CAN FORMANT SHIFTS AND EFFORT CUES ENHANCE BOUNDARY TONE PERCEPTION IN WHISPERED SPEECH?

Willemijn F. L. Heeren

Leiden University Centre for Linguistics, Leiden University, The Netherlands
Utrecht Institute of Linguistics OTS, Utrecht University, The Netherlands
w.f.l.heeren@{hum.leidenuniv.nl;uu.nl}

ABSTRACT

Whispered speech holds cues to speech melody, in spite of the absence of F0. Shifts in the locations of formant peaks have been forwarded as a main cue. Whispering speakers, however, may convey high versus low boundary tones signalling questions versus statements without shifting their formants. Would the addition of formant shifts enhance these natural productions and improve question/statement classification in whisper? Moreover, multiple acoustic correlates tend to vary with pitch or intonation conditions in whispered speech, and may function as listener cues. Here, an attempt was made to better understand the function of one of these 'secondary' cues: intensity. Results show that formant shifts may improve performance, but not dramatically, and that intensity seems more useful when coding increased effort than when being higher across the board to compensate for reduced audibility in whisper.

Keywords: speech perception, whispered speech, cues to intonation, cue enhancement

1. INTRODUCTION

Whispered speech holds cues to speech melody, in spite of the absence of F0, e.g., [1][2][3], and shifts in the locations of formant peaks been forwarded as the main cue to whispered pitch perception, e.g., [1][4][5]. Not all studies into acoustic correlates of whispered intonation have, however, found that speakers systematically shift their formants. In [6], for instance, speakers' productions instead showed systematic changes in the tilts and centres of gravity of the spectra of vowels on which high versus low (H% vs. L%) boundary tones were expressed. Listeners still classified those utterances as statements versus questions around 60% correct, and did so using prosodic information alone.

The authors attributed the absence of formant shifts to the linguistic complexity of the stimuli, in which both a nuclear accent and a boundary tone were produced in close proximity, that is on the same disyllabic word. This case of 'tonal crowding' was thought to challenge speakers: a restricted set of acoustic dimensions was available for expressing segmental, lexical and supra-segmental information at the same time. This was confirmed by an acoustic analysis. If tonal crowding may indeed prevent speakers from producing the supposedly main cue to whispered intonation, i.e. formant shifts, listeners may still benefit from the addition of that acoustic correlate. In this study, the first question was if sentence function (question/statement) recognition in whisper could be improved by enhancing natural question and statement productions by adding formant shifts.

Often, multiple acoustic correlates systematically vary with pitch or intonation conditions in whispered speech, including intensity, e.g., [4][7][8]. An attempt was made to better understand the function of this particular 'secondary' cue, as it is thought to help to explain a response bias observed in [6]. The boundary tone aligns to the end of an utterance, and therefore mainly to the final syllable. If the utterance-final syllable also carried the nuclear accent, questions and statements (H% vs. L% boundary tones) were classified comparably well, and above chance-level. But when the pre-final syllable carried the nuclear accent, listeners tended to classify the utterances as statements. Assuming that 'statement' is the default response, prosodic information seemed to be used less if it occurred on a less prominent final syllable.

The question is how the benefit of stressed, utterance-final syllables can be explained. It would be an auditory advantage if the presence of stress creates a target syllable that is sufficiently prominent for the boundary tone to be heard reliably. Whispered speech generally has less intensity than normal speech does. It would be a linguistic advantage if the effort associated with stress, here implemented as additional intensity, acts as a direct cue to the boundary tone, following [9]'s Effort Code. The spectral tilt changes that speakers produced, i.e. changes in the distribution of energy across the spectrum, are also associated with effort changes, e.g., [10][11][12]. The second question therefore was: do listeners need stimuli to be louder, or to more clearly express differences in effort, for boundary tones to be reliably classified?

2. EXPERIMENT 1

This experiment was intended to evaluate if the addition of formant shifts to cue high vs. low boundary tones is helpful for listeners, and if enhancing audibility of the syllable on which the boundary tone lands improves performance. In a classification task listeners indicated if they perceived a disyllabic target word as question or statement. They did so in two conditions: (1) spectral peaks in the final-syllable vowel of each item were shifted upward (for Q) or downward (for S) by 8%, and (2) in addition to the spectral manipulation, the utterance-final vowel's intensity was increased by 3 dB.

