Phonological effects on the perceptual weighting of voice cues for voice gender categorization

Jebens, Almut; Başkent, Deniz; Rachman, Laura

Published in:
JASA Express Letters

DOI:
10.1121/10.0016601

JASA Express Letters

DECEMBER 21 2022

Phonological effects on the perceptual weighting of voice cues for voice gender categorization

Almut Jebens ; Deniz Başkent ; Laura Rachman

JASA Express Lett. 2, 125202 (2022)
https://doi.org/10.1121/10.0016601

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Phonological effects on the perceptual weighting of voice cues for voice gender categorization

Almut Jebens, Deniz Başkent, and Laura Rachman[a)]
Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
a.n.jebens@student.rug.nl; d.baskent@rug.nl; l.rachman@rug.nl

Abstract: Voice perception and speaker identification interact with linguistic processing. This study investigated whether lexicality and/or phonological effects alter the perceptual weighting of voice pitch (*F*0) and vocal-tract length (VTL) cues for perceived voice gender categorization. *F*0 and VTL of forward words and nonwords (for lexicality effect), and time-reversed nonwords (for phonological effect through phonetic alterations) were manipulated. Participants provided binary "man"/"woman" judgements of the different voice conditions. Cue weights for time-reversed nonwords were significantly lower than cue weights for both forward words and nonwords, but there was no significant difference between forward words and nonwords. Hence, voice cue utilization for voice gender judgements seems to be affected by phonological, rather than lexicality effects.

[Editor: Lina A.J. Reiss] https://doi.org/10.1121/10.0016601

Received: 28 June 2022 **Accepted:** 1 December 2022 **Published Online:** 21 December 2022

1. Introduction

Among the many voice cues, two anatomically related cues, mean fundamental frequency (*F*0) and vocal-tract length (VTL), mainly relate to the perceived gender of a talker's voice (Smith and Patterson, 2005). *F*0 is related to the glottal pulse rate and is perceived as vocal pitch, while VTL correlates with a speaker's body size (Evans *et al.*, 2006; Fitch and Giedd, 1999) and shapes the formant frequencies in the speech signal (Hillenbrand *et al.*, 1995). When synthesized and manipulated, these voice cues affect the perception of voice gender (Fuller *et al.*, 2014). Furthermore, voice cues are important to distinguish different talkers in multi-talker listening conditions and may improve speech intelligibility (Brungart, 2001; Darwin *et al.*, 2003).

Listeners with normal hearing (NH) typically use both *F*0 and VTL cues, and give a perceptual weighting to them, for the assessment of a speaker's voice gender (Skuk and Schweinberger, 2014). In contrast, deaf individuals with cochlear implants (CIs) have shown reduced sensitivity to synthesized VTL cues from syllables (Gaudrain and Başkent, 2018). When utilizing *F*0 and VTL cues to assess voice gender, CI users overly rely on *F*0 cues (Fuller *et al.*, 2014). In daily life, they may benefit from linguistic content for voice perception, such as lexical status, as was shown in CI simulation studies comparing forward and time-reversed words (Koelewijn *et al.*, 2021), or sentential context, when contrasting words and sentences (Meister *et al.*, 2016). Further support for interactions between language and voice comes from studies showing that familiarity with a voice can improve speech intelligibility (Holmes *et al.*, 2018; Holmes and Johnsrude, 2020; Nygaard and Pisoni, 1998). Conversely, listeners are better at identifying and distinguishing voices when they hear them in their native language (Goggin *et al.*, 1991; Johnson *et al.*, 2011). Finally, listeners tend to assign two utterances to the same speaker instead of to two different speakers when they form a lexical compound (e.g., day-dream) compared to when the two combined items lack any meaning (Narayan *et al.*, 2017; Quinto *et al.*, 2020). However, it is not known if such a linguistic influence also exists for voice gender perception, and if so, whether it is driven by lexicality or phonological effects.

In this study, as a first step to explore lexicality and phonological effects on the use of indexical cues, voice gender categorization was measured in NH listeners in different linguistic conditions. Forward Dutch words and nonwords, as well as time-reversed nonwords, were synthesized to manipulate *F*0 and VTL cues and create different voice identities ranging from female-source talkers to synthesized male-like talkers. In order to address phonological effects, we used temporal reversal of speech, which leads to phonetic changes where articulatory and voicing characteristics are altered, such that little of the segmental content remains in the signal (Sheffert *et al.*, 2002).

