Prosodically-conditioned temporal variation in whispered vs. normal speech

Annabelle di Lustro and Sam Tilsen
Cornell University (USA)

Whispered speech prototypically lacks vocal fold vibration and thus exhibits substantial spectral differences from speech with normal vocalization, including the absence of F0-related prosodic cues [1]. Furthermore, whispered speech is aerodynamically disadvantaged: subglottal pressure is exhausted more rapidly during whispered speech, due to higher airflow [2]. These spectral and aerodynamic factors make competing predictions regarding temporal differences between whispered and normal speech: on one hand, speakers may compensate for impoverished prosodic cues by extending duration in prosodically strong positions; on the other hand, they may address aerodynamic needs by lengthening stop closures and shortening high airflow segments like vowels and fricatives. We tested these hypotheses by examining temporal differences between whispered and normal productions of The North Wind and the Sun passage and analyzing their segmental, prosodic, and global temporal differences. We found that speech rate was slower in whispered speech, but most speakers paused less; furthermore, durational increases primarily occurred in phrase-final and phrase-initial prosodic positions. Overall, the results indicate that the main prosodic adjustment of whispered speech is an exaggeration of phrase-final lengthening, suggesting compensation for the absence of F0-related prosodic cues.

Method. Data were collected from 20 native English speakers (16F, 4M) through an online production experiment on Labvanced. The speakers read The North Wind and the Sun passage 5 times in a whispered voice and 5 times in normal vocalized speech. Participants were instructed on how to whisper without voicing. Word- and segment-level alignments were generated with the Montreal Forced Aligner [3] with acoustic models adapted separately to whispered and normal productions. For each recording, we calculated global speech rate (num. words / total word duration of the passage) and unfilled pause proportion (ratio of total silence interval duration to passage duration); each of the variables was regressed with a fixed effect of production mode and random slopes of production mode by speaker. For each word, we regressed word duration with a random factor of speaker, conducted two-sample t-tests of residual word durations from whispered and normal conditions, and used Bonferroni correction to assess whether there was a significant difference. For each segment (defined uniquely by word/occurrence in word), we regressed duration with a fixed effect of condition and random intercepts for speakers.

Results. Nearly all participants produced words more slowly in whispered speech (Fig. 1a) but most paused less, both proportionally (Fig. 1b) and in terms of total pause duration. The regression model fixed effects show that whisper reduced speech rate by 0.6 words/s and decreased pausing by 4% of passage duration. Participants pausing less despite speaking slower suggests that temporal adjustments in whisper are not attributable to the need to inhale more frequently. Instead, word-level analyses (Fig. 2) showed that durational effects were predominantly localized to phrase-final prosodic positions and to a lesser extent, phrase-initial prosodic positions: of the ten largest word effects, eight occur in phrase-final positions. Segment-level analyses showed that whisper was associated with more segmental lengthening than shortening; significantly lengthened segments tended to be word-final, while shortened segments were predominantly initial. Contrary to some previous studies [2], our results suggest that speakers make prosodically conditioned temporal adjustments to facilitate listener comprehension or to augment feedback, rather than segmental timing adjustments to preserve airflow.

