A PERCEPTUAL AND ACOUSTIC STUDY OF MELODY IN WHISPERED CZECH WORDS

Adléta Hanžlová, Tomáš Bořil

Institute of Phonetics, Charles University in Prague
adleta.hanzlova@gmail.com, tomas.boril@ff.cuni.cz

ABSTRACT

The perception of melody in speech depends mainly on the fundamental frequency (f_0) which reflects vocal fold oscillation speed. Whisper is defined by the absence of phonation and therefore the lack of f_0. Intended melody in whisper, however, seems to be discernible regardless.

This paper presents a perception experiment assessing the discernibility of melody in whispered Czech words and words sung in whisper, which proved that melody in whisper in certain cases can in fact be discerned, along with an acoustical analysis of the effect of intended melody in whisper on formant frequencies, formant to formant ratios, center of gravity (CoG) and spectral slope. The parameters affected by intended melody in whispered speech were F2 and CoG of stop-band filtered signal with main formant bandwidths removed. In words sung in whisper, the affected parameters were F2, F3, F2:F1 and F3:F2 ratios, CoG and spectral slope.

Keywords: whisper, melody, absence of phonation, acoustic correlates of intonation, pitch perception

1. INTRODUCTION

Melody is an integral suprasegmental aspect of speech and is mainly dependent on the fundamental frequency (f_0) which reflects vocal fold oscillation. Whisper is defined by the absence of phonation and therefore the lack of f_0.

Notwithstanding claims denying the discernibility of intended melody in whisper (see polemics in [1, 2, 3]), melody has been repeatedly shown to be an aspect perceived in whisper despite the lack of f_0 [1, 3, 4, 5, 6]. Experiments using whispered material from different languages have linked perceived melody in whisper to formant frequencies [7, 8, 9, 10]. Perception experiments [5] and analysis by synthesis [10] suggest that rather than formants alone, the acoustical cues to melody in whisper may be formant-to-formant ratios. Other acoustical properties shown to be linked to melody in whisper are spectral slope [5] and center of gravity (CoG) [11] which changes with intended melody more noticeably in whisper than in modal speech.

The intelligibility of both segmental and suprasegmental aspects of speech in whisper does not have to be based in the same mechanisms as in modal phonation, as can be seen in the worsened abilities of automatic speech recognition models trained on modal phonation when given recordings of whispered speech [12]. Melody in whisper may be discernible on the basis of secondary correlates similar to modal phonation, or by means of compensatory correlates not as prominent in modal phonation [6].

In the last decade, melody in whisper has been studied on French [5] and Dutch [6] words and isolated syllables read by native speakers of Dutch [11]. Recent research focuses on synthesis and automatic recognition of whisper [13, 14], the dominant language of this field being English. Studies of whisper in Czech have so far focused mainly on voicing [15, 16, 17].

This study focuses on the perception and acoustics of melody in whisper in a Czech language environment. A perception experiment was conducted to assess the discernibility of melody in whispered words and words sung in whisper without engaging the vocal cords. An acoustical analysis followed, inspecting the interaction of CoG, spectral slope and formants with intended melody.

2. METHOD

2.1. Material

3 Czech, mainly onomatopoeic, words with CVːCVː and 1 with CVːCV structure repeating the same syllable ([baːba, jɛːjɛː, laːlaː, joːjoː]) were chosen as target words for recording speech material.

For speech, every target word was set in the sentence “Řekl [target] anebo [modified target],” where modified target stands for the target word with two short vowels. Each sentence was realized with 4 different melodic contours as either a statement or question (did he say [target] or [modified target] /

he said [target] or [modified target]). Speakers were asked to repeat the sentences in a shadowing task. Template recordings were in modal phonation with melodic contours manipulated in Praat [18] for the f_0 of vowels in target words to match frequencies of tones to form musical intervals as shown in figure 1.