We hypothesized that if voice cue weighting for the assessment of voice gender is influenced by effects of lexicality, this would result in higher cue weights for forward words compared to forward nonwords. This result would be in

[a)]Author to whom correspondence should be addressed.

line with findings by Xie and Myers (2015), who found better talker identification performance in meaningful vs meaningless sentences, which they suggest may be driven by top-down effects of lexicon knowledge that increase access to talker-specific acoustic-phonetic cues. In addition, if voice cue weighting for voice gender categorization is influenced by phonological effects, we expect higher cue weights for forward nonwords compared to time-reversed nonwords as a result of increased sensitivity to these cues, as was reported by Koelewijn *et al.* (2021).

2. Methods

2.1 Participants

Twenty adult participants (self-reported gender: 3 women, 17 men; mean age: 26.8 years, range 18–49 years) were recruited and reimbursed via the online testing-platform "Prolific" (Palan and Schitter, 2018). All participants reported not having any neurological disorders or history of language or reading impairments. Participants reported to be native speakers of Dutch, and five of the 20 participants were raised multilingually (see Table S1 in the supplementary material[1] for demographic information). Participants reported having normal hearing, which was confirmed with an online version of the digit-in-noise test (DIN; Smits *et al.*, 2006) for 19 out of 20 participants. The study received ethical approval by the Medical Ethical Committee of the University Medical Center Groningen (METc 2018/427). Participants provided informed consent prior to the study and received financial compensation according to Prolific and departmental guidelines.

2.2 Stimuli

Two linguistic manipulations were created. First, we examined the contribution of lexical status by including words and nonwords. Second, we examined phonological effects through phonetic alterations by including forward nonwords and time-reversed nonwords. By time-reversing speech, some phonetic features of a language are altered, for example, because consonant processing is disrupted through time-reversal due to the removal of voice onset time features, or because illicit phoneme sequences may occur (for further discussion, see Perrachione *et al.*, 2019). On the other hand, some features, such as amplitude, duration, and mean *F*0, are preserved, and the formant transition structure of certain phonemes, such as fricatives and long vowels, is roughly mirrored in the reversed signal (Binder *et al.*, 2000). Eight words and eight nonwords with a consonant-consonant-vowel-consonant (CCVC) or a consonant-vowel-consonant-consonant (CVCC) format were selected from the VariaNTS corpus (Arts *et al.*, 2021) for two of the linguistic conditions. The number of stimuli with the consonant cluster at the beginning and the end was balanced across the stimuli. Three utterances of each selected stimulus, produced by three different female speakers, were included in this study. The three source-speakers were selected from the eight available female speakers in the VariaNTS database. Selection was done after applying a 12th order high-pass Butterworth filter with a cut-off at 80 Hz and adding 5 ms of silence at the start of the sound files using Adobe Audition software (Adobe Inc., San Jose, CA). These voices were selected because they sounded the most natural and clearest after *F*0 and VTL manipulations (as follows) according to the authors. The source-speakers' mean *F*0 and their height and weight are presented in Table 1.

The selected words were controlled for morphological and lexical-semantic parameters and both words and nonwords were controlled for phonological parameters. All words are monomorphemic nouns, meaning that they only contain one meaningful word element. In terms of lexical-semantic features, they are rated as highly familiar on a scale from 1 (unfamiliar) to 7 (highly familiar) by Dutch native speakers (Arts *et al.*, 2021). They were classified as high-frequent by the authors of the corpus based on two corpora, ranging from 21 to 515 per million according to the CELEX database (Baayen *et al.*, 1993) and from 24 to 274 per million according to the SUBTELEX database (Keuleers *et al.*, 2010). In terms of their phonological features, all word and nonword stimuli have a low neighborhood density (Marian *et al.*, 2012). The number of neighbors is defined by the words within the language that can be created by deleting, substituting, or adding phonemes to the original item. Nonwords have a high phonotactic probability that is derived from biphone frequencies of their phonemes (Arts *et al.*, 2021). This value refers to the frequency with which two subsequent phonemes in real Dutch words occur based on the CLEARPOND database (Marian *et al.*, 2012). Words and nonwords were controlled for the position of the consonant cluster, either appearing at the beginning or the end of the stimulus. Due to a possible interaction between a speaker's VTL and the formants (Irino and Patterson, 2002), words and nonwords were matched for vowels. Furthermore, phonotactic probability did not significantly differ between words and nonwords [$t(14) = -1.14$, $p = 0.28$]. Controlled parameters for words and nonwords are presented as supplementary material[1] in Tables S2 and S3, respectively. For the third linguistic condition, time-reversed versions of the selected nonwords were created using MATLAB.

