The Role of Word Content, Sentence Information, and Vocoding for Voice Cue Perception

Thomas Koelewijn, Etienne Gaudrain, Thawab Shehab, Tobias Treczoks, Deniz Başkent

The role of word content, sentence information, and vocoding for voice cue perception

Thomas Koelewijn[1,2], Etienne Gaudrain[1,2,3], Thawab Shehab[1,4], Tobias Treczoks[1,5], Deniz Başkent[1,2]

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

2 Research School of Behavioural and Cognitive Neurosciences, Graduate School of Medical Sciences, University of Groningen, Groningen, The Netherlands

3 Lyon Neuroscience Research Center, CNRS UMR5292, Inserm U1028, UCBL, UJM, Lyon, France

4 University of Groningen, faculty of Arts, Neurolinguistics, Groningen, Netherlands

5 Medical Physics and Cluster of Excellence "Hearing4all", Department of Medical Physics and Acoustics, Faculty VI Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Germany

This is the author version.

The published version is available online at:

https://doi.org/10.1044/2023_JSLHR-22-00491

Keywords: linguistic content; vocoder; voice cues perception

Corresponding Author:
Thomas Koelewijn
University Medical Center Groningen – Department of Otorhinolaryngology
Hanzeplein 1, 9713GZ Groningen, The Netherlands
Email: t.koelewijn@rug.nl
Phone: +31 50 3612540

Abstract

Purpose: For voice perception, two voice cues, the fundamental frequency (F0) and/or vocal-tract length (VTL), seem to largely contribute to identification of voices and speaker characteristics. Acoustic content related to these voice cues is altered in cochlear implant transmitted speech, rendering voice perception difficult for the implant user. In everyday listening, there could be some facilitation from top-down compensatory mechanisms such as from use of linguistic content. Recently we showed a lexical content benefit on just-noticeable-differences (JNDs) in VTL perception, which was not affected by vocoding. Whether this observed benefit relates to lexicality or phonemic content and if additional sentence information can affect voice cue perception as well, was investigated in this study.

Method: The current study examined lexical benefit on VTL perception, by comparing words, time-reversed words, and non-words, to investigate the contribution of lexical (words vs. non-words) or phonetic (non-words vs. reversed words) information. In addition, we investigated the effect of amount of speech (auditory) information on F0 and VTL voice cue perception, by comparing words to sentences. In both experiments non-vocoded and vocoded auditory stimuli were presented.

Results: The outcomes showed a replication of the detrimental effect reversed words have on VTL perception. Smaller JNDs were shown for stimuli containing lexical and/or phonemic information. Experiment 2 showed a benefit in processing full sentences compared to single words in both F0 and VTL perception. In both experiments there was an effect of vocoding, which only interacted with sentence information for F0.

Conclusions: In addition to previous findings suggesting a lexical benefit, the current results show more specifically, that lexical and phonemic information improves VTL perception. F0 and VTL perception benefits from more sentence information compared to words. These results indicate that cochlear implant users may be able to partially compensate for voice cue perception difficulties by relying on the linguistic content and rich acoustic cues of everyday speech.

1. Introduction

The ability to perceived differences between the voices of speakers can be of big help during speech perception. Two voice cues, fundamental frequency (F0) and vocal-tract length (VTL), seem to contribute the most to the perception of voice and speaker characteristics (Darwin et al., 2003; Skuk & Schweinberger, 2014; Vestergaard et al., 2011). Because of spectrotemporal degradations inherent to cochlear implant (CI) electric stimulation of the auditory nerve (for a review, please see Başkent, Gaudrain, et al., 2016), users of CIs have difficulties in perceiving the two voice cues, showing higher discrimination thresholds for F0 and VTL compared to NH listeners (Gaudrain & Başkent, 2018; Zaltz et al., 2018). When compared to simulation studies where the spectrotemporal resolution was degraded by means of vocoding, similar results were shown for some vocoding settings (Gaudrain & Başkent, 2015). These deficits in voice cue sensitivity can directly explain abnormal voice gender categorization amongst CI users (Fu et al., 2005; Fuller et al., 2014; Massida et al., 2013; Meister et al., 2016), as these two voice cues are also used for this task. Further, difficulties in other voice-related tasks have also been observed in CI users, such as perception of prosody (for a review, see Everhardt et al., 2020), vocal emotions (e.g., Jiam et al., 2017), indexical cues (e.g., Tamati & Moberly, 2022) and identification of talkers (for a review, see Colby & Orena, 2022). However, it is less clear how specifically the perception of average F0 and VTL may relate to these tasks.