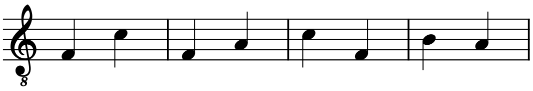

Figure 1: Musical intervals melodic contours of template target words were manipulated to match.

For words sung in whisper, only two target words ([laːlaː, joːjoː]) were used, but each was realized with 8 different melodic contours matched to musical intervals depicted in Figure 2. During recording, the speakers heard through headphones a piano track made in GarageBand [19] playing the target intervals with a metronome. Each interval was played twice, the speakers first listened and then sung along with the track.

Recordings were made in a sound-proof booth using a condenser microphone. Both types of material were recorded first in modal phonation and then in whisper. The templates stayed the same for both types of phonation. Speech material was recorded first, after a short break, the recording resumed with singing. The order of recorded items was randomized for each speaker.

4 female speakers aged 20–24 participated in the recording. All reported having basic musical education and experience with solo or choral singing. Each speaker was recorded in one session which including a break lasted around 25 minutes.

Target words were then extracted from all recordings, labeled using the *align interval* option in Praat and phone boundaries manually corrected according to [20]. Only recordings of whispered target words were used in the following perception experiment and acoustical analysis.

2.2. Perception experiment

A 2AFC perception experiment was conducted online using PsyToolkit software [21, 22] to assess the discernibility of melody in whisper.

Some recorded target words showed signs of vocal fold engagement and were left out from the perception experiment. This elimination resulted in 54 whispered speech and 53 whisper-sung stimuli. Each of these two sets of stimuli was put in a separate experiment. However, both setups were the same and they were embedded in one experiment session, starting with words sung in whisper and continuing with whispered speech. To each set of stimuli, filler stimuli were added to motivate the listeners. For whispered speech, 12 stimuli, preliminarily judged as having recognizable melody, with the lower of the two target musical tones low-pass filtered at 6300 Hz (where visible formant structure ended in the spectrogram) were used to create filler stimuli. For whisper-sung stimuli, 9 additional stimuli recorded during a preliminary session with prominent spectral changes were added.

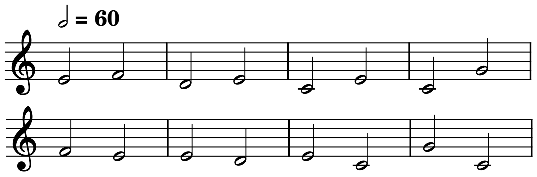

Figure 2: Musical intervals played as template when recording sung and whisper-sung words.

In both experiments, listeners were asked to determine whether they hear a fall or rise in melody by pressing a corresponding button on the screen. Each stimulus was preceded by a beep and could be replayed once. The order of stimuli in each experiment was randomized for each participant. Each experiment begun with a training session (comprising of 6 of the stimuli) and included a short pause in the middle.

1 Slovak and 32 Czech voluntary respondents (13 male) aged 17–63 (median 24) participated in the experiments. They were advised to wear headphones and complete the experiments in a quiet room. The whole experiment session lasted 13–29 minutes (median 19) based on each respondent's pace.

Results of these experiments were analyzed in R [23]. Success rates were calculated as the mean of individual respondents' correct answer rates for each experiment in general as well as for subcategories within them (intended melody and pitch difference). Using binomial tests included in Hmisc package [24], confidence intervals at $\alpha = 0.05$ with Bonferroni correction matched to the number of subgroups were added to each success rate.

2.3. Acoustical analysis

Using a Praat script, the middle third of each vowel was marked based on the labeled whisper recordings. Some recordings showed slight signs of vocal chord engagement. If the spectrogram of a vowel had visible periods in modal speech frequency ranges, a narrow frequency range around a prominent spectral peak was stop-band filtered. The same procedure was applied for low-frequency noises (under 100 Hz) caused by heavy vehicles passing the studio during recording. Within the middle third of each vowel, mean formant frequencies [Hz], center of gravity (CoG, [Hz]) and spectral slope [dB] were obtained using a Praat script.