Table 1. Age, height, weight, and mean *F*0 of the three source-speakers from the VariaNTS corpus (Arts *et al.*, 2021).

Speaker	Age (years)	Height (cm)	Weight (kg)	Mean *F*0 (Hz)
2	20	171	59	214.36
12	22	175	78	191.83
15	21	176	65	199.38

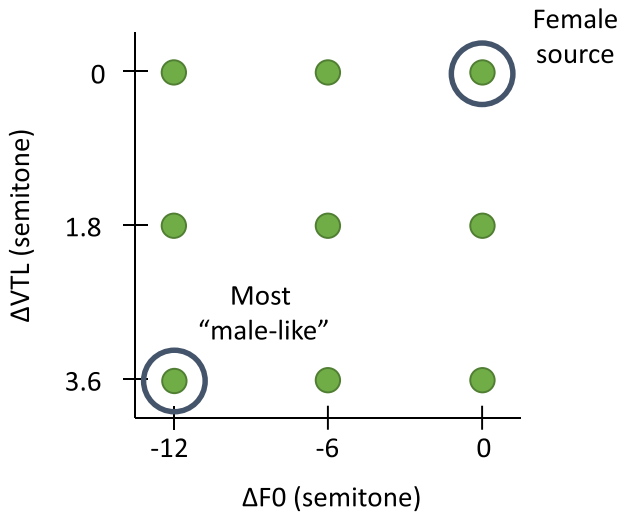

Fig. 1. Nine voice conditions were created by manipulating *F*0 and VTL cues relative to the female source-voices.

For each stimulus item, nine voice conditions were created by manipulating *F*0 and VTL values independently from each other. *F*0 was decreased by 0.0, 6.0, and 12.0 semitones (st) and VTL was increased by 0.0, 1.8, and 3.6 st (Fig. 1). These values were based on previous work on "voice gender categorization" (Smith and Patterson, 2005; Smith *et al.*, 2007). Furthermore, in a study with normal-hearing adult listeners, Fuller *et al.* (2014) have shown that simultaneously decreasing *F*0 by 12 st and increasing VTL by 3.6 st reliably changed the perception from a female voice to a male voice. Stimuli were resynthesized using the PyWorld wrapper (Hsu, 2016) for the WORLD vocoder (Morise *et al.*, 2016), implemented in the Voice Transformation Server (Gaudrain, 2021). To avoid an effect of potential artifacts due to the resynthesis procedure, stimuli for which *F*0 and VTL differences were 0.0 st were resynthesized as well, using the same procedure. Taking all conditions together, this resulted in a total of 648 stimuli that were presented to each participant [9 voice conditions $*$ 3 linguistic conditions $*$ 8 items $*$ 3 speakers].

2.3 Procedure

Data were collected through a remote testing procedure on a web-based platform that was developed using the JavaScript framework JsPsych (de Leeuw, 2015). Participants completed the experiment on their own computers and all participants were requested to wear headphones and to be in a quiet room during the test. Informed consent and demographic information were provided at the start of the experiment. Furthermore, participants were asked to complete an online DIN test (Smits *et al.*, 2006; Smits *et al.*, 2013) via https://www.hoortest.nl/ to verify their hearing status (sufficient: $n = 19$, insufficient: $n = 1$, poor: $n = 0$).

Participants then proceeded with the "voice gender categorization task," consisting of a one-interval two-alternative forced-choice (1I-2AFC) categorization task in which responses were limited to "woman" and "man" and no feedback was provided, similar to Nagels *et al.* (2020) and Fuller *et al.* (2014). The experimental session started with six practice trials during which the practice items from Tables S2 and S3[1] were presented in two voice conditions ($\Delta F0/\Delta VTL = 0.0/0.0$ st and $-12/+3.6$ st). The practice stimuli were produced by a different female speaker than the three included female source-speakers in the main experiment to prevent any adaptation effects. Practice stimuli also stemmed from the VariaNTS corpus (Arts *et al.*, 2021) and were controlled for the same linguistic variables as the testing stimuli.