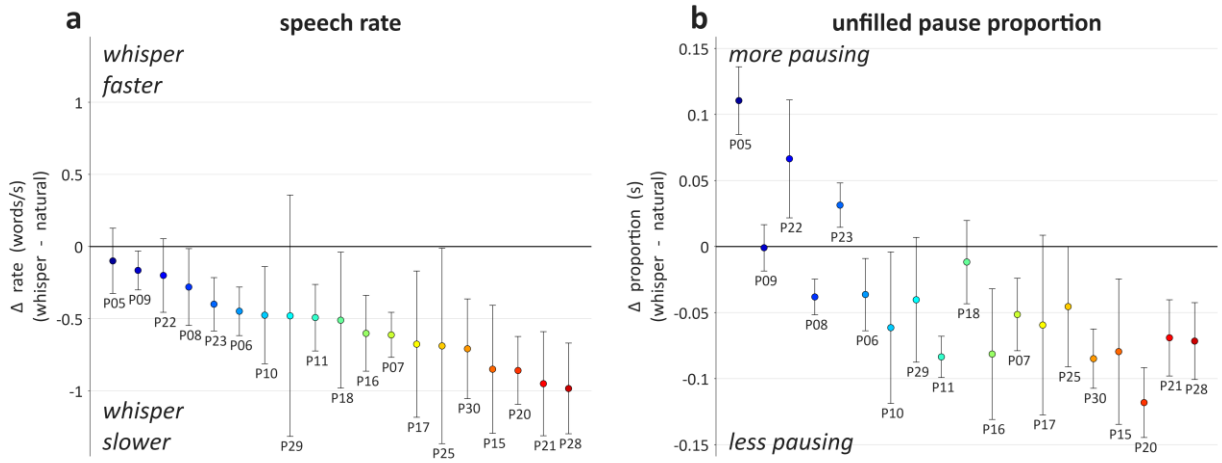

Fig. 1. Difference in speech rate (a) and difference in unfilled pause proportion (b) between whispered and normal trials by speaker.

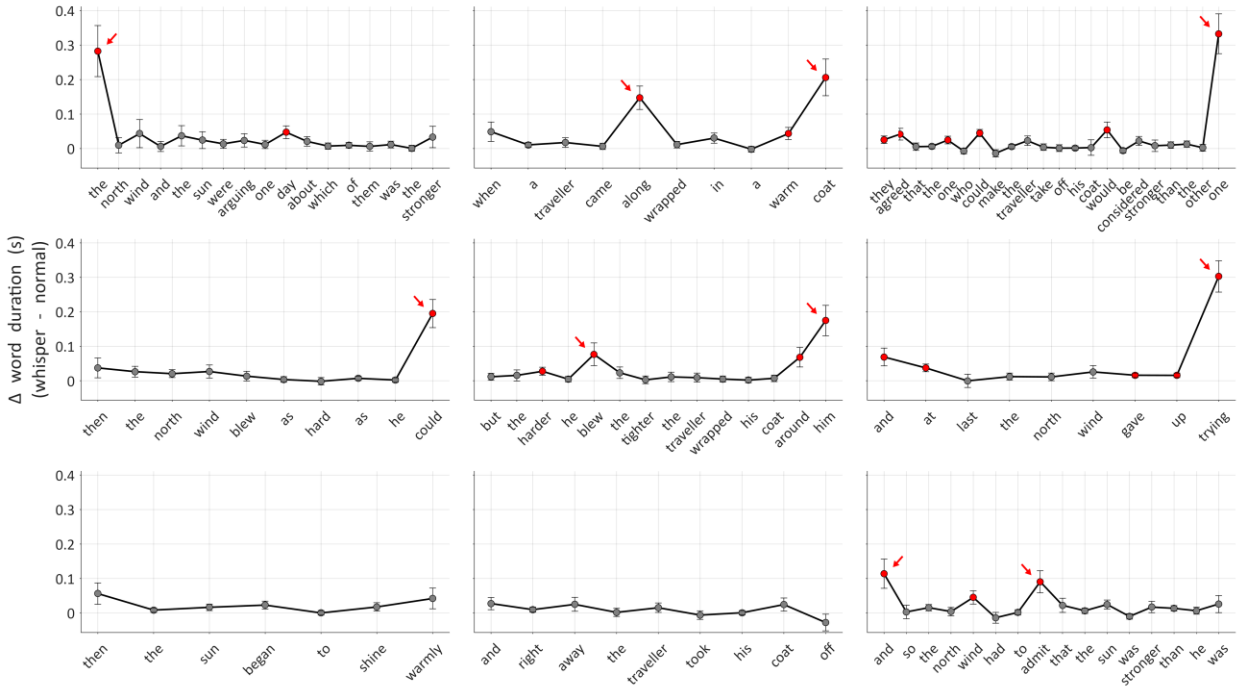

Fig 2. Difference in word duration between whispered and normal trials for The North Wind and the Sun passage, with red denoting significance. Ad-hoc passage portions are presented in order from left to right, top to bottom. Arrows indicate the 10 largest effects.

[1] Tartter, V. C. (1989). What's in a whisper?. *J. Acoust. Soc. Am., 86(5),* 1678–1683.
[2] Schwartz, M. F. (1972). Bilabial Closure Durations for /p/, /b/, and /m/ in Voiced and Whispered Vowel Environments. *J. Acoust. Soc. Am.*, *51(6B)*, 2025–2029.
[3] McAuliffe, M., Socolof M., Mihuc S., Wagner M., and Sonderegger M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.