# The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception

Thomas Koelewijn, Etienne Gaudrain, Terrin Tamati, Deniz Başkent

# The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception

Thomas Koelewijn, Etienne Gaudrain, Terrin Tamati, et al.

Citation: The Journal of the Acoustical Society of America **150**, 1620 (2021); doi: 10.1121/10.0005938
View online: https://doi.org/10.1121/10.0005938
View Table of Contents: https://asa.scitation.org/toc/jas/150/3
Published by the Acoustical Society of America

JASA ARTICLE

# The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception

Thomas Koelewijn,[1,a),b)] Etienne Gaudrain,[2,b),c),d)] Terrin Tamati,[3,b),c)] and Deniz Başkent[1,b),e)]

[1]*Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*

[2]*CNRS Unité Mixte de Recherche 5292, Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics, Institut National de la Santé et de la Recherche Médicale, UMRS 1028, Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France*

[3]*Department of Otolaryngology–Head & Neck Surgery, The Ohio State University Wexner Medical Center, The Ohio State University, Columbus, Ohio, USA*

**ABSTRACT:**

Perceptual differences in voice cues, such as fundamental frequency (F0) and vocal tract length (VTL), can facilitate speech understanding in challenging conditions. Yet, we hypothesized that in the presence of spectrotemporal signal degradations, as imposed by cochlear implants (CIs) and vocoders, acoustic cues that overlap for voice perception and phonemic categorization could be mistaken for one another, leading to a strong interaction between linguistic and indexical (talker-specific) content. Fifteen normal-hearing participants performed an odd-one-out adaptive task measuring just-noticeable differences (JNDs) in F0 and VTL. Items used were words (lexical content) or time-reversed words (no lexical content). The use of lexical content was either promoted (by using variable items across comparison intervals) or not (fixed item). Finally, stimuli were presented without or with vocoding. Results showed that JNDs for both F0 and VTL were significantly smaller (better) for non-vocoded compared with vocoded speech and for fixed compared with variable items. Lexical content (forward vs reversed) affected VTL JNDs in the variable item condition, but F0 JNDs only in the non-vocoded, fixed condition. In conclusion, lexical content had a positive top–down effect on VTL perception when acoustic and linguistic variability was present but not on F0 perception. Lexical advantage persisted in the most degraded conditions and vocoding even enhanced the effect of item variability, suggesting that linguistic content could support compensation for poor voice perception in CI users.

 https://doi.org/10.1121/10.0005938

(Received 21 April 2021; revised 29 July 2021; accepted 2 August 2021; published online 7 September 2021)

[Editor: Jody Kreiman] Pages: 1620–1634

## I. INTRODUCTION

Each human voice is unique, and being able to tell them apart can dramatically improve our ability to understand speech. This is especially true when there is noise in the background or when multiple people are talking at the same time (e.g., Cherry, 1953; Johnsrude *et al.*, 2013). For normal-hearing (NH) listeners, speech-on-speech perception is relatively easy when there is a clear difference between the voices of the speakers who produce the target speech and the masking speech (e.g., Brungart, 2001; Festen and Plomp, 1990; Stickney *et al.*, 2004). Research using target and masking speech produced by the same talker and where the voice difference was introduced artificially through acoustic manipulation has indicated that two acoustic voice cues are particularly important for speech-on-speech performance (Başkent and Gaudrain, 2016; Darwin *et al.*, 2003; Vestergaard *et al.*, 2011): fundamental frequency (F0), which arises from the glottal–pulse rate, and vocal tract length (VTL), which shapes the spectral parameters, such as formant frequencies.

Individual talkers can control their F0 by applying more or less tension on their glottal folds as they speak or sing. However, the average F0 they produce is also constrained by the anatomy of their speech production system.[1] Among adult male talkers, F0 varies with testosterone levels (Dabbs and Mallinger, 1999), whereas both F0 and VTL are known to vary with actual body size and shape (Evans *et al.*, 2006). Additionally, the morphology of the male and female VTL differs, resulting in an overall greater VTL for men (Fitch and Giedd, 1999). This variance in VTL leads to the variation in formant patterns observed for different talkers (Hillenbrand *et al.*, 1994). While speaker size is predominantly judged on VTL; judgment of talkers' sex and age is equally based on F0 and VTL (Smith and Patterson, 2005). Hence, for NH listeners, the F0 and VTL voice cues greatly

[a)]Electronic mail: t.koelewijn@rug.nl
[b)]Also at: Research School of Behavioral and Cognitive Neuroscience, Graduate School of Medical Sciences, University of Groningen, Groningen, The Netherlands.
[c)]Also at: Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.
[d)]ORCID: 0000-0003-0490-0295.
[e)]ORCID: 0000-0002-6560-1451.

https://doi.org/10.1121/10.0005938

help to discriminate between talkers from different or same sexes, identify specific voices belonging to one speaker or another, and help with understanding speech better in multi-talker listening.

### A. Cochlear implants (CIs) and vocoding

Users of CIs have difficulties in perceiving the two voice cues mentioned above, showing higher discrimination thresholds for F0 and VTL compared with NH listeners (Gaudrain and Başkent, 2018; Zaltz *et al.*, 2018), because of spectrotemporal degradations inherent to CI electrical stimulation of the auditory nerve (Başkent *et al.*, 2016b). When compared with simulation studies where the spectral resolution was degraded by means of vocoding, similar results were shown (Gaudrain and Başkent, 2015). Although these deficits in voice cue sensitivity can directly explain abnormal voice gender categorization among CI users (Fu *et al.*, 2005; Fuller *et al.*, 2014; Massida *et al.*, 2013; Meister *et al.*, 2016), it is less clear how they are related to speech-on-speech perception.

Speech-on-speech perception is particularly challenging for CI listeners (Stickney *et al.*, 2004, 2007; Zeng *et al.*, 2005). A number of studies previously used differing talkers for target and masker speech, and when the gender of target and masker talkers differed, some benefit was observed for speech-on-speech perception by some of the CI users (Cullington and Zeng, 2008; in children, Misurelli and Litovsky, 2015). However, when the same speaker was used for target and masker and the voice properties were systematically altered in F0 and VTL, a different picture emerged.

El Boghdady *et al.* (2019) showed that compared with NH listeners, CI users do not necessarily benefit from differences introduced in F0 and VTL cues to better perceive speech on speech. However, El Boghdady *et al.* also showed that this deficit in voice difference benefit was not correlated with the just-noticeable differences (JNDs) for F0 and VTL. Instead, they found that overall speech-on-speech understanding, independently from the voice difference, was correlated with better F0 and VTL JNDs. In a more recent study, El Boghdady *et al.* (2020) showed that spectral contrast enhancement, a stimulation strategy that enhances the contrast between peaks and troughs in the spectrum, could improve speech-on-speech perception. But again, they also found that spectral contrast enhancement did not improve the F0 and VTL JNDs and that it improved speech-on-speech perception by improving the target-to-masker ratio across all voice difference conditions rather than by making the voice cues more salient. Meister *et al.* (2020) used a similar paradigm with bilateral and bimodal CI users, and some of these CI users, especially bimodal ones, were able to derive a benefit from F0 differences (in line with a previous vocoder study that showed better F0 perception with simulated bimodal hearing, Başkent *et al.*, 2018). The relationship between voice discrimination and release from masking in speech on speech is therefore not as straightforward as one would hope, and further investigation is needed to understand better how speech understanding and voice discrimination are articulated in CI listeners and of varying CI configurations. As a first step, using vocoders is useful to explore the signal properties that are detrimental to this phenomenon.

### B. Interaction between voice and linguistic content

The speech signal simultaneously carries linguistic information about the content of the intended utterance and indexical information about the characteristics of the talker (Abercrombie, 1967). Linguistic and indexical information are closely linked in the perceptual processing of speech (Nygaard, 2008; Pisoni, 1997). As a result, even in listeners who have full access to the clean speech signal, the interpretation of acoustic features into voice cues or phonetic content sometimes interferes with each other. This is evidenced by the fact that talker discrimination has been found to be influenced by linguistic factors such as talker and language familiarity. Goggin *et al.* (1991) showed that voice identification is facilitated by language familiarity, an effect most prominent in monolingual listeners. In addition, more recent research suggests that the language familiarity effect is not based on language comprehension *per se* but more on familiarity with the native language's phonology (Fleming *et al.*, 2014). Fleming *et al.* showed that reversed native language sentences were rated as more dissimilar than reversed non-native (Chinese) sentences, confirming the relevance of language-specific phonological features (e.g., vowels and fricatives), which are preserved in reversed speech. The relevance of phonological knowledge to talker discrimination, and possibly to voice cue perception, is further supported by studies on individuals with developmental dyslexia. Perrachione *et al.* (2011) found that individuals with dyslexia show poorer talker discrimination than nondyslexics because of impaired phonological processing. Perea *et al.* (2014) found a similar deficiency in dyslexic children and adults. Both studies found a relationship between reading ability and talker discrimination ability in individual participants. Kadam *et al.* (2016) further found that phonological competence (reading ability) in individuals without dyslexia was related to talker voice learning.

Within the listener's native language, other factors also seem to affect voice perception. Ptacek and Sander (1966) studied the ability of participants to identify whether stimuli were uttered by young or old adults using prolonged vowels, time-reversed words, and words played normally. They observed a performance of 78%, 87%, and 99% correct, respectively. These results suggest that both acoustic "richness" and access to the semantic content of the speech material facilitate talker discrimination. It can be argued that words, being produced through a number of articulatory gestures, contain more information about the speaker than steady-state vowels. From a signal information point of view, it can be expected that acoustic signals that are more variable, i.e., with higher entropy (Shannon and Weaver, 1949)—e.g., evaluated as cochlea-scaled entropy (Stilp and Kluender, 2010)—may be more

favorable for speaker identification because they offer more facets on which to evaluate the identity. However, because it is difficult to manipulate the semantic content without changing the acoustic content as well, it can be challenging to disentangle the effect of acoustic variability from that of semantic variability. Time-reversed speech is useful in this instance as it arguably contains the same acoustic entropy while carrying no semantic content.