Most studies on voice cue perception in CI listeners use simple speech materials, such as isolated syllables or words, for better control of the conditions and simplicity in study design. In everyday listening, however, speech communication provides richer speech cues, with longer speech segments and more linguistic content (e.g., lexical and sentential context). Spectrotemporal degradation of the speech signal in electric hearing can make distinguishing between differences in voice cues or linguistic cues ambiguous. To solve this ambiguity, CI users need to rely on top-down compensatory mechanisms for correct interpretation of the degraded speech cues (Amichetti et al., 2018; Başkent, Clarke, et al., 2016; Nagels et al., 2020; Winn & Moore, 2018). This was evidenced by studies showing that CI users can use linguistic context to make lexical decisions (Gianakas & Winn, 2019; Nagels et al., 2020). It is possible similar compensatory mechanisms could also be employed for voice perception.

In a recent study (Koelewijn et al., 2021), we aimed to investigate the effect of top-down compensatory mechanisms, related to word (lexical) content, on voice cue

discriminability. We assessed the effect of lexical content on just-noticeable-differences (JNDs) in F0 and VTL, and their combination (F0+VTL) for normal and vocoded speech, using an auditory adaptive odd-one-out task (three intervals, three alternatives forced choice; 3I-3AFC). In this study NH participants listened to meaningful Dutch consonant-vowel-consonant (CVC) words that were presented forward or time reversed. While time reversing prevented lexical access (Ptacek & Sander, 1966) talker specific voice cues, such as F0 and VTL and the acoustic entropy, were preserved. This way the amount of lexical/semantic information available was manipulated orthogonally from the acoustic voice cues. In addition, during each trial, either the same word was repeated three times (low-variability context) or three different words were presented (high-variability context), while, in both cases, one of the three items was uttered with a deviant voice. The results showed that within the context of high-variability, when presented with forward words, participants showed smaller VTL JNDs compared to time-reversed words. These outcomes suggest that lexical content may help to resolve ambiguity in whether a difference between items could be related to voice or to phonetic information. An ambiguity that primarily occurred in the variable condition when the phonetic content differed across intervals and when the acoustic features of the VTL voice cue and phonetic content overlapped. This result did not change when vocoding was applied. For F0 JNDs only in the non-vocoded condition with low variability, this lexical content benefit (forward words vs time-reversed words) was shown.

In a follow up study performed by Jebens et al., (2022), the effect of lexical content on F0 and VTL voice cue discriminability was further investigated using a voice gender categorization task. This time by making the district separation between lexicality (word meaning) and phonological (though phonetic alterations) effects. Lexicality was investigated by comparing word to nonword cue weights (outcomes) and phonological effects by comparing non-word to time reversed non-word cue weights of the voice gender categorization task. This showed lower (worse) cue weights for time reversed non-word compared to forward non-words and words, but no difference in cue weights between words and non-words, suggesting voice gender categorization to be affected by phonological rather than lexicality effects. These outcomes (Jebens et al., 2022) raised the question whether the observed effect in our previous study (Koelewijn et al., 2021) relates to, what we in this study will refer to as, “lexical information” (whether a word exists in the lexicon) or “phonemic information” (associated with phonological content including coarticulations within words) as part of the linguistic

content available in words. In addition, it made us wonder whether additional speech (auditory) information available in sentences (in addition to the linguistic content of the individual words) like sentential content (semantic), sentential context, and coarticulation between words, compared to single words (a factor we will call "sentence information") could affect voice cue perception.

