4.6)	160.5 (5.9)	2.6	3.5
11	f	319.4 (58.0)	261.8 (23.0)	209.9 (27.3)	3.4	3.8
12	f	354.9 (4.9)	272.8 (11.9)	204.8 (5.6)	4.6	5.0

[1]Stimulus pairs with one-step differences between fricatives were piloted. The 288 stimulus pairs per speech mode [12 speakers × 2 fricatives × 3 vowels × 2 production orders × 2 pitch target differences (low-mid, mid-high)], discriminated by six normal-hearing female listeners (aged 20–34 yr, including the author), showed that performance across fricatives, vowel contexts, and speech modes was at chance level ($N = 1728$, $p = 1/2$, $Z = -1.4$), with mean scores of 51% correct for whispered and 46% correct for phonated speech.

Abramson, A. S. (**1972**). "Tonal experiments with whispered Thai," in *Papers on Linguistics and Phonetics to the Memory of Pierre Delattre*, edited by A. Valdman (Mouton, The Hague), pp. 31–44.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**, 255–278.

Bates, D., Maechler, M., and Bolker, B. (**2012**). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. Available at http://CRAN.R-project.org/package=lme4 (Last viewed 1 October 2015).

Behrens, S. J., and Blumstein, S. E. (**1988**). "Acoustic characteristics of English voiceless fricatives: A descriptive analysis," J. Phon. **16**, 295–298.

Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer (version 5.3.42) [computer program]," http://www.praat.org/ (Last viewed 2 March 2013).

Cho, T., and McQueen, J. M. (**2005**). "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," J. Phon. **33**, 121–157.

Denes, P. (**1959**). "A preliminary investigation of certain aspects of intonation," Lang. Speech **2**, 106–122.

Fónagy, J. (**1969**). "Accent et intonation dans la parole chuchotée" ("Accent and intonation in whispered speech"), Phonetica **20**, 177–192.

Fougeron, C., and Keating, P. A. (**1997**). "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.

Fujimura, O., Macchi, M. J., and Streeter, L. A. (**1978**). "Perception of stop consonants with conflicting transitional cues: A cross-linguistic study," Lang. Speech **21**, 337–346.

Glave, R. D., and Rietveld, A. C. M. (**1975**). "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?," J. Acoust. Soc. Am. **58**, 875–879.

Gussenhoven, C. (**2002**). "Intonation and biology," in *Liber Amicorum Bernard Bichakjian* (*Festschrift for Bernard Bichakjian*), edited by H. Jakobs and L. Wetzels (Shaker, Maastricht, The Netherlands), pp. 59–82.

Heeren, W. F. L., and Lorenzi, C. (**2014**). "Prosody perception in normal and whispered French," J. Acoust. Soc. Am. **135**, 2026–2040.

Heeren, W. F. L., and Van Heuven, V. J. (**2009**). "Perception and production of boundary tones in whispered Dutch," in *Proc. Interspeech 2009*, Brighton, UK, pp. 2411–2414.

Heeren, W. F. L., and Van Heuven, V. J. (**2014**). "The interaction of lexical and phrasal prosody in whispered speech," J. Acoust. Soc. Am. **136**, 3272–3289.

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

Higashikawa, M., Nakai, K., Sakakura, A., and Takahashi, H. (**1996**). "Perceived pitch of whispered vowels—Relationship with formant frequencies: A preliminary study," J. Voice **10**, 155–158.

Hox, J. J. (**2010**). *Multilevel Analysis: Techniques and applications*, 2nd ed. (Routledge, New York), pp. 1–392.

Jesus, L. M. T., and Shadle, C. H. (**2002**). "A parametric study of the spectral characteristics of European Portuguese fricatives," J. Phon. **30**, 437–464.

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.

Jovičić, S. T., and Šarić, Z. (**2006**). "Acoustic analysis of consonants in whispered speech," J. Voice **22**, 263–274.

Kohler, K. J. (**2011**). "Communicative functions integrate segments in prosodies and prosodies in segments," Phonetica **68**, 26–56.

Kohler, K. J. (**2012**). "Editorial. Bridging the segment-prosody divide in speech production and perception," Phonetica **69**, 5–6.

Kong, Y.-Y., and Zeng, F.-G. (**2006**). "Temporal and spectral cues in Mandarin tone recognition," J. Acoust. Soc. Am. **120**, 2830–2840.

Kraayeveld, H. (**1997**). "Idiosyncrasy in prosody. Speaker and speaker group identification in Dutch using melodic and temporal information," Ph.D. dissertation, Radboud University Nijmegen, pp. 164–166.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge, MA), pp. 18–19.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H & H theory," in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–439.

Liu, S., and Samuel, A. G. (**2004**). "Perception of Mandarin lexical tones when F0 is neutralized," Lang. Speech **47**, 109–138.

Maniwa, K., Jongman, A., and Wade, T. (**2009**). "Acoustic characteristics of clearly spoken English fricatives," J. Acoust. Soc. Am. **125**, 3962–3973.

Meyer-Eppler, W. (**1957**). "Realization of prosodic features in whispered speech," J. Acoust. Soc. Am. **29**, 104–106.

Miller, J. D. (**1961**). "Word tone recognition in Vietnamese whispered speech," Word **17**, 11–15.

Mixdorff, H., and Niebuhr, O. (**2013**). "The influence of F0 contour continuity on prominence perception," in *Proc. Interspeech 2013*, Lyon, France, pp. 230–234.

Monoson, P., and Zemlin, W. R. (**1984**). "Quantitative study of whisper," Folia Phoniatr. **36**, 53–65.

Niebuhr, O. (**2008**). "Coding of intonational meanings beyond F0: Evidence from utterance-final /t/ aspiration in German," J. Acoust. Soc. Am. **124**, 1252–1263.

Niebuhr, O. (**2012**). "At the edge of intonation: The interplay of utterance-final F0 movements and voiceless fricative sounds," Phonetica **69**, 7–27.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Pinheiro, J. C., and Bates, D. M. (**2000**). *Mixed-Effects Models in S and S-PLUS* (Springer, New York), pp. 82–96.

Pisoni, D. B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percept. Psychophys. **13**, 253–260.

R Core Team (**2012**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (Last viewed 1 October 2015).

Repp, B. H. (**1977**). "Perceptual integration and selective attention in speech perception: Further experiments on intervocalic stop consonants," Haskins Laboratories Status Reports on Speech Research SR-49, pp. 37–69.

Sluijter, A. M. C., and Van Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress," J. Acoust. Soc. Am. **100**, 2471–2485.

Soli, S. D. (**1981**). "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation," J. Acoust. Soc. Am. **70**, 976–984.

Uchanski, R. M. (**2008**). "Clear speech," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 207–235.