Although acoustic or linguistic variability—or entropy—within a stimulus may be advantageous for talker identification or categorization because it provides more cues, it may be detrimental for talker discrimination when the variability also occurs across the items that are being compared. For instance, Cleary and Pisoni (2002) observed that talker discrimination in pediatric CI users was better when identical sentences were used across talkers compared with when two different sentences were used. Cleary *et al.* (2005) failed to replicate this result using an adaptive procedure; however, they reported that "performance was much more variable in the presence of linguistic variability." Recently, Narayan *et al.* (2017) examined in more detail the effects of acoustic and linguistic variability across comparison items in a talker discrimination task by using rhyming words (close acoustic match but distant semantic content), the two elements of compound words (e.g., "day" and "dream," no acoustic match but related semantic content), or two entirely unrelated words. They found not only that rhyming words provided the highest sensitivity to voice differences ($d' = 1.98$, standard error [s.e.] $= 0.05$, $p < 0.03$ for all comparisons) but also that semantically related words ($d' = 1.87$, s.e. $= 0.05$) yielded higher sensitivity than unrelated words ($d' = 1.56$, s.e. $= 0.05$, $p < 0.001$).[2] They also found that participants were more biased toward reporting no talker difference when the words were rhyming ($c = 0.27$, s.e. $= 0.03$) or when they were semantically related as part of a compound word pair ($c = 0.18$, s.e. $= 0.03$) than when they were unrelated ($c = -0.12$, s.e. $= 0.03$, $p < 0.001$ for both cases). Quinto *et al.* (2020) replicated these results using only same-sex talkers and confirmed that sensitivity and bias are both affected by the acoustic/linguistic relationship between the compared words: When the words were rhyming or when they were semantically related, participants were at the same time more sensitive to the voice difference and biased toward judging them as originating from the same speaker. In that case, it seems that reducing acoustic or linguistic entropy across items helps to discriminate voices.

All the studies listed above involved actual talkers, which means that the perceptual difference between the talkers was not parametrically controlled, and the nature of the voice difference across talkers spans across multiple acoustic features. Quinto *et al.* (2020) limited their talkers to a single sex class in an effort to restrict the type of cue that may be available. However, it remains unclear whether these effects of acoustic and linguistic variability would persist when individual voice cues are manipulated (as was done by Cleary and Pisoni, 2002; and Cleary *et al.*, 2005).

Moreover, the studies by Narayan *et al.* (2017) and Quinto *et al.* (2020) concern NH listeners. While Cleary and Pisoni (2002) seemed to have observed some similar effects in pediatric CI users, they could not replicate this finding in subsequent observations (Cleary *et al.*, 2005). Therefore, it also remains unclear whether the presence of linguistic information can also improve voice perception in degraded conditions that resemble CI listening.

### C. Compensatory mechanisms in CI listening

Unlike NH listeners, CI users are routinely presented with speech that is affected by spectrotemporal degradations inherent to electrical stimulation of the ear (Başkent *et al.*, 2016b). These degradations bring further ambiguity in the signal, notably in the partition between voice cues and linguistic cues. When faced with such ambiguity, CI listeners are thought to rely heavily on cognitive compensation to correctly interpret the degraded signal and improve intelligibility (Amichetti *et al.*, 2018; Başkent *et al.*, 2016a; Nagels *et al.*, 2020a; Winn and Moore, 2018).

Top–down compensation can be evidenced using *phonemic restoration*, which was shown to occur in CI users with good speech in quiet performance (Bhargava *et al.*, 2014, but see also Jaekel *et al.*, 2021). This supports the idea that CI users can make effective use of linguistic context in a sentence to make lexical decisions, although some seem to be able to do this more efficiently than others (Nagels *et al.*, 2020a). On the other hand, degradation of the speech seems to reduce and delay semantic integration (Wagner *et al.*, 2016). Winn (2016) showed that high semantic context in a sentence reduces cognitive processing load (listening effort) relative to processing sentences with low semantic context. This decrease in listening effort, reflected by the pupillary response, was slightly delayed in CI users and more delayed in NH listeners listening to vocoded speech. Overall, these results indicate that CI users may be relying more on compensatory mechanisms based on linguistic context and knowledge than their NH counterparts. There is thus a possibility that voice perception could be more strongly affected by linguistic cues in CI users, or in general, by extension, in spectrotemporally degraded speech.

### D. Current study

The current study was designed to assess the effect of lexical content on voice cue sensitivity for normal and vocoded speech. The study was modeled after that of Gaudrain and Başkent (2015), except that the stimuli in the current study were meaningful words instead of meaningless consonant–vowel (CV) triplets. During the experiment, JNDs in F0, VTL, and their combination (F0 + VTL) were obtained using an auditory adaptive odd-one-out task (three intervals, three alternatives forced choice [3I-3AFC]). The stimuli were meaningful Dutch consonant–vowel–consonant (CVC) words presented in their forward (normal) or reversed time direction. Time reversing the speech signal disrupts the acoustic–phonetic attributes of speech that rely

on transients and prevents lexical access to the original words that were uttered (Ptacek and Sander, 1966). However, time-reversed speech preserves many talker-specific vocal cues, allowing for talker identification using voice cues alone without relying on talker-specific dynamic articulatory features (Perrachione *et al.*, 2019; Sheffert *et al.*, 2002; Van Lancker *et al.*, 1985). Crucially, the acoustic entropy of a reversed word is nearly identical to that of the forward version of that same word. By comparing forward with reversed words, the amount of acoustic entropy within items is controlled for while manipulating the amount of semantic information available.

The acoustic/linguistic variability across items (the item variability) was also manipulated orthogonally. During each trial, the three intervals could be either the same word repeated three times (one being uttered with a deviant voice) or three different words (one being uttered with a deviant voice). In conditions where the same word is presented to detect a voice cue difference, listeners can rely on the same word template from one interval to the next, whereas when three different words are presented, each word has to be compared with a learned template based on lexical knowledge.

The current study made use of vocoding in NH listeners as opposed to recruiting actual CI listeners. First, recruiting young adult NH listeners instead of CI participants, who are typically older, helps to reduce the variability that may be due to cognitive aspects related to age or to long periods of auditory deprivation. In addition, using vocoders not only removes the variability caused by physiological differences or differences in implant models, stimulation strategies, etc., that is found in actual implants but also makes it possible to control the type of degradation imposed. All these help with pinpointing the cause of a certain observed behavior to a specific sensory deficit. Gaudrain and Başkent (2015) have suggested that voice cue perception, and in particular VTL, is affected by spectral resolution. In actual implants, the spread of excitation that occurs in the cochlea defines how much interaction happens between individual channels of the implant (Black and Clark, 1980), thus constraining the spectral resolution of the evoked excitation pattern. In line with Gaudrain and Başkent (2015), we used two vocoders (one simulating low spread of excitation [LS-vocoder] and one simulating high spread of excitation [HS-vocoder]) and a no-vocoder condition. In both vocoder conditions, the 12 channels simulated the same number of electrode contact points, but the differences in filter spread represented different amounts of channel interaction (spectral smearing).

To summarize, the current experiment resulted in a dependent JND for each voice cue (F0, VTL, F0 + VTL) for words or time-reversed words presented fixed or variably across intervals for any of the three vocoding conditions.

### E. Hypotheses

In the current study, we investigated how lexical content (forward vs reversed words) and item variability (same or different items across intervals) affect the JNDs for F0 and VTL in normal and vocoded conditions. Our first hypothesis (H1) was that access to lexical content would lead to smaller JNDs for voice cues, namely, when forward words are presented compared with reversed words. For processing words, our internal lexical representation can be used as a reference, as discussed above, which could provide a benefit for voice cue perception compared with reversed words. With our second hypothesis (H2), we predicted that when there is no acoustic/linguistic (item) variability, i.e., when the presented words during the 3AFC task are all the same, a difference in voice cues would be more easily detected. When the same word is repeated, any detectable acoustic deviation can be used to identify the odd one out, and the F0 contour and formant patterns will be exactly identical. When the variability across intervals is high because three different words are presented, there is no direct acoustic reference, and mean F0 and VTL have to be extracted from different signals. Therefore, we hypothesized that voice cue discrimination would be better (smaller JNDs) when the same words are presented instead of three different words. Finally, our third hypothesis (H3) was that the effect of lexical content and item variability would interact with each another. We expected the benefit of lexical content on voice cue perception to be most prominent when item variability is highest when different words are presented relative to the conditions where the same words are presented. Indeed, when there is no item variability, the participants can solve the task without relying on lexical processing but only through comparison of acoustic templates.

Based on previous research (e.g., Gaudrain and Başkent, 2015), we expected through a fourth hypothesis (H4) a prominent decremental effect of vocoding on voice cue perception. Also, Gaudrain and Başkent (2015) showed smaller VTL JNDs for steep (12th-order) compared with shallower (4th-order) filters. Our fifth hypothesis (H5) was that the benefit of perceiving forward words compared with reversed words would be most observable during the vocoded conditions because listeners may rely more on lexical content to compensate for the relatively ambiguous acoustic–phonetic information to make their judgments (e.g., Başkent *et al.*, 2016a). Additionally, similar to the lexical content effect, with item variability, we predicted through our sixth hypothesis (H6) that the detrimental effect of item variability would be most prominent in the vocoded conditions. The ambiguity induced by the variability of words across intervals is harder to resolve when the stimuli are vocoded and fewer acoustic details are available.

Finally, although this might be difficult to assess reliably, we expected that the effects of linguistic variability and lexical content would be most prominent for VTL compared with F0 JNDs [hypothesis 7 (H7)]. VTL perception relies on the spectral envelope and likely more specifically on formant frequencies, which also carry phonetic properties of vowels (Chiba and Kajiyama, 1941). In contrast, average F0 seems less relevant to phonetic cues, offering less potential for interference between vocal and phonetic cues.

However, there are potential shortcomings in evaluating these interactive effects, as described at the end of Sec. II.