A study by Meister et al. (2016) tested CI users and NH listeners for speaker voice gender categorization in single words and sentence stimuli. Findings from this study revealed that changing the F0 and VTL voice cues combined (F0+VTL) had a more substantial influence on the speaker's gender perception for NH listeners, than changing only the F0 voice cue. NH listeners relied on both F0 and VTL because these cues conflict when manipulated singularly. On the other hand, CI users showed ambiguous responses for speaker gender recognition, which can be explained by the limited spectral resolution of CI devices, affecting the ability to detect VTL differences (also see, Fuller et al., 2014). Besides, they found that performance in sentences was better than single words or four-word sequences in both groups, which might be explained by having more information in sentences that allows for more detailed analysis. However, CI listeners did not make use of VTL cues regardless of stimulus type. Whether a similar sentence benefit will be observed at the level of voice cue discrimination, and whether or not also VTL JNDs would improve by full sentence processing is unknown.

What we do know is that speech contains both linguistic and indexical information (Abercrombie, 1967), which seems to interact in a way that processing word or phoneme information can improve our sensitivity to voice cue differences (Koelewijn et al., 2021). This is in line with previous studies that already showed this link between linguistic and indexical information in speech perception (Nygaard, 2008; Pisoni, 1997). By time-reversing speech, most lexical content especially semantic information (word meaning) can no longer be retrieved, resulting in words that sound unfamiliar as they do not occur naturally in reversed form. According to Binder et al. (2000), in reversed speech some phonetic features that are temporally symmetrical (fricatives and vowels) may be preserved. However, plosive consonants because of their strong abrupt onset and a more gradual decay, are most likely not preserved when this signal is time reverse. In addition, coarticulations are reversed and therefore may sound unfamiliar. This means that the part of the lexical content benefit observed previously by Koelewijn et al., (2021), could in addition to the absence of word meaning be attributed to disrupted processing phonemic information (e.g., consonants) in reversed speech.

Hence, our previous study only using words and reversed words could not dissociate between, the negative effect of reversed speech on VTL voice cue discriminability, being related to lexical information, phonemic information, or both.

2. Experiment 1 – Lexical Content

The first experiment expanded on the characterization of potential lexical content benefit effect previously observed on VTL perception (Koelewijn et al., 2021), by comparing words, time-reversed words, and non-words, presented both in non-vocoded and vocoded versions. The lexical content benefit, referring to the participants' ability to detect smaller changes in the VTL voice cue for words compared to time-reversed word, could be related to lexical information and/or phonemic information. There is a difference between real words, non-words, and reversed words based on phonological and lexical components (Jebens et al., 2022). Words have both phonemic and lexical information that facilitates recognition. Non-words only obey the phonological constraints of a natural language but contain no lexical information. The time reversed words, as was used by (Koelewijn et al., 2021), do not follow the Dutch language's phonological constraints, and do not carry information to be represented in the mental lexicon. We hypothesized that spoken word recognition involves phonological and lexical components that facilitate lexical retrieval. Therefore, smallest JNDs were predicted with words followed by non-words and reversed words respectively. Any significant difference in JNDs between real words and non-words would signal the distinct representation of real words in the mental lexicon. In line with Koelewijn et al. (2021), we expect larger VTL JNDs for the vocoded conditions and no interaction with word type.