Linguistic conditions were presented in separate blocks, with three block repetitions per linguistic condition, resulting in a total of nine blocks. The block order was randomized, as well as the order of speakers and voice conditions within each block. Participants completed the experiment in approximately 50 min.

2.4 Cue weighting

To quantify how participants used voice cues in different linguistic conditions, perceptual weighting of *F*0 and VTL was calculated for each linguistic condition using RStudio (Version 1.2.5042) and the lme4 package (Bates *et al.*, 2020). We first normalized the *F*0 and VTL differences in st relative to the voices of the source talkers, defined as $\delta F0 = -\Delta F0/12 - 0.5$ and $\delta VTL = \Delta VTL/3.6 - 0.5$. This procedure was performed to make sure that, despite differences in range and quality, *F*0 and VTL were functionally equivalent for subsequent model fitting. After normalizing the voice cues, the original female voices had a $\delta F0$ value of -0.5 and a δVTL value of -0.5, while the intermediate voice differences (-6 st for *F*0 and $+1.8$ st for VTL) had a value of 0.0, and the most male-like voice differences had a value of $+0.5$ (-12 st for *F*0 and $+3.6$ st for VTL). Cue weight coefficients were then extracted for each voice cue and each linguistic condition by

fitting a mixed-effects logistic regression model with random intercepts and slopes for δF0 and δVTL per voice condition and per participant, in lme4 syntax: `response ~ (1+(δF0 + δVTL) * linguistic condition | participant)`. Finally, coefficients for δF0 and δVTL for every linguistic condition were converted into Berkson (Bk) units for each st, which corresponds to a $\log_2$ odds ratio per st (i.e., one Bk per st equals doubling the categorization of a given stimulus as "man" (see Hilkhuysen *et al.*, 2012; Nagels *et al.*, 2020), and factors were sum coded. The Bk units of each participant were then analyzed by fitting a generalized linear mixed-effects model with random intercepts per participant. In the next step, we compared models in a backward stepwise model selection with an analysis of variance (ANOVA) Chi-Square test, and based on their significance ($p = 0.05$), factors were kept in the model. We started with the full factorial model and a two-way interaction between the fixed effects of *voice cue* (*F*0 and VTL) and *linguistic condition* (words, nonwords, reversed nonwords) and a random intercept per participant and *cue weight* in Bk/st as an outcome variable, in lme4 syntax: `cue weight ~ voice cue * linguistic condition + (1|participant)`. Finally, a *post hoc* analysis was performed based on the best-fitting model using *emmeans* from the emmeans package (Lenth *et al.*, 2018), through pairwise comparisons with Bonferroni correction.

3. Results

Figure 2 shows the average percentages of "female" ratings at each *F*0 and/or VTL manipulation for each linguistic condition. Figure 3 shows participants' cue weights for *F*0 and VTL in each linguistic condition. Model comparison showed that the full model with a random intercept per participant had a significantly better fit than the full model without a random intercept [$\chi^2(1) = 23.3$, $p < 0.001$]. Backward stepwise selection showed that the best-fitting and most parsimonious model was the model with voice cue and linguistic condition as fixed effects. This model did not differ significantly from the model with a two-way interaction between voice cue and linguistic condition [$\chi^2(2) = 2.4$, $p = 0.30$], while it had a significantly better fit than the models with only voice cue [$\chi^2(2) = 14.8$, $p < 0.001$] or only linguistic condition as fixed effect [$\chi^2(1) = 28.6$, $p < 0.001$].

Follow-up pairwise comparisons revealed that cue weights for time-reversed nonwords were significantly lower than cue weights for forward words and nonwords [$ps < 0.001$], while the cue weights did not differ between forward words and forward nonwords [$p = 1$]. This indicates that voice gender categorization responses of time-reversed stimuli were less affected by *F*0 and VTL changes than categorization responses of forward stimuli. It should be noted that absolute cue weights should only be compared between the linguistic conditions, but not across voice cues. *F*0 and VTL perception rely on different mechanisms and as such, a direct comparison of the use of these cues is not straightforward. While the conversion of cue weights to Bk/st allows for performing statistical analyses on these physical units, it is not known whether and how this may map onto perceptual units. For this reason, the effect of voice cue was not submitted to further statistical testing.