## II. METHODS

### A. Participants

Fifteen NH adults (self-reported gender: one transman, seven males, seven females; age range, 20–31 yr; median age, 22 yr), recruited at the University of Groningen and the University Medical Center Groningen, participated in the study. NH was defined as pure-tone thresholds ≤20 dB hearing level (HL) at octave frequencies between 0.25 and 8 kHz. Participants reported normal or corrected-to-normal vision and did not report dyslexia, epilepsy, or history of developmental disorders. They all were native Dutch speakers and provided written informed consent in accordance with the ethics committee of the University Medical Center Groningen (METc 2018/427). Participants received an hourly compensation of 8 euros in accordance with our departmental guidelines for participant reimbursement.

### B. Stimuli

Participants performed a 3AFC task. In this task, the participant had to choose which one of the three consecutive stimuli (words) sounded different from the other two. Although they could use any cue available, depending on the specific condition, the "voice cue" manipulated was F0 only, VTL only, or F0 + VTL. Note that most design choices considering the different stimuli manipulations mentioned below are based on previous research by Gaudrain and Başkent (2015, 2018).

The words were randomly picked from recordings of 144 Dutch meaningful CVC words, representing the first 12 word lists of the Nederlandse Vereniging voor Audiologie (NVA) corpus (Bosman and Smoorenburg, 1995). Within trials, three words were randomly selected without replacement, whereas between trials, words were replaced. These words were uttered by a native Dutch female talker and had a duration ranging from 386 to 1107 ms. The talker's voice had an average F0 of 242 Hz as measured over all the words. Because some of the word recordings stood out because of excessive sub–100-Hz levels (proximity effect), all audio files were high-pass filtered at 125 Hz (Butterworth eighth order) and matched for total root mean square (RMS). The words were presented in their forward (normal) or reversed direction, referred to as the manipulation of "lexical content." Items were presented at 65 dB sound pressure level (SPL) and in sequence with an interval of 450 ms between each of them.

The processing of voice cue changes was done with STRAIGHT (Kawahara and Irino, 2005). The voice cues (F0, VTL, or F0 + VTL) started at a voice distance corresponding to an adult male's voice (see Fig. 1), which was defined as having a VTL of 24.5% (3.8 semitones [st]) longer than that of the reference female voice, and an F0 half of that of the reference voice (−12 st). For more details on voice cue manipulation, see Gaudrain and Başkent (2015, 2018).

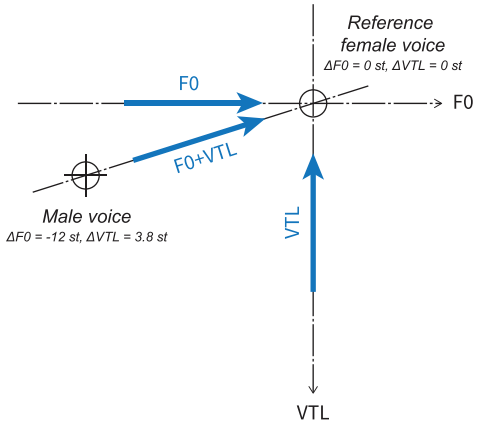


FIG. 1. (color online) F0-VTL plane with the reference female voice in the center and the male voice in the lower left corner. The thick (blue) arrows describe how the voice difference evolve during the adaptive procedure.

In analyzing data, we have used differing approaches for results of F0 and VTL JNDs and F0 + VTL JNDs. In real life, voice cues of F0 and VTL often change across differing talkers; hence, the F0 + VTL condition is useful for measuring such voice variability in the general population. However, these results by themselves are not informative about individual mechanisms of F0 and VTL because these covary. Therefore, to use it in a more informative way, we used the F0 + VTL condition as a reference condition and did not analyze it in the same way as the F0 and VTL conditions. Instead, we took advantage of having the individual F0 and VTL data and used these data to predict F0 + VTL to gain insight into the individual contribution of each cue to the overall voice perception.

During each trial, the three words presented were always from the same lexical content category (all *forward* or all *reversed*). However, within this category, the three items presented could be derived from the same word (*fixed*), or all three items could be derived from different words (*variable*), referred to as "item variability." Note that variability in words entails variability in acoustic content and phonological content in addition to variability in lexical content. However, the lexical content factor offers a way to disentangle the latter from the two others: When lexical content is not present (in the reversed condition), only acoustic and, to some extent, phonological variability persist. The participants were instructed to report which item was uttered with a different voice independently from the fact that the words the items were derived from could also vary within a trial.

Participants listened to items in a non-vocoding, LS-vocoding, and HS-vocoding condition. The vocoders were implemented in MATLAB (Gaudrain, 2016). For both vocoding conditions, the 12 analysis filters were 12th order [72 dB/octave (oct.)] zero-phase Butterworth filters spanning from 150 to 7000 Hz uniformly divided in terms of cochlear place of excitation (Greenwood, 1990). The synthesis filters were identical to the analysis filters for the LS-vocoding

https://doi.org/10.1121/10.0005938

condition, and were fourth-order filters (24 dB/oct.) for the HS-vocoding condition. In each frequency band, the temporal envelope was extracted by halfwave rectification and low-pass filtering (zero-phase fourth-order Butterworth filter) with a cutoff frequency of half the bandwidth of each band, with a maximum of 300 Hz (Gaudrain and Başkent, 2015).

### C. Procedure and apparatus

The JNDs were estimated separately for each condition using a two-down/one-up adaptive procedure (Levitt, 1971), which converged toward an average proportion of correct responses of 70.7% for each individual test. At the beginning of each test (adaptive run), the manipulated voice of the deviant item started away from the reference female voice and became progressively more similar by means of the adaptive procedure. For all cues, the initial distance was 12 st, calculated as the Euclidian distance in the F0–VTL plane (see Fig. 1). Each test started with a 2-st step size. After every 15 trials or when the voice difference became smaller than twice the step size, the step size was reduced by a factor of $\sqrt{2}$. The adaptive run ended after eight reversals or 150 trials. The JND was calculated as the average difference in semitones over the last six reversals.

During each trial of the 3AFC task, participants heard the three items play in sequence while at the same time three corresponding buttons lit up onscreen. To respond, they chose the button corresponding to the deviant item by means of a mouse click. Participants received visual feedback from the selected button with feedback blinking green for a correct response and red for an incorrect response. JNDs were obtained in a blockwise fashion for each voice cue (F0, VTL, F0 + VTL) for forward or reversed words presented fixed or variable at any of the three vocoder conditions.

Each of the 36 conditions was presented twice as separate adaptive tests, which resulted in a total of 72 tests presented for each participant in a random order divided over three 2-h sessions. At the start of the first session, the participant's audiogram was recorded, and the participant filled in a questionnaire concerning language, hearing, and demographic status. Before performing the first test of the first session, the participants performed a short training. This consisted of the first three trials of each condition presented in order of increasing vocoder difficulty (non-vocoding, LS-vocoding, and HS-vocoding, respectively), with the exception of the voice cue F0 + VTL conditions, resulting in a total of three times 24 practice trials, which in total took 10–15 min. During each 2-h session, approximately 24 tests were performed (in random order), which each took 4–5 min. If a test failed to converge because either the maximum number of trials was exceeded or the voice difference became too large, the same condition was attempted again. Halfway through each session, after 12 tests, participants took an obligatory 10-min break, and they were allowed to take short breaks between tests.

All testing was performed in a sound-treated room. During the experiment, participants were seated in front of a computer screen at a comfortable (approximately 60–80-cm) viewing distance. Audio was presented diotically through Sennheiser HD650 headphones *via* a MOTU (Ultra Lite mk4) soundcard connected to a DA10 D/A converter (Lavry Engineering, Poulsbo, WA). All tests were presented using a MATLAB script that ran on a computer (Mac mini; Apple, Cupertino, CA).

### D. Statistical analyses

As done in previous studies (e.g., El Boghdady *et al.*, 2018), the JNDs were log-transformed to improve the homogeneity of variance across conditions. As will be seen in Sec. III, the F0 and VTL JNDs in the non-vocoded, fixed condition were relatively similar (0.69 st and 0.85 st, respectively); however, the intersubject variance was very different for the two cues [0.21 and 0.08, respectively[3]; Levene's test $F_{(2,28)} = 5.75$, $p < 0.05$]. For this reason, the different voice cues were analyzed separately to avoid the risk of spurious interactions but complicating the assessment of H7.

For each voice cue, we performed a $2 \times 2 \times 3$ repeated-measures analysis of variance (ANOVA) on the log-transformed JNDs with lexical content (forward, reversed), item variability (fixed, variable), and vocoding (no, LS, HS) as the within-subject factors. All analyses were performed in R v4.3.0 (R Core Team, 2020); ANOVAs were computed with the ez package v4.4.0 (Lawrence, 2016) using type III sums of square. When the sphericity assumption was violated (as indicated by Mauchly's test), the Greenhouse–Geisser (GG) correction of degrees of freedom was applied while computing the $p$ values, which were then indexed with GG in Sec. III. Effect sizes are reported as generalized $\eta^2$ [$\eta_g^2$ (Bakeman, 2005)].

Because we were primarily interested in the effect of lexical content, additional planned comparisons were performed for each combination of voice cue, item variability, and vocoder, the effect of lexical content. Multiple comparisons were performed using two-tail $t$ test and using the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). Effect sizes for these tests, when present, are reported as Cohen's $d$.

The JNDs for the combined voice condition F0 + VTL were analyzed in a different way. We used a linear mixed-effects model to estimate the JNDs in the F0 + VTL condition as a function of the individual F0 and VTL JNDs; i.e., the log-JNDs for F0 + VTL were expressed as a linear combination of the log-JNDs for F0 and VTL. Because different participants may give different weights to different cues and because vocoders may lead them to re-evaluate the role of each cue, the coefficients were estimated per participant $i$ and per vocoder $v$:

$$\log\left(JND_{\mathrm{F0+VTL},i,v}\right) = \alpha_{i,v} \cdot log(JND_{\mathrm{F0},i,v}) + \beta_{i,v} \cdot \log(JND_{\mathrm{VTL},i,v}) + \varepsilon. \quad (1)$$

The $\alpha_{i,v}$ and $\beta_{i,v}$ coefficients were obtained with the `lmer` function of `lme4` package v.1.1.26 (Bates *et al.*, 2015).