a. Methods

1. Participants

Of the 23 participants that initially signed up, 21 started the online experiment. A total of 14 participants performed all adaptive runs of the online experiment, of which three participants were excluded due to producing unusable data in some runs. Data from the remaining 11 participants (self-reported gender 6 females and 5 males; age range 22-41 years,

mean age 29 years) was analyzed. Additional demographics were collected to further describe the participant population. It is not clear at this point, based on literature, if any of these factors may or may not affect voice perception with or without vocoding, but for the purpose of completeness, we provide these demographics details. From the 11 participants, 2 reported MBO (secondary vocational education), 1 HBO (higher vocational education), and 8 university level education (according to the Dutch schooling system), all reported to have learned 2 or more languages in addition to their first language (3 learned 3 additional languages, 2 learned 4 additional languages), no participants reported to be raised bilingual, 5 participants played a music instrument of which 3 received formal music education before the age of 10, 1 participant reported to speak with a regional accent, and 4 participants lived 1 year or more outside the Netherlands. All participants reported normal hearing and normal or corrected-to-normal vision. Participants did not report dyslexia, epilepsy, and/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). Participation was voluntary, meaning that participants received no monetary compensation. Participants were recruited via word of mouth at the University of Groningen and the University Medical Center Groningen.

2. Stimuli

Three sets of audio stimulus items used contained: words, non-words, and time-reversed words. Each set consisted of seventy-five stimuli and the Dutch words and non-words were selected from the VariaNTS corpus (Arts et al., 2021). The VariaNTS corpus contains 11 linguistic categories based on lexical frequency, phonotactic probability, and neighborhood density. The selected words had a high frequency and low density (easy words) and the non-words had a low phonotactic probability and high density (hard non-words). Phonotactic probability refers to the frequency of occurrence of combinations of phones in language. Non-words with a low phonotactic probability do still follow these language rules but are less common. The terms 'easy' and 'hard' in the VariaNTS corpus refer to the processing demands of the linguistic information. Easy words and hard non-words were selected to maximize the difference in effect they might have on VTL JNDs. The VariaNTS corpus contains recordings of 8 female and 8 male native talkers of standard Dutch. The audio files selected for this experiment were from a 20-yr-old native Dutch female normal hearing speaker with no regional

accent, a height of 171 cm, a weight 59 kg, and with an average F0 of 214.36 Hz. The time-reversed word items were created in Adobe Audition (2020) by time reversing the audio files of the set word items.

Voice cue changes were processed online with WORLD (Morise et al., 2016). For more details on voice cue manipulation see Gaudrain and Başkent (2015, 2018). In the adaptive procedure described below, VTL of the stimuli was varied in each trial. Three randomly selected stimuli were resynthesized with WORLD using the new VTL parameter and word duration was normalized to 600 ms. The stimuli were resynthesized even when the VTL was unchanged compared to the original voice.

The auditory stimuli were presented without and with vocoding. Vocoding was coded in Python (as part of the VTServer v2.2; Gaudrain, E., 2021), in line with a previous implementation in MATLAB (Gaudrain & Başkent, 2015) (Gaudrain, 2016), and ran in real-time on a dedicated online sound processing server. In line with our previous study (Koelewijn et al., 2021), we used a vocoder simulating low spread of excitation (LS-vocoder), and a vocoder simulating high spread of excitation (HS-vocoder) (Koelewijn et al., 2021). Both vocoding conditions used 12 analysis filters. These were 12^{th} order (72 dB/oct.) zero-phase Butterworth filters with a range spanning from 150 to 7000 Hz, which were uniformly divided in terms of cochlear place of excitation (Greenwood, 1990). For the LS-vocoding condition the synthesis filters were identical to the analysis filters, while for the HS-vocoding condition 4^{th} order filters (24 dB/oct.) were used. The temporal envelope was extracted, in each frequency band, by half-wave rectification and low-pass filtering. These low-pass filters (zero-phase 4^{th} order Butterworth) had an effective cut-off frequency of half the bandwidth of each band, with a maximum of 300 Hz (Gaudrain & Başkent, 2015). For both vocoding conditions noise was used as a carrier signal.

3. Procedure

Participants performed a three alternative forced choice (3AFC) task. During each trail three consecutive stimuli items were presented, and participants chose the one that sounded different relative to the other two. Although they could use any cue available to choose the odd-one-out, depending on the specific experiment and/or condition, the 'voice cue'

manipulated was VTL only. Note that the content of each of the three items was different, which entail variability in acoustic content, phonological content, and lexical content for words.