4. Discussion

This study investigated whether perceptual weighting of *F*0 and VTL cues for the categorization of voice gender is affected by linguistic effects. We manipulated *F*0 and VTL cues of monosyllabic stimuli in three linguistic conditions: forward words, forward nonwords, and time-reversed nonwords. Our results show that cue weighting of *F*0 and VTL was significantly different for the time-reversed stimuli compared to the two forward-presented stimulus conditions, while cue weighting did not significantly differ between the forward words and forward nonwords. These results indicate that cue weighting of *F*0 and VTL for voice gender categorization is altered by stimulus time-reversal, but not by lexical status, and imply that voice gender perception and linguistic processing are altered by phonological processes as a result of the phonetic manipulations.

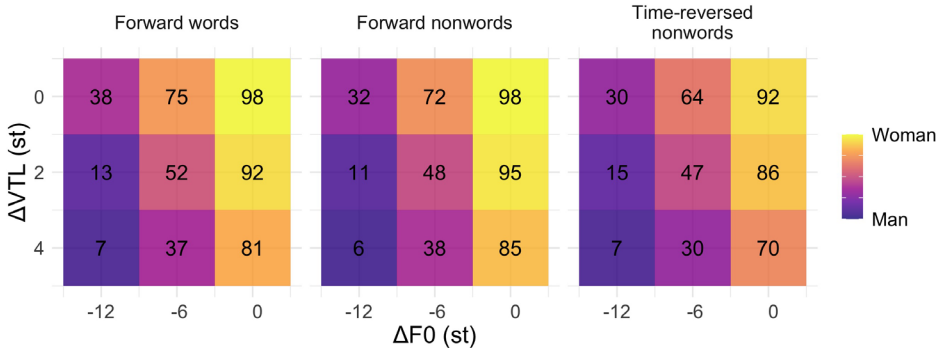

Fig. 2. Average voice gender categorization judgements as a function of differences in *F*0 (*x* axis) and VTL (*y* axis) in st for each linguistic condition. The numbers indicate the average "woman" responses in percentages. The colors provide a gradient between 100% "man" responses (blue) and 100% "woman" responses (yellow).

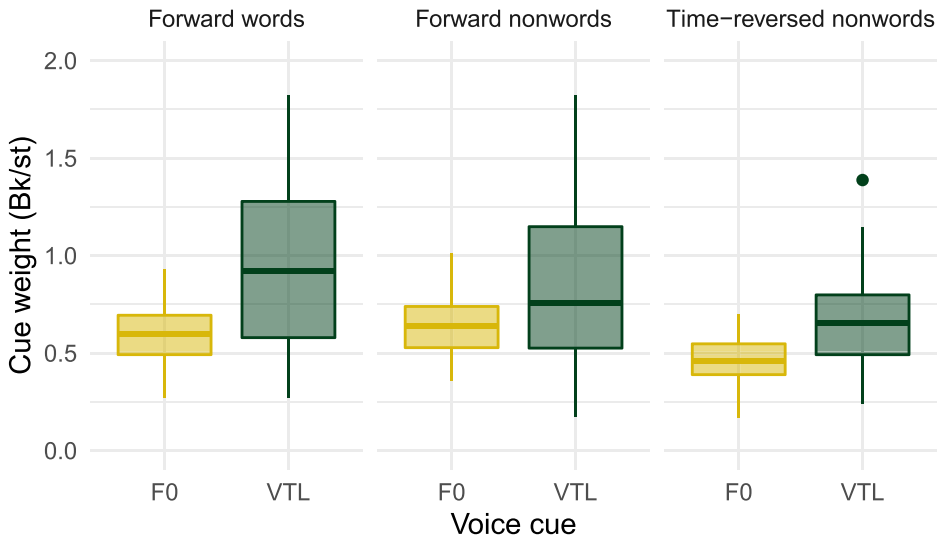

Fig. 3. *F*0 and VTL cue weights shown for the three linguistic conditions. Cue weights are presented as Berkson per semitone (Bk/st). The boxplots show the median cue weights per voice cue and per linguistic condition. Boxes extend from the lower to the upper quartiles (interquartile range, IQ) The whiskers indicate the lowest and highest data points within plus or minus 1.5 times the IQ. The dot indicates an outlier with a value larger than 1.5 times the IQ.

Our results differ from those reported by Abu El Adas and Levi (2022) and Perrachione *et al.* (2015), who found that listeners were better at identifying talkers when presented with words compared to nonwords. Furthermore, Xie and Myers (2015) also found that listeners were more accurate at identifying talkers when presented with meaningful compared to meaningless sentences. On the other hand, Zarate *et al.* (2015) found no difference in talker identification when comparing word vs nonword conditions. It is worth noting that our test differs from talker identification tasks used in the abovementioned studies because the phonological representation that listeners have to access in identification tasks is likely much more precise than the phonological representations for voice gender perception. Consequently, the extent to which lexical status affects performance in these tasks may also differ.