## III. RESULTS

Figure 2 shows participants' JNDs for F0 (left) and VTL (right) as a function of lexical content, item variability, and vocoding. The results for the three vocoder conditions are shown side by side within each panel from left to right in order of least to most degraded. The item variability conditions can be visualized by comparing dark (purple) boxes for fixed items across intervals to light (yellow) boxes for variable items across intervals. The lexical content factor is represented with hatches: JNDs for forward words are shown with plain boxes, whereas JNDs for reversed words are shown with striped boxes. The outcomes of the two three-way ANOVAs, one for each F0 and VTL, are shown in Table I, with references to the specific relevant hypotheses.

The lexical advantage, i.e., the comparison between forward and reversed words, for each vocoder and item variability condition is shown in Fig. 3 and Table II.

### A. F0 JNDs

As expected from previous studies, the largest significant effect on the JNDs was imposed by the vocoding [$F_{(2,28)} = 254.15$, $p < 0.001$, $\eta_g^2 = 0.70$]: The JNDs in the HS-vocoder condition (7.7 st) were more than five times larger than those observed in the no-vocoding condition (1.5 st). Item variability had the second largest effect [$F_{(1,14)} = 135.89$, $p < 0.001$, $\eta_g^2 = 0.58$]: Fixed items over intervals [2.2 st, dark (purple)] yielded JNDs three times smaller than those obtained with variable items (6.6 st). However, the F0 JNDs were, on average, not significantly affected by the lexical content [$F_{(1,14)} = 4.35$, $p = 0.056$, $\eta_g^2 = 0.01$]: The average JND across all conditions for forward words (plain boxes) was 3.7 st against 4.0 st for the reversed words (striped boxes).

TABLE I. Repeated-measures ANOVA for the F0 and VTL JNDs as a function of voice cue, lexical content (LC, forward/reversed), item variability (IV, fixed/variable), and vocoding.[a]

| | | $F$ | $p$ | $\eta_g^2$ |
|---|---|---|---|---|
| F0 | | | | |
| H1 | LC | 4.35 [1,14] | 0.056 | 0.01 |
| H2 | IV | 135.89 [1,14] | **<0.001** | 0.58 |
| H3 | LC × IV | 0.03 [1,14] | 0.864 | <0.01 |
| H4 | Vocoding | 254.15 [2,28] | **<0.001** | 0.70 |
| H5 | LC × vocoding | 4.56 [2,28] | $_{(GG)}$ **0.022** | 0.02 |
| H6 | IV × vocoding | 21.58 [2,28] | $_{(GG)}$ **<0.001** | 0.13 |
| | LC × IV × vocoding | 3.28 [2,28] | $_{(GG)}$ 0.061 | 0.02 |
| VTL | | | | |
| H1 | LC | 64.17 [1,14] | **<0.001** | 0.12 |
| H2 | IV | 229.67 [1,14] | **<0.001** | 0.59 |
| H3 | LC × IV | 12.84 [1,14] | **0.003** | 0.03 |
| H4 | Vocoding | 380.20 [2,28] | **<0.001** | 0.75 |
| H5 | LC × vocoding | 0.58 [2,28] | $_{(GG)}$ 0.549 | <0.01 |
| H6 | IV × vocoding | 3.91 [2,28] | $_{(GG)}$ **0.035** | 0.03 |
| | LC × IV × vocoding | 0.96 [2,28] | $_{(GG)}$ 0.382 | <0.01 |

[a]The hypotheses corresponding to each test are indicated in the first column. LC and linguistic variability: H1, LC yields smaller JNDs; H2, when there is no item variability, JNDs are smaller; and H3, the benefit of lexical content on JNDs is most prominent when there is item variability. Vocoding: H4, vocoding yields larger JNDs; H5, LC helps most in the vocoded conditions; and H6, IV is most detrimental when there is vocoding. Boldface indicates significant results.

The effects of lexical content and item variability were independent from each other [$F_{(1,14)} = 0.03$, $p = 0.86$, $\eta_g^2 < 0.01$]. However, the effect of item variability decreased when the amount of degradation imposed by the vocoding increased [$F_{(2,28)} = 21.58$, $p < 0.001$, $\eta_g^2 = 0.13$]: Without vocoding, the JNDs obtained in the variable condition were more than four times larger than those obtained in the fixed condition, whereas with the HS-vocoder, they were only

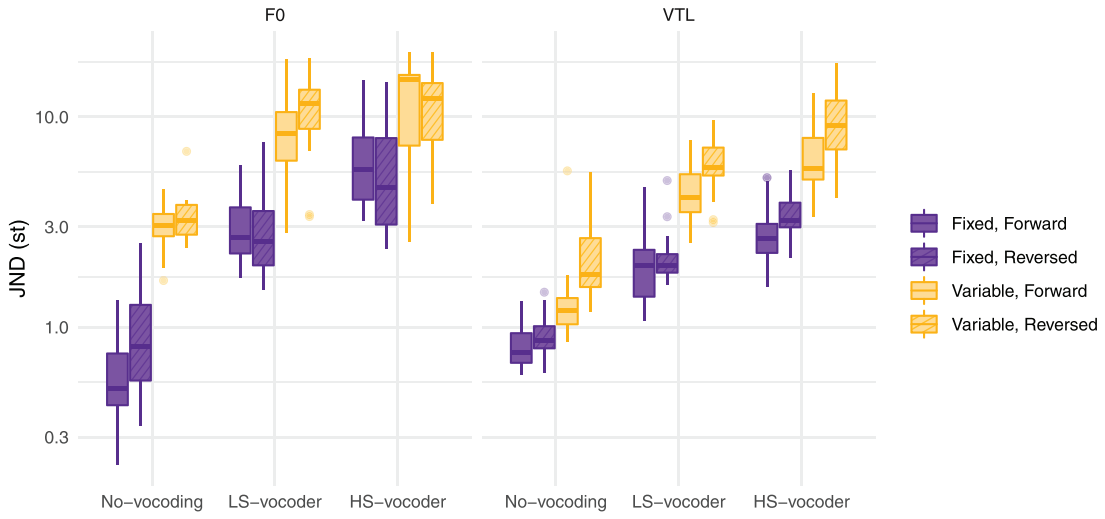


FIG. 2. (color online) F0 (left) and VTL (right) JNDs shown for each lexical content (forward, reversed), item variability (fixed, variable), and vocoding (no, LS, HS) condition. Boxes extend from the lower to the upper quartile (the interquartile range [IQ]), and the midline indicates the median. The whiskers indicate the highest and lowest values no greater than 1.5 times the IQ, and the dots indicate the outliers, i.e., data points larger than 1.5 times the IQ. The $y$ axis is log-spaced.

https://doi.org/10.1121/10.0005938

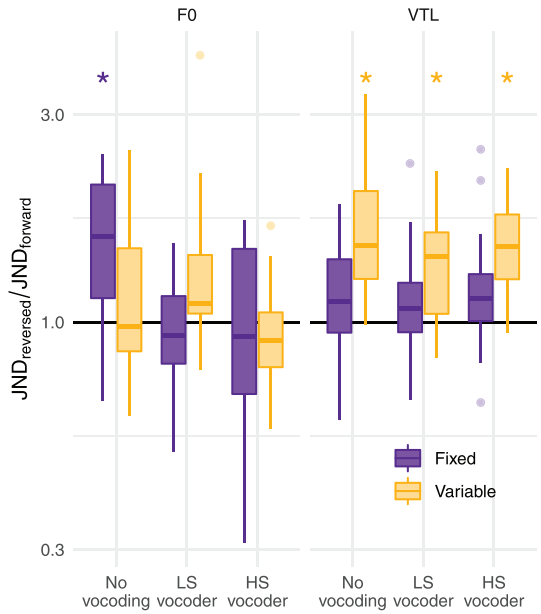


FIG. 3. (color online) Lexical advantage (calculated as the ratio of JNDs for reversed words over JNDs for forward words) for each voice cue (left, F0; right, VTL) and for each item variability (fixed, variable) and vocoding (no, LS, HS) condition. See Fig. 2 for the details of the boxes and whiskers. The solid black line represents equality between the two conditions (ratio = 1.0). Values >1.0 indicate an advantage (smaller JNDs) when lexical content is present (in the forward condition). The *y* axis is log-spaced. The asterisks indicate significant differences between forward and reversed words.

less than two times larger. Finally, the effect of lexical content also depended weakly on the level of vocoding [$F_{(2,28)} = 4.56$, $p_{GG} = 0.022$, $\eta_g^2 = 0.02$). In the no-vocoding condition, the forward words yielded JNDs that were 1.3 times smaller than in the reversed condition [$t_{(14)} = -3.45$, $p_{FDR} = 0.012$], whereas for the two vocoded conditions, the forward and reversed JNDs were not different from each other (LS-vocoder: $p_{FDR} = 0.35$; HS-vocoder: $p_{FDR} = 0.39$). This is the opposite of what was predicted in H5. The three-way interaction among vocoding, lexical content, and item variability was not significant [$F_{(2,28)} = 3.28$, $p_{GG} = 0.061$, $\eta_g^2 = 0.02$].

The effect of lexical content was only significant in the fixed, no-vocoding condition [$t_{(14)} = 3.53$, $p$FDR $< 0.01$, $d = 0.91$; all other conditions: $p_{FDR} > 0.10$, $d < 0.58$].

TABLE II. The outcomes of 12 planned comparisons in the form of paired-samples *t* tests, FDR corrected, between forward and reversed words for voice cue (F0, VTL), vocoding (no, LS, HS), and item variability (fixed, variable) as the within-subject factors.

| Voice cue | Vocoding | Item variability | *t* | $p_{FDR}$ |
|---|---|---|---|---|
| F0 | No | Fixed | 3.53 | 0.010[a] |
| | | Variable | 1.04 | 0.424 |
| | LS | Fixed | –0.85 | 0.445 |
| | | Variable | 2.24 | 0.101 |
| | HS | Fixed | –0.53 | 0.605 |
| | | Variable | –0.88 | 0.445 |
| VTL | No | Fixed | 1.08 | 0.424 |
| | | Variable | 5.06 | 0.001[a] |
| | LS | Fixed | 1.46 | 0.285 |
| | | Variable | 3.62 | 0.010[a] |
| | HS | Fixed | 1.84 | 0.174 |
| | | Variable | 6.03 | <0.001[a] |

[a]Significant results.