For each condition, JNDs were estimated separately using a 2-down-1-up staircase procedure (Levitt, 1971), resulting in approximately 71% correct response for each test (adaptive run). For each adaptive run the VTL of the deviant item started at a 12 st absolute distance from the female reference voice of the other two items. Each adaptive run started with a 2 st step size, which reduced by a factor of $\sqrt{2}$ every 15 trials or when the voice difference became smaller than twice the step-size. The test ended after 8 reversals, after 150 trials, or when it reached a difference of 25 st, and the JND was calculated as the mean over the last 6 reversals and the difference that would have been presented in the following trial. If 8 reversals were not reached in 150 trials or the staircase reached a difference of 25 st the procedure was aborted, and the data was unusable.

In the 3AFC task participants had to select the odd-one-out based on any perceivable voice difference. During each trial participants listened to the three items play in sequence while at the same time the three corresponding buttons, presented on screen from left to right, briefly lighted up. Participants responded by a mouse click on the button corresponding to what they perceived as the deviant item. At the end of each trial, they received visual feedback from the selected button by blinking in green or red for correct or incorrect responses respectively. Each condition was presented in a block wise fashion of which the order was randomized separately for each participant.

4. Online testing and apparatus

For running the experiment online, a platform was used developed by the dB SPL research group within the University Medical Center Groningen. The platform was coded in JavaScript using the jPsych library (v6.2; de Leeuw, 2015) and was accessible through a web browser. Participants were advised to run the experiment on a desktop or laptop computer, to use a good set of headphones, and to perform the experiment in a quiet environment. The experiment started with general project information about the hearing research conducted in our lab, followed by an informed consent form, to which participants had to agree before they could continue. Next, participants had to fill out an online questionnaire containing questions on demographics relevant for voice and speech perception research. Subsequently, participants

received experiment-specific instructions, adjusted the sound levels presented through their headphones to a comfortable level, followed by a short practice session before the start of the actual experiment.

5. Statistical analysis

All JNDs were log-transformed to improve homogeneity of variance across conditions. We performed a 3x3 repeated-measure analysis of variance (ANOVA) on the log-transformed VTL JNDs with lexical content (forward, reversed, non-word) and vocoding (no, LS, HS) as the within subject factors. Effect sizes are reported as generalized eta-squared (Bakeman, 2005). For planned comparisons on the data of Experiment 1, paired samples t-tests were used. The Holm-Bonferroni correction for multiple comparisons (Holm, 1979) was used, and adjusted p-values are reported. Effect sizes for these t-tests are reported as Cohen's d. The statistics were computed in R v4.3.0 (R Core Team, 2020) with the ez package v4.4.0 (Lawrence, 2016) using type III sums of squares.

b. Results

The outcome showed a significant main effect of vocoder [$F_{(2,20)} = 68.0$, $p < .001$, $\eta_g^2 = .65$], and a significant effect of word status [$F_{(2,20)} = 22.0$, $p < .001$, $\eta_g^2 = .21$] on VTL JNDs (see Figure 1). However, there was no significant interaction [$F_{(4,40)} = .80$, $p = .479$, $\eta_g^2 = .02$]. For planned comparisons between word status conditions, the thresholds were averaged over all vocoder conditions. Three paired samples t-tests were used for comparing JNDs between the word status conditions. A significant difference was found between words and reversed words [$t_{(10)} = -5.83$, $p_{adj.} < .001$, d = 1.76], showing larger (worse) JNDs for reversed words compared to words. Also, non-words and reversed words were significantly different from each other [$t_{(10)} = 5.15$, $p_{adj.} < .001$, d = 1.55], showing larger JNDs for reversed words compared to non-words. The words and non-words were also significantly different [$t_{(10)} = -2.43$, $p_{adj.} = .035$, $d = .73$], but note that the effect size was only moderate. These results showed that both lexical and more phonological information was associated with smaller (better) JNDs. Additionally, for planned comparisons between vocoder conditions, the thresholds were averaged over all word status conditions. Three paired samples t-tests were used for comparing JNDs between the vocoder conditions. The results were significantly different between the No