The results presented here also suggest that the reduced sensitivity for *F*0 and VTL cues in words as reported by Koelewijn *et al.* (2021) could be related to the distorted phonetics induced by time-reversing the stimuli. In a follow-up study, Koelewijn *et al.* (2022) showed a reduced sensitivity to *F*0 and VTL cues in time-reversed words compared to forward words and nonwords. These results are in line with the current findings, as they also point to phonological effects, as a result of phonetic manipulations, on voice cue sensitivity. Furthermore, the results of Koelewijn *et al.* (2021) suggest that the differences in the perceptual weighting of *F*0 and VTL for voice gender categorization may be a result of reduced access to these cues when the speech signal is time-reversed.

The current findings have implications for other relevant areas as well. A language effect on voice perception was previously implied in a speaker discrimination study where listeners rated native-language speakers as more dissimilar than speakers of an unfamiliar language (Fleming *et al.*, 2014). As this study also used time-reversed stimuli, this language familiarity effect (LFE) was proposed to be related to a greater familiarity with the phonetic and phonological structure of the listener's native language, rather than the ability to understand the linguistic content. On the other hand, linguistic processing models describing the LFE propose that linguistic competence (e.g., word recognition and speech comprehension abilities) further increases voice perception (Goggin *et al.*, 1991; Xie and Myers, 2015). It should also be noted that the LFE is sensitive to task design, as shown by different variations of talker identification tasks and talker discrimination tasks, even when using the same stimuli in different types of tasks (for reviews, see Levi, 2019; Perrachione, 2019). Two commonly used tasks to address LFEs are voice lineup tasks, in which listeners have to pick out a learned voice among a set of different voices, and voice discrimination tasks (Levi, 2019). The voice gender categorization task used in this study is neither a pure recognition or identification task, nor a discrimination task. Instead, listeners are asked to make broad categorizations (man vs woman) based on varying voice cues. These differences in experimental design should therefore be carefully considered when comparing findings of interactions between voice and language perception. Nevertheless, while various studies have suggested that sufficient indexical cues remain accessible in time-reversed speech samples for speaker recognition (Bricker and Pruzansky, 1966; Fleming *et al.*, 2014; Sheffert *et al.*, 2002; Van Lancker *et al.*, 1985), our results suggest a difference in accessibility to these cues for voice categorizations due to the partial disruption of the phonetic structure.

The difference in perceptual cue weighting of *F*0 and VTL cues for time-reversed stimuli is also of relevance when considering disorders such as developmental dyslexia. This reading disorder is considered to be driven by impaired phonological abilities (Snowling, 1998), although other more cognitive abilities have been suggested to be impaired, such

as time perception or executive functions (Gooch *et al.*, 2011). In this clinical group, voice perception abilities have been found to be impaired (Perrachione *et al.*, 2011). By controlling for lexical status, by contrasting words with nonwords, we explicitly tested for lexicality effects and could assign the difference in perceptual weighting between forward and time-reversed nonwords to phonological processes as a result of our phonetic manipulations. This emphasizes the link between voice perception abilities and language impairments such as dyslexia with an underlying phonological impairment. Future research could investigate the voice cue weightings in phonological disorders, and how these are affected by phonetic manipulations.

Finally, our results imply that the linguistic content could influence how voice cues are perceived and utilized to derive speaker-related characteristics such as voice gender, which could benefit CI users in their daily communication. In a previous study, Meister *et al.* (2016) reported that CI users made better use of *F*0 and VTL cues in a voice gender categorization task when using sentences compared to single words or word repetitions. While it has been suggested that longer speech samples also provide the listener with more time to interpret speaker-related information, the results by Meister *et al.* are likely driven by the richer phonetic inventory of sentences compared to word repetitions, which carry more redundant information. Previous work on voice identity learning showed that voice learning did not transfer well between words and sentences (Nygaard and Pisoni, 1998). Follow-up work should include sentence stimuli to investigate whether the phonetic manipulation effects on voice cue weighting reported here transfer to longer stimuli. Together with the results of Meister *et al.* (2016) and Koelewijn *et al.* (2021), these findings point to the possibility that top-down mechanisms driven by phonological processing could be utilized by CI users as a compensation strategy (Başkent *et al.*, 2016).