## B. VTL JNDs

Similar to F0 JNDs, the VTL JNDs were primarily affected by the vocoder [$F_{(2,28)} = 380.2$, $p < 0.001$, $\eta_g^2 = 0.75$] and the HS-vocoder (4.8 st), yielding JNDs more than four times larger than without vocoding (1.18 st). Item variability also had an overall significant effect [$F_{(1,14)} = 229.7$, $p < 0.001$, $\eta_g^2 = 0.59$]: Variable items across intervals (4.0 st) yielded JNDs more than twice as large as in the fixed condition (1.7 st). Finally, unlike for F0, lexical content did affect VTL JNDs significantly [$F_{(1,14)} = 64.17$, $p < 0.001$, $\eta_g^2 = 0.12$]: JNDs for the reversed words (3.0 st) were 30% larger than for the forward words (2.3 st).

This lexicality advantage effect did not depend significantly on the vocoder [$F_{(2,28)} = 0.58$, $p = 0.57$, $\eta_g^2 < 0.01$], but it did depend on the item variability, although that effect was small [$F_{(1,14)} = 12.84$, $\rho < 0.01$, $\eta_g^2 = 0.03$]: In the variable condition, the VTL JNDs for reversed words (4.8 st) were almost 50% larger than for the forward words (3.3 st), whereas in the fixed condition, the JNDs for the reversed words (1.9 st) were <20% larger than for forward words (1.6 st). Similarly, the effect of item variability was slightly magnified through vocoding [$F_{(2,28)} = 3.91$, $p < 0.05$, $\eta_g^2 = 0.03$]: The JNDs in the variable condition were 1.9 times larger than those in the fixed condition when no-vocoding was applied, whereas this ratio increased to 2.4 and 2.5 for the LS- and HS-vocoder conditions, respectively. The three-way interaction was not significant [$F_{(2,28)} = 0.96$, $p = 0.39$, $\eta_g^2 < 0.01$].

## C. Comparison of voice cues and lexicality advantage

Because the variance was widely different for the two voice cues in the non-vocoded, fixed condition, the results were analyzed using two separate ANOVAs, preventing direct comparisons of F0 with VTL JNDs. However, a lexicality advantage can be defined by calculating the individual differences between the (log) JNDs obtained with the forward and reversed words (Fig. 3). For this derived measure, the variance is homogenous enough across conditions to allow comparisons across the voice cues [Levene's test, $F_{(11,168)} = 1.15$, $p = 0.32$].

A repeated-measures ANOVA was performed on the lexicality advantage, with voice cue (F0 vs VTL), variability (fixed vs variable), and vocoder (no-vocoding, LS-vocoder, HS-vocoder) as repeated factors. The lexicality advantage was larger for VTL than for F0 [$F_{(1,14)} = 7.81$, $p < 0.05$, $\eta_g^2 = 0.05$]: For F0 JNDs, lexical content gave an advantage of almost 10% on average, whereas it gave an advantage of almost 30% for VTL. Item variability also significantly boosted the lexicality advantage [$F_{(1,14)} = 11.9$, $p < 0.01$, $\eta_g^2 = 0.03$]: The lexical advantage was about 15% for the

fixed condition against 36% for the variable condition. These two factors interacted significantly [$F_{(1,14)} = 5.05$, $p < 0.05$, $\eta_g^2 = 0.03$]: The lexicality advantage on F0 JNDs only increased from 8.5% to 9.6% when variability was introduced, whereas the lexicality advantage on VTL JNDs more than tripled from 13% to 46% when variability was introduced. Interestingly, the vocoder did not have any significant effect [$F_{(2,28)} = 2.77$, $p = 0.08$, $\eta_g^2 = 0.04$] and did not interact with any factor [$F_{(2,28)} < 3.61$, $p_{GG} > 0.055$, $\eta_g^2 \leq 0.05$].

Note that the reported effects, whether significant, are all small. This is also visible in Fig. 3: The variability across participants is large compared with the lexicality advantage. In fact, the lexicality advantage was significant only in 4 of the 12 tested conditions: for F0 JNDs in the fixed, no-vocoding condition [$t_{(14)} = 3.53$, $p_{FDR} < 0.01$, $d = 0.91$] and for VTL JNDs in the variable condition only for the three vocoder conditions [no-vocoding: $t_{(14)} = 5.06$, $p_{FDR} < 0.01$, $d = 1.31$; LS-vocoder: $t_{(14)} = 3.62$, $p_{FDR} < 0.01$, $d = 0.94$; HS-vocoder: $t_{(14)} = 6.03$, $p_{FDR} < 0.001$, $d = 1.56$; all other conditions: $p_{FDR} > 0.10$, $d < 0.58$]. Therefore, despite the small size of the effect, these results provide general support for H7.

### D. Predicted and observed F0 + VTL JNDs

The JNDs in the F0 + VTL condition were predicted using a linear mixed-effects model and the individual F0 and VTL JNDs. The correlation between the predicted and the observed F0 + VTL JNDs, shown in Fig. 4(a), yielded a correlation coefficient of 0.91 across participants, vocoders, lexical content, and item variability. Figure 4(b) shows that the pattern of predicted F0 + VTL JNDs across conditions is consistent with that of the observed values. The average of coefficients across participants and vocoders was 0.55 for F0 and 0.43 for VTL. As can be seen in Fig. 4(c), the F0 coefficients [the coefficient $\alpha$ from Eq. (1)] decreased from 0.60 to 0.51 when the amount of degradation applied with the vocoder increased from no-vocoding to HS-vocoder [$F_{(2,28)} = 7.17$, $p < 0.01$, $\eta_g^2 = 0.25$]. In contrast, [4] the VTL coefficients [the coefficient $\beta$ from Eq. (1)] did not significantly depend on the vocoder [$F_{(2,28)} = 2.45$, $p = 0.10$, $\eta_g^2 = 0.12$].

## IV. DISCUSSION

In this study, we investigated the effect of lexical content, item variability, and vocoding on the JNDs of the F0 and VTL voice cues. As was predicted in the hypotheses, the presence of item variability (H2) and the introduction of acoustic degradations through vocoding (H4) made F0 and VTL discrimination worse. The effect of lexical content was also as predicted in H1, except that it was limited to VTL JNDs (H1, H7) and to when item variability was present (H3), i.e., when the items used in the odd-one-out task differed across intervals. Vocoding did not enhance the benefit of lexical content (not confirming H5), but it did slightly increase the deficit induced by item variability (confirming H6).

### A. Acoustic vs linguistic variability

As predicted by H2, the presence of item variability yielded larger JNDs with variable compared with fixed items across intervals. This is in line with previous research showing that linguistic variability hinders voice discrimination (Cleary *et al.*, 2005; Cleary and Pisoni, 2002). However, as described by these authors, linguistic and acoustic variability were confounded in these studies. Thus, it remains unclear whether voice discrimination is hindered by the acoustic variability that the item variability introduces or the linguistic relationship that exists between the compared elements. Narayan *et al.* (2017), and more recently Quinto *et al.* (2020), further investigated this issue by comparing rhyming words (which differ in semantic content but are similar acoustically) to the elements of compound words (which are highly related semantically but differ acoustically). They found that the acoustic and semantic relationships both provided an advantage compared with unrelated words that differ both semantically and acoustically (although Quinto *et al.*, 2020, only found this effect

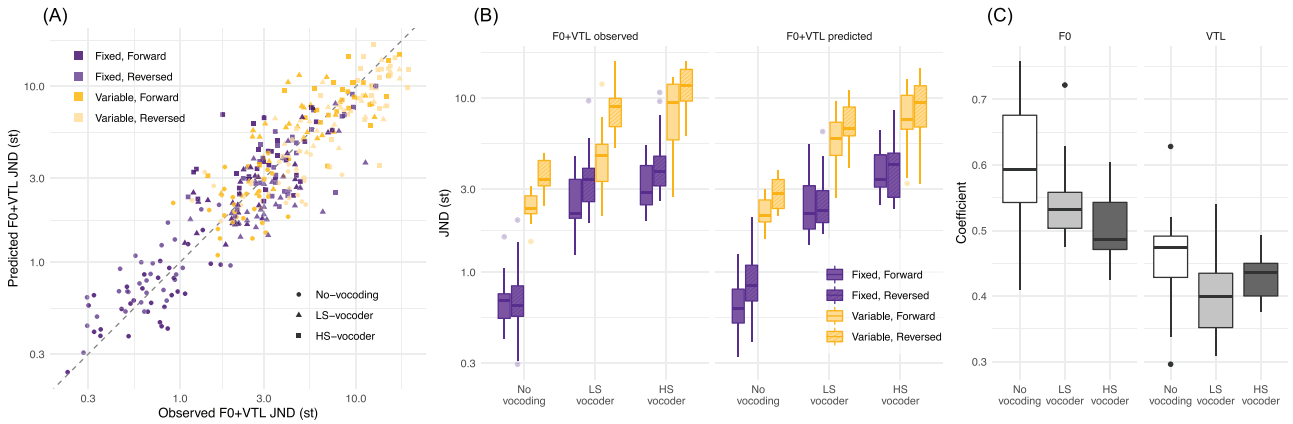


FIG. 4. (color online) (a) Predicted F0 + VTL JNDs as a function of observed F0 + VTL JNDs (see the text for details). The dashed diagonal line represents equality between the two. (b) Same as Fig. 2 for observed F0 + VTL JNDs (left) and predicted F0 + VTL JNDs (right). (c) Coefficients attributed to the individual F0 (left) and VTL (right) JNDs in the prediction of F0 + VTL JNDs. See Fig. 2 for a description of the boxes and whiskers.

https://doi.org/10.1121/10.0005938

when compound word pairs were presented in reverse order). Moreover, Narayan *et al.* (2017) found that the rhyming relationship provided a larger advantage than the compound word pair relationship.