vocoder and LS Vocoder [$t_{(10)}$ = -8.45, $p_{adj.}$ < .001, d = 2.55] and between the No vocoder and the HS Vocoder [$t_{(10)}$ = -9.59, $p_{adj.}$ < .001, d = 2.89]. Finally, the two vocoders were also found to be significantly different [$t_{(10)}$ = -2.47, $p_{adj.}$ < .05, d = .74], but note that the effect size was only moderate.

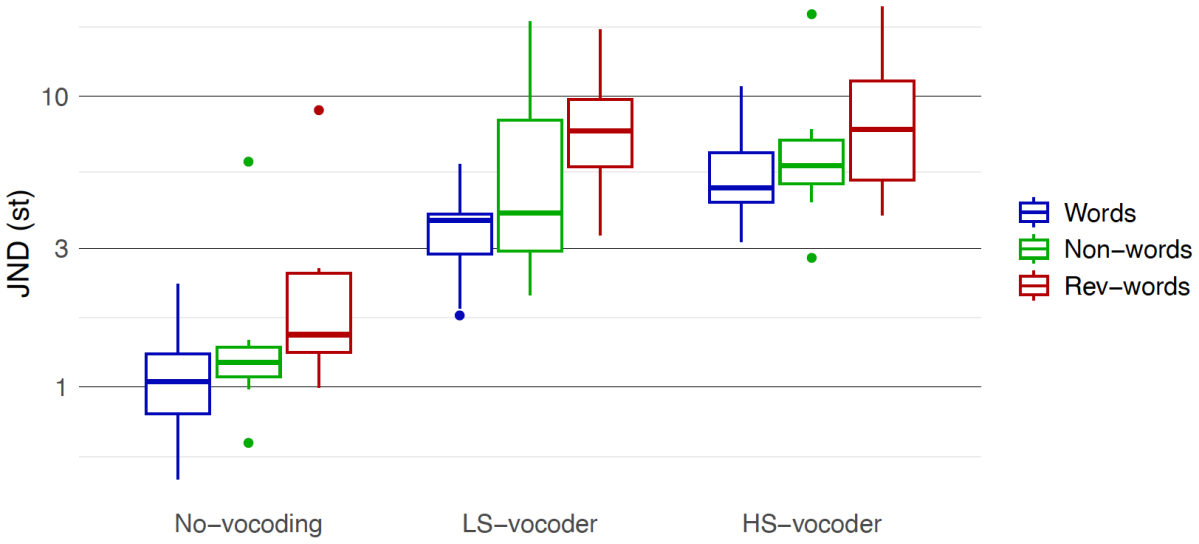

Figure 1 — Median JNDs on the y-axis (log-spaced) for each lexical content (forward, reversed, non-word), vocoding (no, low spread = LS, high spread = HS), and voice cue (VTL) condition. Boxes extend from the lower to the upper quartile (the interquartile range, IQ), and the midline indicates the median. 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.

c. Discussion

The results of Experiment 1 show that in line with our hypotheses the lexical content benefit as observed previously, when presenting variable items (Koelewijn et al., 2021), is related to lexical information available in words, and phonological content available in both words and non-words. In our previous study (Koelewijn et al., 2021), presenting variable items across the three alternatives resulted in larger VTL JNDs compared to presenting the same item (fixed) three times. This is because, for the fixed items conditions, variations in formant frequencies between items were only related to shifts in VTL, which made these differences stand out. In contrast, in the variable condition different CVC words were presented, each