Formants were analyzed using the *burg* method between 0 and 5.5 kHz with maximum number of peaks set to 5. Additionally, ratios of mean formant frequencies (F2:F1, F3:F2, F3:F1) were calculated in R. CoG was measured from each vowel's spectrum with default Praat settings. For recordings of whispered speech, a second CoG value was measured after using a stop-band filter between 1 and 6.3 kHz, removing main formant structure visible in the spectrogram. Spectral slope was obtained from LTAS spectra of the vowels using Praat's *get slope* function (method *energy*, bandwidths 50 Hz–2 kHz and 2–8 kHz).

Phone duration as a parameter was omitted in this analysis, as it was likely affected by recording speech via a shadowing task as well as singing with a metronome. Even though duration may be one of the distinguishing factors of pitch in tone languages [25], it was not proven to be a major acoustic cue to melody in whisper recorded in a controlled environment [11].

The effects of each acoustical parameter on melody in whisper were analyzed using linear mixed-effects models in R [23] with lme4 [26] and emmeans [27] packages. The dependent variable for each model was one of the aforementioned acoustical parameters (F1–F3, formant ratios, CoG, spectral slope). Intended pitch movement (rise/fall) of a word in interaction with vowel position within the word (in first/second syllable) was used as a fixed factor (further referred to as intended melody). Random factors encompassed speakers and target words with random intercepts. Formant and CoG values were logarithmed to prevent heteroscedasticity. Statistical significance of the fixed factors' effects was determined using likelihood ratio tests.

3. RESULTS AND DISCUSSION

3.1. Discernibility of melody in whisper

Respondents were able to discern the intended melody in 52–83 % (median 72 %) for whispered speech and 66–94 % (median 83 %) for words sung in whisper.

As shown in Figure 3, intended melody was discernible with statistical significance in words sung in whisper, but the success rate in whispered speech was significant only in stimuli with falling melody. Greater ease of discerning falling rather than rising melody may be caused by the fact that whispered melody seems to be more easily discernible when the higher of the two tones matches with the position of word stress [6]. Due to the nature of the shadowing task used while recording and Czech stress being word-initial, the speakers may have been placing prominent word-initial stress on target words, which would make it correspond to higher tones in words with falling melody.

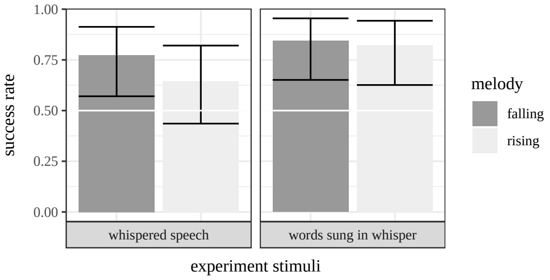

Figure 3: Success rates of perception experiments by intended melody with confidence intervals at $\alpha = 0.05$ with Bonferroni correction for $n = 2$.

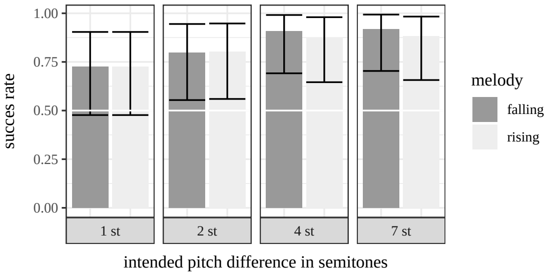

Figure 4: Success rates of discerning melody of whisper-sung words by intended melody and pitch difference with confidence intervals at $\alpha = 0.05$ with Bonferroni correction for $n = 8$.

In addition to falling melody being discerned more easily, greater differences in intended pitch of the syllables within a word lead to higher success rates, as can be seen in Figure 4, which shows success rates of discerning individual melodic contours in words sung in whisper.