In the present study, like in the studies of Cleary and Pisoni (2002) and Cleary *et al.* (2005), the compared items either repeated or differed, thus confounding acoustic and linguistic variability. However, the voice differences that the participants had to detect differed in how likely they were to interact with linguistic content. In stress accent languages, such as Dutch and English, F0 contours, as part of the prosodic cues, matter for stress patterns of individual words (Cutler *et al.*, 1997). While not very common, the lexical stress can contribute to linguistic content and help to disambiguate word meaning, even if uttered in isolation (van Leyden and van Heuven, 1996). Mean F0, however, does not seem to contribute to linguistic content *per se* in these languages. VTL, on the other hand, affects formant frequencies and is therefore more prone to interact with linguistic content (Ladefoged and Broadbent, 1957), yet, our results show an effect of item (acoustic/linguistic) variability not only on VTL JNDs but also on F0 JNDs. This finding suggests that acoustic similarity between items helps in the odd-one-out task.

Furthermore, we also manipulated the presence of lexical content by using forward and reversed words. For the F0 JNDs, the presence of lexical content did not affect the JNDs, except when there was no acoustic/linguistic variability and the stimuli were not vocoded. When variability was introduced among items, or when vocoding was applied, the presence of lexical content had no effect on the JNDs. This pattern of result is broadly consistent with the idea that it is acoustic variability, rather than linguistics, that is responsible for the inflated JNDs in the variable condition. However, the reliance on acoustic templates alone would not explain the difference between forward and reversed words in the fixed, no-vocoding condition.

A potential explanation for the effect of lexical content in this specific condition hinges on the fact that the task requires participants to compare the F0 contours that are not flat (see Fig. 5). When the mean F0 is modified, the entire F0 contour is shifted by a number of semitones. There are two ways listeners can then perform the task: (i) They can extract the mean F0 from that F0 contour for each stimulus and compare these (average and compare), or (ii) if the contours are identical but simply shifted, they can estimate the pitch difference point by point along the contour and average these comparisons (compare and average). The first method would suffer from the amount of variability that exists within the contour. In a signal-detection theory framework (Green and Swets, 1988), the estimate of the average F0 obtained through temporal integration would be a normal distribution with a width related to the variance of the F0 within the contour. Because the average standard deviation of the F0 contours around their mean is 4.1 st, the F0 difference corresponding to a $d'$ of 1 would be 4.1 st. In other words, if the average-and-compare method was used, we would not expect F0 JNDs to be much better than 4 st. Because we observed JNDs of $<1$ st in the fixed condition, it seems more likely that in this condition, listeners were using the second strategy, i.e., compare and average.

With this strategy, assuming that the F0 contour of the odd stimulus is exactly identical but shifted, limitations to how small a difference can be detected are related to purely sensory limitations and to the ability of the listener to keep an accurate trace of the contour in memory. In the latter case, linguistic knowledge may provide some assistance by helping the listener to predict or encode the shape of the contour. Examining the F0 contours (see Fig. 5), while most show a deviation of about 5 st away from the mean, this deviation does not take place at the same time point. The contours can be grouped into two classes based on when the positive excursion above the mean takes place: before 300 ms, which coincides with words starting with the letter

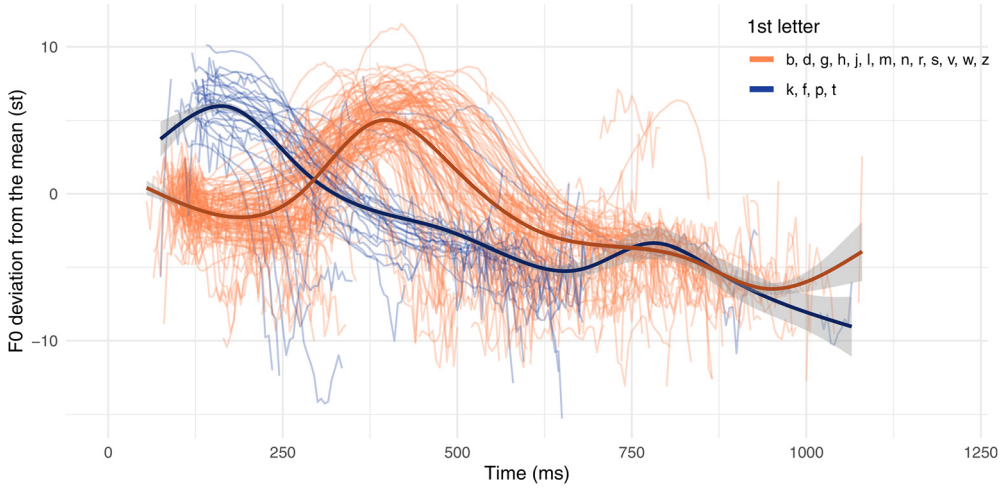


FIG. 5. (color online) F0 contours of the NVA words used to measure the JNDs. The F0 is expressed in semitones relative to the average F0 for each word. The darker (blue) lines correspond to words starting with the letters *k*, *f*, *p*, or *t*, whereas the lighter (orange) lines correspond to the other words. The thick, solid lines are spline fits for each of the two groups. The shaded area indicates the 95% confidence interval.

*k*, *f*, *p*, or *t* (short, unvoiced consonants), or after 300 ms, which coincides with words starting with other letters. It can be speculated that when the words are presented in a forward fashion, upon hearing the first sound, the listeners are able to predict the general shape of the F0 contour. However, when the words are presented time reversed, the listeners cannot use this association to predict the shape of the F0 contour. In other words, the presence of lexical content may facilitate the retention of the contour in memory because it offers support for the location of the inflection along the contour.

In the other conditions (variable and/or vocoded), the F0 JNDs seem insensitive to lexical content and seem on par with what would be predicted from purely acoustic considerations. In other words, apart from the fixed, non-vocoded condition, F0 JNDs seem to be dominated by acoustic, rather than linguistic, variability. However, the situation is different for VTL JNDs because the acoustic cues that support the perception of VTL are also directly related to the perception of phonetic information.

### B. Ambiguity resulting from the overlapping role of formant cues

The pattern of results observed for the VTL JNDs is consistent with the idea that when there is ambiguity in whether a difference between items could be related to voice (which participants are asked to detect) or to phonetic information (which participants are asked ignore), lexical content may help to resolve it. This type of ambiguity primarily occurs when the phonetic content differs across intervals, i.e., in the variable condition, and when the voice cue is coded on acoustic features that overlap with acoustic features coding for phonetic content, i.e., formants, when the voice cue is VTL. In the rest of the discussion, these effects are described in more detail and in relation to vocoding.

In the variable condition, different CVC words were presented, each potentially containing a different vowel. Different vowels are produced through articulation by shaping the mouth cavity with the tongue and lips, which affects the formant frequencies (and shapes). Formant patterns, that can be defined as the relative formant spacings is a strong acoustic correlate to vowel identity (Chiba and Kajiyama, 1941). In the present study, in addition to the different vowel-related formant spacings, VTL was manipulated by translating the formant pattern, as is on a log-frequency axis. This means that all formant frequencies were multiplied by the same ratio or shifted by the same amount in semitones. Therefore, in the presence of linguistic variability, i.e., in the variable condition, the listener has to deal with two sources of entropy across intervals that are both related to formant frequencies. The listener's task is to ignore the vowel-related formant variability and detect the VTL-related formant variability. In contrast, in the fixed condition, the only source of formant variability is the VTL manipulation they are tasked with detecting, thus yielding smaller JNDs.

Furthermore, small VTL changes—as those happening close to the JND—may result in ambiguous stimuli where the same acoustic manipulation could be interpreted either as a change in vowel identity or as a change in VTL. Consistent with H1, the presence of lexical content (in the forward condition) seems to help listeners to resolve this ambiguity. Lexical knowledge indeed provides an indication about the likelihood of a specific CVC pattern to be part of the language. Using this information, listeners can better decide how likely it is that a given formant pattern results from a given vowel with a given VTL rather than another vowel with a differing VTL. Here is a hypothetical example: The Dutch word *kip* (meaning chicken), pronounced /kɪp/, may be heard as /kʏp/ (which would be spelled *kup*) after the VTL is artificially elongated, which results in the formants being shifted down. However, because the word *kup* does not exist in Dutch, the listeners would be likely to correctly identify that the sound /kʏp/ is in fact /kɪp/ with a longer VTL, thus detecting the odd one out in the task.

When lexical content cannot be accessed because the words have been time reversed, the ambiguity cannot be lifted using lexical information anymore. Yet previous studies, consistent with our results, have indicated that talker-specific vocal cues can still be perceived in reversed speech (Perrachione *et al.*, 2019; Sheffert *et al.*, 2002; Van Lancker *et al.*, 1985). Furthermore, Fleming *et al.* (2014) found that the language-familiarity advantage to speaker discrimination was preserved when lexical content is rendered inaccessible through time reversal. These observations suggest that phonetic information, as opposed to lexical content, may still support VTL discrimination in the variable, reversed condition: If through VTL manipulation a vowel's identity is modified toward a vowel that is not contrastive in the language, the listener will be more likely to interpret it as a VTL change than a shift in vowel identity.

Finally, when the three items in the task are identical—in the fixed condition—there is no such ambiguity that arises because any perceivable difference between the intervals becomes relevant to the task.

In short, linguistic variability and VTL differences both affect formant frequencies. Lexical content allows for a better decoding of changes in formant frequencies related to vowel identity in the different CVC words, which in turn allows for better VTL difference detection.

### C. Compensation for degradations caused by vocoding

In this study, we used vocoding to degrade the speech signal to simulate some aspects of degradations CI users experience and which impair phonological processing. As predicted by H4 and in line with previous research (e.g., Gaudrain and Başkent, 2015), vocoding the presented speech stimuli clearly hindered the participants' ability to detect small differences in both F0 and VTL voice cues. Remarkably, for VTL JNDs, the lexical content effect occurred with and without vocoding, suggesting that listeners did not rely more, or less, on lexical information to

compensate for the relatively degraded acoustic–phonetic information to make their judgments. This goes against H5 that the effect of lexical content would be most observable during the vocoding conditions as a compensatory mechanism for reduced voice cues due to vocoding (Başkent *et al.*, 2016a). One possible explanation is that as vocoding is applied and the partition between voice and phonetic cues becomes ambiguous, so does the semantic content of the items. In other words, while reliance on top–down semantic knowledge may increase when the stimuli are vocoded, the quality of the information that would be used to trigger this knowledge is also degraded, thus hindering the efficiency of such top–down compensation. This is consistent with previous reports indicating that top–down restoration can take place when the stimuli are lightly degraded but may decrease when the degradation becomes substantial (Bhargava *et al.*, 2014; Clarke *et al.*, 2016, but also observed between mild and moderate levels of hearing impairment as in Başkent *et al.*, 2010). The fact that the detrimental effect of acoustic/linguistic variability was enhanced with vocoding also supports the idea that participants were less able to resolve ambiguities in vocoded conditions.