potentially containing different vowels that would additionally contribute to formant variation. Only the variable item conditions showed a benefit of lexical content, suggesting that word content, only available in forward speech, helped to disambiguate between differences in formants related to either shifts in VTL or variable word items. The current results now specifically showed a lexical and phonological benefit in VTL discriminability. The later explained by change in transients, resulting in for instance the absence of plosives in consonants, and affected coarticulations in time- reversed speech. Note that vowels and fricatives are preserved in time- reversed speech (Binder et al., 2000). This was revealed by comparing the three conditions: words that contain both lexical and phonological content, non-words that have no lexical but full phonological content, and time-reversed words that have no lexical and impoverished phonological content. In other words, it seems that the absence lexical information, and of plosives and coarticulations, affecting phoneme processing, resulted in a significant decrease in VTL discriminability. Note that the current study did not investigate the effect of lexical content on F0 JNDs but focused on VTL sensitivity instead. This, since in our previous study (Koelewijn et al., 2021) only VTL JNDs were affected by lexical content when vocoded, which may indicate that top-down compensatory mechanisms can improve voice perception.

In line with the outcomes from our previous study (Koelewijn et al., 2021), the observed semantic and phonological benefit on VTL perception was shown for both vocoder conditions. This suggests the possibility that CI users could rely on phonological content to compensate for the difficulties in perceiving VTL, shown by Gaudrain and Başkent (2018), and using it for voice gender recognition, shown by Fuller et al. (2014). On the one hand, it might be that some of these previous studies only using syllables, instead of meaningful words or sentences (e.g., Zaltz et al., 2018), underestimated VTL discrimination in CI users. On the other hand, it is known that access to phonological content in CI users is often hampered when it comes to perceiving steady state vowels (Fielden et al., 2015). Vowel perception in general is difficult for CI users since they cannot easily distinguish the voiced cues of harmonic structure or hear spectral peaks, and they confuse place cues (Fu et al., 1998). Dynamic cues are easier to hear by CI users because cues move and the transition of a peak from one electrode to another would make it more audible, but steady vowels do not provide this. However, contradicting evidence is also shown (Donaldson et al., 2015). Still, as discussed earlier, it is most likely the plosives in consonants and coarticulations that are affected by the reversal of the speech signal. Hence,

the current results indicate a possibility for better VTL perception in CI users when it comes to real life listening situations than what is measured in lab experiments with short speech segments or syllables.

3. Experiment 2 – Words vs. Sentences

Experiment 1 shows that lexical and phonemic content in words had an influence on VTL JNDs. In Experiment 2 we look at the effect of adding additional sentence information (content, context, and coarticulation) by comparing JNDs with single words to with full sentences. Non-vocoded as well as vocoded speech was presented, as full sentence processing was previously observed to have an effect on voice gender categorization in CI listeners (Meister et al., 2016). Since HS-vocoding conditions in Experiment 1 were shown to be very difficult, even resulting in unusable data for some of the participants, in Experiment 2 only LS-vocoding was implemented. Both F0 and VTL voice cues were independently manipulated since CI users, in that same study, as well as others (e.g., Fuller et al., 2014), seemed only to make use of F0 voice cue differences. The pitch contours (e.g., F0) are more complex in sentences than single words and therefore might contribute detectability of voice differences. Experiment 2 addressed the question if additional speech (acoustic, lexical, and/or semantic) information conveyed through sentences compared to words influence the participant's ability of discriminating voice cues (F0 and VTL). We predicted smaller JNDs in both F0 and VTL voice cues when sentences are presented compared to words. Again, in line with Koelewijn et al. (2021) we expect larger JNDs for the vocoded conditions.

a. Methods

1. Participants

Of the 13 participants that signed up and started the online experiment, a total of 12 participants performed all adaptive runs, of which one participant was excluded due to producing unusable data in some runs. For the remaining 11 participants (self-reported gender 8 females and 3 males; age range 26-52 years, mean age 32 years) the data was analyzed. Additional demographics collected are the following. From the 11 participants 1 reported

HAVO (senior general secondary), 1 HBO (higher vocational education), and 9 university level education, all reported to have learned 2 or more languages in addition to their first language (7 learned 3 additional languages, 1 learned 4 additional languages), 1 participant reported to be raised bilingual (Dutch and English), 3 participants played an instrument of which 1 received formal music education before the age of 10, 1 participant reported to speak with a regional accent, and 7 participants lived 1 year or more outside the Netherlands. All participants reported normal hearing and normal or corrected-to-normal vision. Participants did not report dyslexia, epilepsy, and/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). Participation was voluntary, meaning that participants received no monetary compensation. Participants were recruited via word of mouth at the University of Groningen and the University Medical Center Groningen.