Acknowledgments

This work was funded by an NWO ZonMw VICI grant (918-17-603), the Heinsius Houbolt Foundation, and a Rosalind Franklin Fellowship. The study was conducted as a master's research project of the first author (Research Master's Program in Behavioral and Cognitive Neuroscience). We thank Etienne Gaudrain for advice on data analysis, and Carolyn McGettigan, Matt Davis, and Jan Wouters for their suggestions as members of the scientific advisory board for the VICI project. We are grateful to Jennifer Breetveld for research support and Christina Elsenga for help in data sharing. The data presented in this study will be made openly available in DataverseNL at https://doi.org/10.34894/ICA98X.

References and links

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0016601 for table with demographic information and for tables with controlled parameters for used stimuli.

Abu El Adas, S., and Levi, S. V. (**2022**). "Phonotactic and lexical factors in talker discrimination and identification," Atten. Percept. Psychophys. **84**(5), 1788–1804.

Arts, F., Başkent, D., and Tamati, T. N. (**2021**). "Development and structure of the VariaNTS corpus: A spoken Dutch corpus containing talker and linguistic variability," Speech Commun. **127**, 64–72.

Baayen, R., Piepenbrock, R., and van Rijn, H. (**1993**). *The CELEX Lexical Database (CD-ROM)* (University of Pennsylvania, Linguistic Data Consortium, Philadelphia, PA).

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., and Gaudrain, E. (**2016**). "Cognitive compensation of speech perception with hearing impairment, cochlear implants, and aging," Trends Hear. **20**, 233121651667027–233121651667016.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., and Fox, J. (**2020**). "Linear mixed-effects model using 'Eigen' and S4, R Package Version 1.1- 23," https://github.com/lme4/lme4/ (Last viewed June 21, 2022).

Binder, J., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., and Possing, E. T. (**2000**). "Human temporal lobe activation by speech and nonspeech sounds," Cereb. Cortex **10**(5), 512–528.

Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**(6), 1441–1449.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**(5), 2913–2922.

de Leeuw, J. R. (**2015**). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," Behav. Res. Methods **47**(1), 1–12.

Evans, S., Neave, N., and Wakelin, D. (**2006**). "Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice," Biol. Psychol. **72**(2), 160–163.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**(3), 1511–1522.

Fleming, D., Giordano, B. L., Caldara, R., and Belin, P. (**2014**). "A language-familiarity effect for speaker discrimination without comprehension," Proc. Natl. Acad. Sci. U.S.A. **111**(38), 13795–13798.

Fuller, C. D., Galvin, J. J., Free, R. H., and Başkent, D. (**2014**). "Musician effect in cochlear implant simulated gender categorization," J. Acoust. Soc. Am. **135**(3), EL159–EL165.

Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., and Başkent, D. (**2014**). "Gender categorization is abnormal in cochlear implant users," J. Assoc. Res. Otolaryngol. **15**(6), 1037–1048.
Gaudrain, E. (**2021**). "*egaudrain/VTServer* (V2.2)," Zenodo, https://doi.org/10.5281/zenodo.5801906 (Last viewed December 15, 2022).
Gaudrain, E., and Başkent, D. (**2018**). "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**(2), 226–237.
Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Memory Cogn. **19**(5), 448–458.
Gooch, D., Snowling, M., and Hulme, C. (**2011**). "Time perception, phonological skills and executive function in children with dyslexia and/or ADHD symptoms," J. Child Psychol. Psychiatry **52**(2), 195–203.
Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M. (**2012**). "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," J. Acoust. Soc. Am. **131**(1), 531–539.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.
Holmes, E., Domingo, Y., and Johnsrude, I. S. (**2018**). "Familiar voices are more intelligible, even if they are not recognized as familiar," Psychol. Sci. **29**(10), 1575–1583.
Holmes, E., and Johnsrude, I. S. (**2020**). "Speech spoken by familiar people is more resistant to interference by linguistically similar speech," J. Exp. Psychol. Learn. Memory Cogn. **46**(8), 1465–1476.
Hsu, J. C. C. (**2016**). "Python wrapper for World Vocoder," github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder (Last viewed December 15, 2022).
Irino, T., and Patterson, R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Commun. **36**(3–4), 181–203.
Johnson, E. K., Westrek, E., Nazzi, T., and Cutler, A. (**2011**). "Infant ability to tell voices apart rests on language experience," Dev. Sci. **14**(5), 1002–1011.
Keuleers, E., Brysbaert, M., and New, B. (**2010**). "SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles," Behav. Res. Methods **42**(3), 643–650.
Koelewijn, T., Gaudrain, E., Shehab, T., Treczoks, T., and Baskent, D. (**2022**). "The effects of phonological content, sentence context, and vocoding on voice cue perception," J. Acoust. Soc. Am. **151**(4), A277.
Koelewijn, T., Gaudrain, E., Tamati, T., and Başkent, D. (**2021**). "The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception," J. Acoust. Soc. Am. **150**(3), 1620–1634.
Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (**2018**). "emmeans: Estimated marginal means, aka least-squares means," R package version 1.1.3, https://cran.r-project.org/web/packages/emmeans/emmeans.pdf (Last viewed December 15, 2022).
Levi, S. V. (**2019**). "Methodological considerations for interpreting the language familiarity effect in talker processing," Wiley Interdiscipl. Rev. Cogn. Sci. **10**(2), e1483.
Marian, V., Bartolotti, J., Chabal, S., and Shook, A. (**2012**). "CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities," PLoS One **7**(8), e43230.
Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., and Walger, M. (**2016**). "The use of voice cues for speaker gender recognition in cochlear implant recipients," J. Speech. Lang. Hear. Res. **59**, 546–556.
Morise, M., Yokomori, F., and Ozawa, K. (**2016**). "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," IEICE Trans. Inf. Syst. **E99.D**(7), 1877–1884.
Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., and Başkent, D. (**2020**). "Development of voice perception is dissociated across gender cues in school-age children," Sci. Rep. **10**, 5074.
Narayan, C. R., Mak, L., and Bialystok, E. (**2017**). "Words get in the way: Linguistic effects on talker discrimination," Cogn. Sci. **41**(5), 1361–1376.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376.
Palan, S., and Schitter, C. (**2018**). "Prolific.ac—A subject pool for online experiments," J. Behav. Exp. Finance **17**, 22–27.
Perrachione, T. K. (**2019**). "Speaker recognition across languages," in *The Oxford Handbook of Voice Perception*, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, UK), pp. 515–538.
Perrachione, T. K., Del Tufo, S. N., and Gabrieli, J. D. E. (**2011**). "Human voice recognition depends on language ability," Science **333**(6042), 595.
Perrachione, T. K., Dougherty, S. C., Mclaughlin, D. E., and Lember, R. A. (**2015**). "The effects of speech perception and speech comprehension on talker identification," in *Proceedings of the 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK, pp. 1–4.
Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," J. Acoust. Soc. Am. **146**(5), 3384–3399.
Quinto, A., Abu El Adas, S., and Levi, S. V. (**2020**). "Re-examining the effect of top-down linguistic information on speaker-voice discrimination," Cogn. Sci. **44**(10), e12902.
Sheffert, S. M., Fellowes, J. M., Pisoni, D. B., and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave, and reversed speech samples," J. Exp. Psychol. Hum. Percept. Perform. **28**(6), 1447–1469.
Skuk, V. G., and Schweinberger, S. R. (**2014**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech. Lang. Hear. Res. **57**(1), 285–296.
Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**(5), 3177–3186.
Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007**). "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," J. Acoust. Soc. Am. **122**(6), 3628–3639.
Smits, C., Goverts, S. T., and Festen, J. M. (**2013**). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," J. Acoust. Soc. Am. **133**(3), 1693–1706.

Smits, C., Merkus, P., and Houtgast, T. (**2006**). "How we do it: The Dutch functional hearing-screening tests by telephone and internet," Clin. Otolaryngol. **31**(5), 436–440.

Snowling, M. (**1998**). "Dyslexia as a Phonological Deficit: Evidence and Implications," Child Psychol. Psychiatr. Rev. **3**(1), 4–11.

Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices," J. Phon. **13**(1), 19–38.

Xie, X., and Myers, E. B. (**2015**). "General Language Ability Predicts Talker Identification," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, July 22–25, Austin, TX, 2697–2702.

Zarate, J. M., Tian, X., Woods, K. J. P., and Poeppel, D. (**2015**). "Multiple levels of linguistic and paralinguistic features contribute to voice recognition," Sci. Rep. **5**, 11475.