### D. Effects of vocoding on voice cues

As described earlier, directly comparing the JNDs for two voice cues is problematic. First, conceptually, there is no reason to expect that F0 JND should be comparable to the VTL JNDs, even if they are both expressed in semitones for data analyses purposes, since they stem from potentially very different physiological mechanisms. Second, statistically, while it may be interesting to compare the effects of some factors across the voice cues (rather than directly comparing the JNDs themselves), this is complicated by situations where the variance is unequal across voice cues. Qualitatively, the average F0 JNDs in the fixed, no-vocoding condition seem smaller than the VTL JNDs, but vocoding seems to make F0 JNDs somewhat larger than the VTL JNDs in both vocoded conditions.

Another way of assessing the relative degradations on F0 and VTL is to estimate how much weight listeners put in each cue in the F0 + VTL condition and how this weighting changes with vocoding. In this analysis (Fig. 4), we found that the weights given to VTL did not depend on the amount of vocoding, whereas the weight given to F0 did and decreased significantly.

This result is in line with previous research involving vocoders (Gaudrain and Başkent, 2015) or actual CI users (El Boghdady *et al.*, 2019, 2020; Gaudrain and Başkent, 2018), where the effects of spectral degradations on the voice JNDs tend to me more dramatic for F0 than for VTL. As proposed in these studies, VTL concerns broad spectral cues that are less likely to be affected by spectral resolution. In contrast, the spectral cues to F0 fundamentally rely on the perception of harmonic structure, which consists of small spectral details.

Gaudrain and Başkent (2015) also suggested that depending on vocoder parameters, F0 could also be provided through temporal periodicity cues in this task. However, our LS-vocoder condition is identical to the 12-band condition of Gaudrain and Başkent's experiment 1, where they found that temporal pitch cues were likely too weak to be useful compared with their four-band condition. It is also worth noting that while temporal modulation cues are enhanced when spectral resolution is limited by reducing the number of channels, this is not the case when it is reduced by means of increased spread of excitation. Fewer channels over the same frequency range mean that the analysis filters are broader and thus capture a larger number of harmonics, yielding deeper modulation. In contrast, a larger spread of excitation mixes channels that have slightly different patterns of modulation, which results in some temporal smearing and thus shallower modulation. Results in CI users are compatible with these considerations: There is evidence that CI users do not strongly rely on temporal cues to perceive voice pitch (Fielden *et al.*, 2015), and Nogueira *et al.* (2020) have found that using parallel stimulation across multiple electrodes, which increases the interaction between channels, did not improve F0 JNDs but, instead, had an adverse effect.

However, it is worth noting that this relationship between F0 and VTL is limited to a measure of sensitivity and likely does not generalize to other pragmatic use of these cues in other tasks. For instance, Fuller *et al.* (2014) found that both NH participants listening to vocoded stimuli and CI listeners relied on F0 more than VTL compared with NH participants listening to non-vocoded stimuli while estimating the gender of a voice.

### E. Comparison with other studies using different stimuli

The current study was partly a follow-up to Gaudrain and Başkent (2015) with a focus on the effect of lexical content using meaningful words instead of meaningless CV triplets. Although sensitivity to the VTL voice cue is impaired for CI users, the present results show that the effect of lexical content on VTL JNDs was not affected by vocoding, meaning that NH listeners benefited from lexical content while listening to both non-vocoded and vocoded speech. Previous studies using CV triplets instead of words might have missed this top–down effect since they only activated a limited assortment of linguistic processing.

For non-vocoded stimuli and using a fixed mode of presentation, Gaudrain and Başkent (2015) reported an average F0 JND of 2.68 st (although one outlier was pulling the average up; without this one subject, the average drops to 1.40 st), which is quite larger than the JNDs obtained in the fixed, non-vocoded condition with an average of 0.69 st. But other studies found more similar average F0 JNDs: Başkent *et al.* (2018) found an average of 0.81 st, and Nagels *et al.* (2020b) found an average of 0.79 st. Regarding VTL, Gaudrain and Başkent (2015) reported an average VTL JND of 1.62 st under the same conditions vs an average of 0.85 st

in the present study. Başkent *et al.* (2018) found an average of 1.24 st, and Nagels *et al.* (2020b) found an average of 1.08 st.

Thus, the JNDs reported in previous studies with CV triplets tended to be slightly larger than those reported here. The different nature of the stimuli may potentially explain this discrepancy. The CV triplets were composed of three 200-ms CV syllables. The average F0 of each CV was adjusted such that the triplet formed a contour that had a rather small magnitude (one third of a semitone) but was variable across intervals, thus making it more difficult to use the compare-and-average method described previously. Moreover, even without this artificial contour across CVs, with three syllables being involved, the F0 contour may have been more interrupted and more variable than in the CVC words used in the present study, thus further complicating the extraction of average F0. With these considerations in mind, it is not surprising that the F0 JNDs obtained with CV triplets fall in between the JNDs reported here for the fixed and variable conditions.

The same comment on variability may apply to the VTL JNDs: The stimuli used in the present experiment contained only one vowel or diphthong, whereas the CV triplets used by Gaudrain and Başkent (2015) and subsequent studies had three vowels. In a fixed condition, this extra variability on formants could provide more cues to extract VTL. However, in the current study, it seems that this variability works against the listener and makes the extraction of VTL more difficult. It is also worth mentioning that the CVs in these previous studies were assembled randomly without any consideration of syllable co-occurrence frequency. It is also possible that the CV triplets contained unlikely combinations that may have been distracting for the participant, further reducing performance.

Gaudrain and Başkent (2015) also compared low- and high-spread vocoders in their experiment 3 but only for VTL JNDs. For their 12-band noise vocoder, they observed VTL JNDs of 2.97 st for 12th-order filters (corresponding to our LS-vocoder condition) and 4.59 st for 4th-order filters (corresponding to our HS-vocoder condition). In contrast, the average VTL JNDs from the present study in the fixed condition were 2.03 st and 3.04 st for the LS- and HS-vocoder conditions, respectively. Strikingly, the ratio of the JNDs between these two vocoder conditions of 1.5 is precisely the same in the two studies.

In contrast, using full sentences as stimuli, Zaltz *et al.* (2018) reported smaller JNDs than the ones reported here: an average F0 JND of 0.63 st and an average VTL JND of 0.58 st. Their design was such that not only the same sentence was used in the three intervals, as in our fixed condition, but also the sentence remained the same for the entire JND measurement. This configuration provides a wide variation of articulatory gestures by providing a segment of a full sentence while also providing the highest predictability and by using the same sentence across trials. One can argue, however, that the predictability was mostly acoustic and phonetic, perhaps lexical, but that the semantic content was not highly predictable because the sentences were extracted from a matrix corpus.

Future research using words with various linguistic properties and relationships, or full sentences that carry semantic context, could provide more insight into how cognitive mechanisms interact with voice perception. In CI users, such situations might show a higher degree of cognitive compensation, leading to a magnified benefit of linguistic relationships between intervals in the task. In addition, complementary measures that reflect the actual processing load during this task could be used to investigate the impact of the lexical content and linguistic variability on listening effort.

### F. Conclusions

This study showed that there is an interaction between lexical content and voice perception, even when cues were degraded by means of vocoding. Lexical content seemed to have a positive top–down effect on VTL perception when linguistic variability was present but not on mean F0 perception. Interestingly, the lexical advantage remained even for the most degraded conditions. This could suggest that top–down mechanisms relying on linguistic content could be used by CI users as a compensatory strategy (Başkent *et al.*, 2016a). Still, it is important to note that NH participants had only short-term exposure to vocoded speech, whereas CI users might show cortical plasticity over time. Future research has to show that relying on lexical content or linguistic relationships could also benefit actual CI users, improving their voice discrimination and, in turn, perceived voice gender categorization (e.g., Fuller *et al.*, 2014) and speech-on-speech listening (e.g., El Boghdady *et al.*, 2019), other perceptual tasks that rely on voice perception and that pose a challenge for implant users.

## ACKNOWLEDGMENTS

The authors thank Olivier Crouzet for his help on aspects related to phonetics. This work was funded by a Vici grant (918-17-603) from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw) to D.B., a Veni grant (275-89-035) from the NWO to T.T., the Heinsius Houbolt Foundation, and a Rosalind Franklin Fellowship. The study was performed within the framework of the Laboratoire d'Excellence Centre Lyonnais d'Acoustique (ANR-10-LABX-0060) of Université de Lyon within the program "Investissements d'Avenir" (ANR-16-DEX-0005) operated by the French National Research Agency (ANR) and is part of the research program of the Department of Otorhinolaryngology, University Medical Center Groningen: Healthy Aging and Communication. The raw data presented here can be accessed online at https://doi.org/10.34894/7TLCU9.

[1]Unless stated otherwise, further mentions of F0 will refer to *average* F0 rather than to the instantaneous F0 variations that are used, for instance, to carry intonation.

[2]In the original article, hit rates and correct rejection rates are reported. The authors generously shared their data for us to compute the sensitivity $d'$ and criterion $c$ (Green and Swets, 1988) reported here. The comparisons are Holm-corrected $t$ tests.

[3]Note that this is the intersubject standard deviation of the $\log_{10}$(JND) and is hence reported without units.