2.1. Recording and manipulation of the materials

The materials recorded in [6] were used, that is four Dutch, disyllabic lexical stress minimal pairs: 'ca·non/ ka'non 'canon/ cannon' (/kanɔn/), 'Ser·visch/ ser'vies 'Serbian/ crockery set' (/sɛrvis/), 'Pla·to /pla'teau 'Plato/ plateau' (/plato/), and 'voor·naam/ voor'naam 'first name/ respectable' (/vornam/). These were recorded from 12 native speakers of Dutch (6 male) in neutral carrier sentences "He said…" (Dutch: 'Hij zei …'). Orthographically, sentences ended either in a full stop (to elicit a low boundary tone L%) or a question mark (to elicit a high boundary tone H%), and forced the nuclear accent onto the target word in final position, thus establishing prosodic crowding. Recordings were made in a sound-treated booth at Leiden University using an Edirol R-44 portable recorder and Røde NTG-2 condenser microphone with 'deadcat' windscreen (44.1 kHz, 24 bits).

Stimuli were presented to the speaker one by one and in written form on a computer screen, in a pseudo-random order. To prompt the speaker to use listener-directed speech, for each speaker, a different listener was present, who provided live feedback on whether an utterance was perceived as a statement or question. The listener was seated outside the booth in a silent room, wearing Sennheiser HD 414 SL headphones, and used a keyboard to classify each of the speaker's utterances by pressing one of two dedicated keys. The correctness of this response was immediately visually presented to the speaker.

Participants received written instructions, and completed a short practice session with different minimal pairs. Both normal and whispered speech were recorded, but only the 192 whispered items are used here (4 minimal pairs × 2 sentence functions × 12 speakers). To create the two stimulus conditions the final vowel of each disyllabic target word was manipulated in Praat [13]. Upward and downward shifts of spectral peaks were established by stretching or shrinking final-vowel spectra by 8%, and correcting for durational change. To enhance audibility in one condition, the intensity of all final vowels was increased by 3 dB.

2.2. Participants and procedure

There were 16 listeners (aged 18-26, 8 males) who were hearing-screened to have normal hearing at octave frequencies between 0.125 and 8 kHz. Informed consent was obtained and they received a small fee for their voluntary participation.

Each of the 192 items was presented once to each listener in a blocked design over conditions. Half of the subjects heard the first half of the materials in the formant shift (F-shift) condition, and the second half of the materials in the formant shift + intensity change (F-shift&I-change) condition. The other half of the subjects listened to the complementary sets. The set of materials was halved by including only one boundary tone realization, either statement or question, per speaker and per target word in each half. Hearing screening, explanation, practice and testing were completed within 45 minutes.

2.3. Analysis, results and discussion

The outcome variable response accuracy was modelled as a function of the fixed predictors Manipulation (F-shift, F-shift&I-change), Sentence Function (question, statement) and Stress Position (initial, final) using mixed-effects logistic regression, implemented in the *lmer()* function from the lme4 package [14] in R [15]. The base model was an empty model containing only the maximal random effects structure justified by the model [16]. The optimal model included the predictor Manipulation to assess an effect of manipulation condition and an interaction of the predictors Sentence Function and Stress Position to assess the response bias. Models were compared using likelihood ratio tests [17].

Figure 1 shows listener responses and Table 1 gives the results of the fixed part of the extended model. This model showed improvement over the base model ($\chi^2(4) = 15.9$, $p = .003$), and the interaction in the model was justified ($\chi^2(1) = 15.6$, $p < .001$). The significant intercept indicated that responses on average were above chance level (= 50%). There was no effect of Manipulation, meaning that the final-syllable intensity increase did not help listeners (F-shift: 62.4%; F-shift & I-change: 64.7%). And relative to results reported in [6], performance did not increase by much.

Stress Position and Sentence Function interacted: listeners showed a preference for classifying initial-stress words as statements, and final-stress words as questions. Including the three-way interaction of Manipulation × Sentence Function × Stress Position did not improve the model ($\chi^2(3) = 1.1$, $p = .77$); the response bias did not reduce or alter with either manipulation, and was comparable to that found before, [6].

Figure 1: Classification accuracy by condition. In each panel, sentence function scores are given for initial-stress and final-stress words separately. Error bars indicate the 95% CI of the mean.

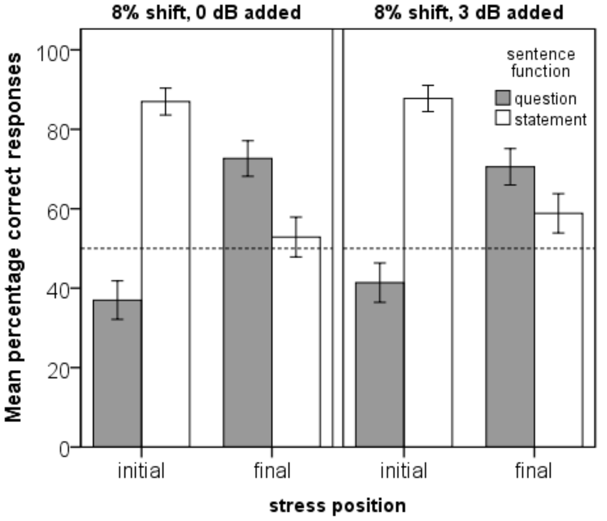

Table 1: Fixed effects parameter estimates of the extended model for Experiment 1, N = 3072.