3.2. Formants and formant ratios

In whispered speech, only F2 was significantly affected by intended melody ($\alpha = 0.05$, $p = 0.046$). For words sung in whisper, intended melody significantly affected F2 ($p < 0.001$), F3 ($p = 0.008$) as well as F2:F1 ($p < 0.001$) and F3:F2 ($p = 0.012$) ratios. The effect of intended melody on formant ratios in words sung in whisper points to the

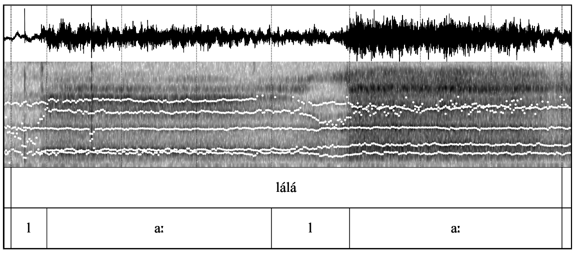

Figure 5: [laːlaː] sung in whisper with a rising melody with an intended pitch difference of 7 st. Dotted lines denote formants. The spectrogram frequency range is 0–8 kHz, time step 5 ms.

movement of F2 being more prominent than those of F1 or F3. One realization of such F2 movement can be observed in Figure 5, which shows the spectrogram of a stimulus recognized by 100 % of respondents as having a rising melody. This is in line with the results of I. B. Thomas [8] who determined F2 to be a direct correlate of pitch in whisper. More recent research has also established formants as a correlate of melody in whisper. Perception experiments using synthesized vowels pointed to the movement of F2 corresponding to perceived melody with the effect being stronger when F1 was shifted as well [10]. A study of whispered Dutch then reported a statistically significant effect of melody on the first three formants [11]. F1 was, however, not proven to be significantly affected by intended melody in the present study.

The movement of F2 and its interaction with F1 and F3 was more significant in words sung in whisper than in whispered speech. A possible explanation lies in the characteristics of regular singing and speech. While intonation in speech is based on relative pitch movements, sung melody is intended to match specific musical notes. Speech may also be more precise in maintaining vowel quality, preventing formants from shifting to a degree where formant relations would interfere with those of a different vowel. It should also be noted that the recorded vowels sung in whisper were longer in duration than vowels in whispered speech, so speakers had more time to reach the intended target pitch.

3.3. Center of gravity

Center of gravity was significantly affected by intended melody only in words sung in whisper ($\alpha = 0.05$, $p = 0.002$). In whispered speech, intended melody significantly affected CoG of the signal after using a stop-band filter removing the main formant bandwidths ($p = 0.022$). This corresponds to [7] stating high-frequency spectral noise as a correlate of melody in some whispered vowels, as well as newer studies which found the main acoustical cues to melody in whisper likely to be in frequency ranges above 1.5 kHz [5].

3.4. Spectral slope

Perception experiments with spectral slope manipulations proved that spectral slope can affect perceived pitch [28]. Spectral slope was also found to be a correlate of intended melody in whisper [11], which is in line with the hypothesis that if higher pitch requires higher effort, it will also show a less negative spectral slope [6].

In the present study, the effect of intended melody in whisper on spectral slope was statistically significant only in words sung in whisper ($\alpha = 0.05$, $p = 0.0069$), but not in whispered speech ($p = 0.24$). A possible explanation may lie in the setup of the recording. Speakers were asked to repeat the heard melody in both tasks, it is however likely that while they attempted to be precise in whisper-singing, they repeated the speech materials in their own voice range which was more comfortable for them and therefore required less effort.

4. CONCLUSION

This paper presented an insight into the production and perception of melody in whisper in a Czech-language environment. Two-syllable words realized with different intended melodic contours were recorded and used as data for this study. A perception experiment showed that intended melody can be more easily discerned with greater difference in intended pitch of the two syllables. Falling melody was also more successfully recognized than rising melody.

Several acoustical parameters were then analyzed as possible cues to intended melody in whisper using linear mixed-effects models. In whispered speech, F2 and spectral slope of stop-band filtered signal removing visible formant structure were significantly affected by intended melody. In words sung in whisper, acoustical parameters significantly affected by intended melody were F2, F3, F2:F1 and F3:F2 ratios, which point to greater movements of F2 in comparison to surrounding formants, as well as center of gravity and spectral slope.