[4]Strictly speaking, to conclude that the effect of vocoding differs for the two voice cues, the interaction between the vocoding factor and the voice cue factor has to be examined: $F_{(2,28)} = 6.34$, $p < 0.01$, $\eta_g^2 = 0.06$ (despite the significance level, note the very small effect size). However, it is worth noting that there is no reason for the F0 coefficient and the VTL coefficient to be identical, and although they are numerically not wildly different, there is indeed a strong main effect of voice cue on the coefficients [$F_{(1,14)} = 155.3$, $p < 0.001$, $\eta_g^2 = 0.47$]. Caution is thus mandated while interpreting this interaction, and we thought it was best to keep it at the simple comparison of the two vocoding effects.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine Publishing Co, Chicago), pp. 1–17.

Amichetti, N. M., Atagi, E., Kong, Y.-Y., and Wingfield, A. (**2018**). "Linguistic context versus semantic competition in word recognition by younger and older adults with cochlear implants," Ear Hear. **39**, 101–109.

Bakeman, R. (**2005**). "Recommended effect size statistics for repeated measures designs," Behav. Res. Methods **37**, 379–384.

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., and Gaudrain, E. (**2016a**). "Cognitive compensation of speech perception with hearing impairment, cochlear implants, and aging: How and to what degree can it be achieved?," Trends Hear. **20**, 233121651667027.

Başkent, D., Eiler, C. L., and Edwards, B. (**2010**). "Phonemic restoration by hearing-impaired listeners with mild to moderate sensorineural hearing loss," Hear. Res. **260**, 54–62.

Başkent, D., and Gaudrain, E. (**2016**). "Musician advantage for speech-on-speech perception," J. Acoust. Soc. Am. **139**, EL51–EL56.

Başkent, D., Gaudrain, E., Tamati, T. N., and Wagner, A. (**2016b**). "Perception and psychoacoustics of speech in cochlear implant users," In A. T. Cacace, E. de Kleine, A. G. Holt, and P. van Dijk (Eds.), in *Scientific Foundations of Audiology: Perspectives From Physics, Biology, Modeling, and Medicine* (Plural Publishing, Inc, San Diego), pp. 285–319.

Başkent, D., Luckmann, A., Ceha, J., Gaudrain, E., and Tamati, T. N. (**2018**). "The discrimination of voice cues in simulations of bimodal electro-acoustic cochlear-implant hearing," J. Acoust. Soc. Am. **143**, EL292–EL297.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, v067i01.

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the False discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B Stat. Methodol. **57**, 289–300.

Bhargava, P., Gaudrain, E., and Başkent, D. (**2014**). "Top–down restoration of speech in cochlear-implant users," Hear. Res. **309**, 113–123.

Black, R. C., and Clark, G. M. (**1980**). "Differential electrical excitation of the auditory nerve," J. Acoust. Soc. Am. **67**, 868–874.

Bosman, A. J., and Smoorenburg, G. F. (**1995**). "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment," Int. J. Audiol. **34**, 260–284.

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

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Chiba, T., and Kajiyama, M. (**1941**). *The Vowel, Its Nature and Structure* (Tokyo-Kaiseikan Pub. Co., Tokyo).

Clarke, J., Başkent, D., and Gaudrain, E. (**2016**). "Pitch and spectral resolution: A systematic comparison of bottom-up cues for top-down repair of degraded speech," J. Acoust. Soc. Am. **139**, 395–405.

Cleary, M., and Pisoni, D. B. (**2002**). "Talker discrimination by prelingually deaf children with cochlear implants: Preliminary results," Ann. Otol. Rhinol. Laryngol. **111**, 113–118.

Cleary, M., Pisoni, D. B., and Kirk, K. I. (**2005**). "Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants," J. Speech Lang. Hear. Res. **48**, 204–223.

Cullington, H. E., and Zeng, F.-G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.

Cutler, A., Dahan, D., and Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang. Speech **40**(Pt 2), 141–201.

Dabbs, J. M., and Mallinger, A. (**1999**). "High testosterone levels predict low voice pitch among men," Pers. Individ. Differ. **27**, 801–804.

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**, 2913–2922.

El Boghdady, N., Başkent, D., and Gaudrain, E. (**2018**). "Effect of frequency mismatch and band partitioning on vocal tract length perception in vocoder simulations of cochlear implant processing," J. Acoust. Soc. Am. **143**, 3505–3519.

El Boghdady, N., Gaudrain, E., and Başkent, D. (**2019**). "Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?" J. Acoust. Soc. Am. **145**, 417–439.

El Boghdady, N., Langner, F., Gaudrain, E., Başkent, D., and Nogueira, W. (**2020**). "Effect of spectral contrast enhancement on speech-on-speech intelligibility and voice cue sensitivity in cochlear implant users," Ear Hear. **42**, 271–289.

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**, 160–163.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Fielden, C. A., Kluk, K., Boyle, P. J., and McKay, C. M. (**2015**). "The perception of complex pitch in cochlear implants: A comparison of monopolar and tripolar stimulation," J. Acoust. Soc. Am. **138**, 2524–2536.

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**, 1511–1522.

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

Fu, Q.-J., Chinchilla, S., Nogaki, G., and Galvin, J. J. 3rd (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**, 1711–1718.

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**, 1037–1048.

Gaudrain, E. (**2016**). "Vocoder: Basal," Zenodo.

Gaudrain, E., and Başkent, D. (**2015**). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**, 1298–1308.

Gaudrain, E., and Başkent, D. (**2018**). "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**, 226–237.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cognit. **19**, 448–458.

Green, D. M., and Swets, J. A. (**1988**). *Signal Detection Theory and Psychophysics*, reprint ed. (Peninsula, Los Altos, CA).

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Hillenbrand, J., Getty, L. A., Wheeler, K., and Clark, M. J. (**1994**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **95**, 2875–2875.

Jaekel, B. N., Weinstein, S., Newman, R. S., and Goupell, M. J. (**2021**). "Access to semantic cues does not lead to perceptual restoration of interrupted speech in cochlear-implant users," J. Acoust. Soc. Am. **149**, 1488–1497.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**, 1995–2004.

Kadam, M. A., Orena, A. J., Theodore, R. M., and Polka, L. (**2016**). "Reading ability influences native and non-native voice recognition, even for unimpaired readers," J. Acoust. Soc. Am. **139**, EL6–EL12.

Kawahara, H., and Irino, T. (**2005**). "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Boston), pp. 167–180.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
Lawrence, M. A. (**2016**). "ez: Easy analysis and visualization of factorial experiments [computer program]," https://CRAN.R-project.org/package=ez (Last viewed 2021-08-24).
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Massida, Z., Marx, M., Belin, P., James, C., Fraysse, B., Barone, P., and Deguine, O. (**2013**). "Gender categorization in cochlear implant users," J. Speech Lang. Hear. Res. **56**, 1389–1401.
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.
Meister, H., Walger, M., Lang-Roth, R., and Müller, V. (**2020**). "Voice fundamental frequency differences and speech recognition with noise and speech maskers in cochlear implant recipients," J. Acoust. Soc. Am. **147**, EL19–EL24.
Misurelli, S. M., and Litovsky, R. Y. (**2015**). "Spatial release from masking in children with bilateral cochlear implants and with normal hearing: Effect of target-interferer similarity," J. Acoust. Soc. Am. **138**, 319–331.
Nagels, L., Bastiaanse, R., Başkent, D., and Wagner, A. (**2020a**). "Individual differences in lexical access among cochlear implant users," J. Speech Lang. Hear. Res. **63**, 286–304.
Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., and Başkent, D. (**2020b**). "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**, 1361–1376.
Nogueira, W., Boghdady, N. E., Langner, F., Gaudrain, E., and Baskent, D. (**2020**). "Effect of channel interaction on vocal cue perception in cochlear implant users," Trends Hear. (in press); PsyArXiv 12 Mar. 2020.
Nygaard, L. C. (**2008**). "Perceptual integration of linguistic and nonlinguistic properties of speech," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishing, Malden, MA), pp. 390–413.
Perea, M., Jiménez, M., Suárez-Coalla, P., Fernández, N., Viña, C., and Cuetos, F. (**2014**). "Ability for voice recognition is a marker for dyslexia in children," Exp. Psychol. **61**, 480–487.
Perrachione, T. K., Del Tufo, S. N., and Gabrieli, J. D. E. (**2011**). "Human voice recognition depends on language ability," Science **333**, 595–595.
Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," J. Acoustical Soc. Am. **146**, 3384–3399.
Pisoni, D. B. (**1997**). "Some thoughts on 'normalization' in speech perception," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic Press, San Diego).
Ptacek, P. H., and Sander, E. K. (**1966**). "Age recognition from voice," J. Speech Hear. Res. **9**, 273–277.
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**, e12902.
R Core Team (**2020**). "R: A language and environment for statistical computing [computer program]," https://www.r-project.org (Last viewed 2021-08-24).
Shannon, C. E., and Weaver, W. (**1949**). *The Mathematical Theory of Communication* (University of Illinois Press, Champaign, IL).
Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave, and reversed speech samples," J. Exp. Psychol. Human Percept. Perform. **28**, 1447–1469.
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**, 3177–3186.
Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F.-G. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.
Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. U.S.A. **107**, 12387–12392.
Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters part I: Recognition of backward voices," J. Phon. **13**, 19–38.
van Leyden, K., and van Heuven, V. J. (**1996**). "Lexical stress and spoken word recognition: Dutch vs English," Linguistics Netherlands **13**, 159–170.
Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2011**). "The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening," J. Acoust. Soc. Am. **130**, 429–439.
Wagner, A. E., Toffanin, P., and Başkent, D. (**2016**). "The timing and effort of lexical access in natural and degraded speech," Front. Psychol. **7**, 398.
Winn, M. B. (**2016**). "Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants," Trends Hear. **20**, 1–17.
Winn, M. B., and Moore, A. N. (**2018**). "Pupillometry reveals that context benefit in speech perception can be disrupted by later-occurring sounds, especially in listeners with cochlear implants," Trends Hear. **22**, 1–22.
Zaltz, Y., Goldsworthy, R. L., Kishon-Rabin, L., and Eisenberg, L. S. (**2018**). "Voice discrimination by adults with cochlear implants: The benefits of early implantation for vocal-tract length perception," J. Assoc. Res. Otolaryngol. **19**, 193–209.
Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.