Fixed effects	*β (SE)*	*Z*	*p*
Intercept	−0.527 (0.186)	−2.82	.005
Manipulation	0.110 (0.098)	1.12	.263
StressPosFinal	1.510 (0.261)	5.79	< .001
SentFunc S	2.841 (0.472)	6.02	< .001
StressPosFinal : SentFunc S	−3.605 (0.506)	−7.13	< .001

Final-stress words were responded to correctly slightly more often than initial-stress words (63.7% vs. 63.3%), and statements got more correct responses than questions (71.6% vs. 55.4%), which has been observed before in similar tasks using whispered speech, e.g., [7][8].

These results provide no evidence for the audibility hypothesis: more intensity in the final syllable does not reduce the imbalance in responses to initial stress words. Alternatively, the +3 dB increase in final syllables was not sufficient to enhance perception of the boundary tone. A larger intensity increase was not implemented, however, as pilot testing showed that stress position perception, that is the interpretation of word meaning, then risked being affected.

3. EXPERIMENT 2

If extra intensity in the utterance-final syllable would contribute to coding effort, rather than support audibility, an intensity increase would only be helpful in questions. Therefore, in this experiment the intensity increase was only applied to questions, not statements.

The same classification task was used. To enhance the sentence function contrast in one condition, the final-syllable vowel of each item was altered by shifting spectral peaks upward (for Q) or downward (for S) by 8%, and instead of raising intensity in all stimuli, this was only done in questions. In the second condition, the un-altered recordings were presented for comparison.

3.1. Method

There were 16 participants (aged 19-26, 6 males) who were hearing-screened to have normal hearing at octave frequencies between 0.125 and 8 kHz. Informed consent was obtained and they received a small fee for their voluntary participation. The procedure was the same as before.

In the manipulated condition, all stimuli had 8% formant shifts and a +3 dB intensity increase in final vowels of questions only. This condition was compared with non-altered materials, where the latter had been processed using the same script as the manipulated stimuli but with a ±0% shift and a 0 dB intensity change. This was done to exclude an effect of stimulus processing on perception.

3.2. Analysis, results and discussion

The outcome variable response accuracy was modelled as a function of the fixed predictors Manipulation (No-change, F-shift & I-change-in-Q), Sentence Function (2) and Stress Position (2) using mixed-effects logistic regression, as before (see section 2.3.). The base model was an empty model containing only the maximal random effects structure justified by the model. The optimal model included the predictor Manipulation and an interaction of the predictors Sentence Function and Stress Position to assess the response bias.

Figure 2 shows listener responses and Table 2 gives results for the fixed part of the extended model. This showed improvement over the base model ($\chi^2(4) = 29.2$, $p < .001$), and the interaction in the model was justified ($\chi^2(1) = 10.7$, $p = .0001$). The significant intercept showed that performance was above chance level (= 50%). More importantly, listeners did better with manipulated materials (64.5%) than with the un-altered ones (57.0%). This suggests that enhancement of the sentence function contrast supports perception.

Figure 2: Classification accuracy by condition. In each panel, sentence function scores are given for initial-stress and final-stress words separately. Error bars indicate the 95% CI of the mean.

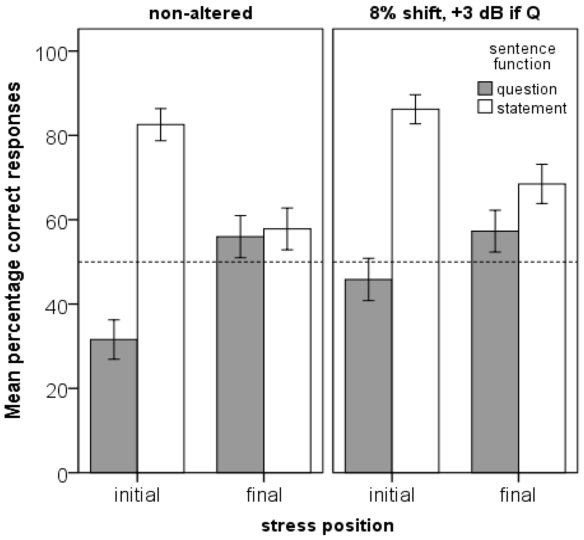

Table 2: Fixed effects parameter estimates of the extended model for Experiment 2, N = 3072.