5. ACKNOWLEDGMENTS

This work was supported by the Cooperatio research program of Charles University.

6. REFERENCES

[1] F. Giet, *Zur Tonität nordchinesischer Mundarten*, ser. Anthropos-Institut Sankt Augustin: Studia Instituti Anthropos. Verlag der Missionsdruckerei St. Gabriel, 1950.

[2] G. Panconcelli-Calzia, "Das Flüstern in seiner physio-pathologischen und linguistischen Bedeutung," *Lingua*, vol. 4, pp. 369–378, Jan. 1954.

[3] F. Giet, "Kann man in einer Tonsprache flüstern?" *Lingua*, vol. 5, pp. 372–381, Jan. 1955.

[4] M. Kloster Jensen, "Recognition of word tones in whispered speech," *WORD*, vol. 14, no. 2-3, pp. 187–196, 1958.

[5] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2026–2040, Apr. 2014.

[6] W. F. L. Heeren and V. J. van Heuven, "The interaction of lexical and phrasal prosody in whispered speech," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3272–3289, Dec. 2014.

[7] W. Meyer-Eppler, "Realization of Prosodic Features in Whispered Speech," *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 104–106, jan 1957.

[8] I. B. Thomas, "Perceived Pitch of Whispered Vowels," *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 468–470, Aug. 1969.

[9] J. Fónagy, "Accent et intonation dans la parole chuchotée," *Phonetica*, vol. 20, no. 2-4, pp. 177–192, 1969.

[10] M. Higashikawa and F. D. Minifie, "Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 3, pp. 583–591, 1999.

[11] W. F. L. Heeren, "Vocalic correlates of pitch in whispered versus normal speech," *The Journal of the Acoustical Society of America*, vol. 138, no. 6, pp. 3800–3810, Dec. 2015.

[12] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, Feb. 2005.

[13] Z. Raeesy, K. Gillespie, C. Ma, T. Drugman, J. Gu, R. Maas, A. Rastrow, and B. Hoffmeister, "LSTM-based whisper detection," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 139–144.

[14] K. Phapatanaburi, W. Pathonsuwan, L. Wang, P. Anchuen, T. Jumphoo, P. Buayai, M. Uthansakul, and P. Uthansakul, "Whispered speech detection using glottal flow-based features," *Symmetry*, vol. 14, no. 4, p. 777, 2022.

[15] P. Machač and P. Šturm, "The phonological contrast of voicing in whispered Czech and its phonetic correlates – A preliminary study," *20th Czech-German Workshop - Speech Processing*, pp. 34–43, 2010.

[16] R. Skarnitzl, P. Šturm, and P. Machač, "The phonological voicing contrast in Czech: an EPG study of phonated and whispered fricatives." in *14th Annual Conference of the International Speech Communication Association*, 2013, pp. 3191–3195.

[17] M. Svatošová and T. Bořil, "Duration as a cue for phonological voicing contrast in whispered Czech," in *Proc. 20th ICPhS*, 2023. [in print].

[18] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]. Version 6.1.56*, 2021. [Online]. Available: http://www.praat.org/

[19] Apple Inc, *GarageBand for Mac. Version 10.4.6*, 2022. [Online]. Available: https://www.apple.com/mac/garageband/

[20] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Epocha, 2010.

[21] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[22] ——, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] F. E. Harrell Jr, *Hmisc: Harrell Miscellaneous. Version 4.7-0*, 2022. [Online]. Available: https://CRAN.R-project.org/package=Hmisc

[25] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," *Language and Speech*, vol. 47, no. 2, pp. 109–138, 2004.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[27] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[28] J. Kuang and M. Liberman, "The effect of spectral slope on pitch perception," in *16th Annual Conference of the International Speech Communication Association*, 2015.