Fixed effects	*β (SE)*	*Z*	*p*
Intercept	−0.666 (0.101)	−6.59	< .001
Manipulation	0.404 (0.095)	4.27	< .001
StressPosFinal	0.777 (0.220)	3.53	< .001
SentFunc S	2.523 (0.329)	7.67	< .001
StressPosFinal : SentFunc S	−2.217 (0.568)	−3.91	< .001

Across conditions, initial-stress words received somewhat more correct responses (61.6%) than final-stress words (59.9%), and statements got more correct responses (73.8%) than questions (47.7%). Inclusion of a three-way interaction of Manipulation × Sentence Function × Stress Position marginally improved the model ($\chi^2(3) = 6.8$, $p = .079$), confirming that there may be a tendency for the response bias to be smaller in the manipulated condition. In fact, the F-shift & I-change-in-Q condition was the only one in which questions on initial-stress words were *not* classified below chance level ($N = 384$, p = ½, $Z = -1.5$, $p = .11$). This also indicates that the bias is reduced in the case that intensity was meant to contribute to effort coding.

4. DISCUSSION AND CONCLUSION

Can formant shifts, assumed to be a main cue to intonation in whisper, help perception in a case where speakers do not produce them themselves. In this tonal crowding setting, in which both lexical stress realized as a nuclear accent and a boundary tone landed in close proximity, listeners furthermore showed a bias. In utterance-final, disyllabic words, boundary tones were only reliably heard when the word had final stress. The question was how this can be explained: is the tone-bearing syllable not audible enough if unstressed, or would listeners be helped by higher intensity expressing added effort?

With 8% changes in formant peak locations, listeners classified statements vs. questions about 63% correct in experiment 1 and slightly better in experiment 2. This does not provide compelling evidence that for these 'tonal crowding' stimuli the addition of a cue that listeners do not produce themselves is very helpful. Even though a significant improvement was found in experiment 2, mean performance does not exceed performance on the original recordings by much (see also [6]). Moreover, in [6] enhancement of the naturally-produced spectral tilt contrast yielded a comparable performance improvement of a few percentage points. In [18], however, a clear advantage was found for changes in spectral peak locations relative to changes in spectral tilt as cues for discriminating pitch differences in whisper. For now, it is tentatively concluded that the better cues may vary with linguistic environment and also with task, and that the supposedly best cue may not add information in all cases.

The experiments presented above were not designed to evaluate the question if the formant shift per se was sufficient for improved performance relative to the un-altered recordings. It is therefore not possible to make definitive claims about how much of the improvement in experiment 2 should be attributed to the formant shifts and how much to the intensity increase in questions only. A comparison of both experiments' results, however, suggests that the main contribution came from the former.

When listeners do not receive sufficient evidence for interpreting a whispered utterance as a question, they interpret it as a statement, e.g., [7][8], consistent with the more frequently occurring sentence function. A complication is that there is no one-to-one correspondence between H% and questions, and L% and statements. Earlier, boundary tones were classified above chance level when occurring on stressed syllables, but not on unstressed syllables. This investigation into the weak intensity of the latter syllable type as an explanation for the bias showed that enhancing audibility did not affect performance, but that enhancing intensity in questions only, that is effort, did. Only in the latter case, the bias was reduced, but not resolved.

5. ACKNOWLEDGEMENT

This work was supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO).

6. REFERENCES

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

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

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

[4] Meyer–Eppler, W. 1957. Realization of prosodic features in whispered speech. *J. Acoust. Soc. Am.* 19, 104–106.

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

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

[7] Fónagy, J. 1969. Accent et intonation dans la parole chuchotée [accent and intonation in whispered speech]. *Phonetica* 20, 177–192.

[8] Heeren, W. F. L., Van Heuven, V. J. J. P. 2009. Perception and production of boundary tones in whispered Dutch. *Proc. Interspeech* Brighton UK, 2411–2414.

[9] Gussenhoven, C. 2002. Intonation and biology, In: H. Jakobs, L. Wetzels (eds.), *Liber Amicorum Bernard Bichakjian*. Maastricht: Shaker, 59–82.

[10] Glave, R. D., 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.

[11] Gauffin, J., Sundberg, J. 1989. Spectral correlates of glottal voice source waveform characteristics. *J. Speech Hear. Res.* 32, 556–565.

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

[13] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341–345.

[14] Bates, D., Maechler, M., Bolker, B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4

[15] R Core Team 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[16] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278.

[17] Pinheiro, J. C., Bates, D. M. 2000. *Mixed effects models in S and S-plus*. New York: Springer, 82–96.

[18] Heeren, W. F. L., Pacilly, J. J. A. 2014. Local versus global spectral change as pitch cue in whispered speech. *Proc. 6th International Conference on Tone and Intonation in Europe*, 3–4.