2. Stimuli

The stimuli consisted of audio words and sentences. For sentence stimuli the original recordings of Dutch everyday sentences from the VU98 corpus, by a native Dutch female talker (referred to as talker HB) were used (for a full description see, Versfeld et al., 2000). For word stimuli 28 Dutch meaningful CVC words, part of the Nederlandse Vereniging voor Audiologie (NVA) corpus (Bosman & Smoorenburg, 1995), were extracted from the original recordings of the VU98 corpus. In that way resulting word and sentence stimuli were uttered by the same female talker (HB) with an average F0 of 175 Hz.

Voice cue changes were processed online with WORLD (Morise et al., 2016) the same way as in Experiment 1. Three randomly selected stimuli were resynthesized with WORLD using the new F0 and/or VTL parameters. The stimuli were resynthesized even when the F0 and/or VTL were unchanged compared to the original voice. The auditory stimuli were presented without and with noise vocoding. Experiment 2 only used the LS-vocoding condition, and all settings were the same as in Experiment 1.

3. Procedure

Experiment 2 was again run online the same as Experiment 1. Participants performed a similar three alternative forced choice (3AFC) task as described previously. Each adaptive run

started with a 12 st difference and a 2 st step size for each voice cue, which reduced by a factor of $\sqrt{2}$ every 15 trials or when the voice difference became smaller than twice the step-size. Again, the content of each of the three items was different, which entail variability in sentence content and context. For each condition, JNDs were estimated separately similarly as in Experiment 1.

4. Statistical analysis

All JNDs were log-transformed to improve homogeneity of variance across conditions. In line with our previous research (Koelewijn et al., 2021) data for each voice cue was separately analyzed, since intersubject variance tends to be different for F0 and VTL. Hence, for each of the voice cues we performed a separate 2×2 ANOVA on the log-transformed JNDs with sentence information (words, sentences), and vocoding (no, LS) as the within subject factors. For post-hoc analysis, paired samples t-tests were used. The Holm-Bonferroni correction for multiple comparisons (Holm, 1979) was applied. ANOVAs were performed in R v4.3.0 (R Core Team, 2020) with the ez package v4.4.0 (Lawrence, 2016) using type III sums of squares. Effect sizes are reported as generalized eta-squared (Bakeman, 2005).

b. Results

The outcomes (see Figure 2) for F0 JNDs showed a significant main effect of sentence information [$F_{(1,10)} = 8.70$, $p = .015$, $\eta_g^2 = .23$] and vocoder [$F_{(1,10)} = 175$, $p < .001$, $\eta_g^2 = .68$]. There was a significant interaction between vocoder and sentence information [$F_{(1,10)} = 7.13$, $p = .023$, $\eta_g^2 = .06$], but note that the effect is small. Post hoc analysis in the form of two separate pairwise t-tests showed a significant difference between words and sentences in the non-vocoded conditions [$t_{(10)} = -3.34$, $p_{adj.} = .015$, $d = 1.01$] but not for the vocoded conditions [$t_{(10)} = -1.72$, $p_{adj.} = .117$, $d = .52$].

The outcomes for VTL JNDs showed a significant main effect of sentence information [$F_{(1,10)} = 17.9$, $p < .01$, $\eta_g^2 = .30$] and vocoder [$F_{(1,10)} = 74.8$, $p < .001$, $\eta_g^2 = .41$]. However, there was no significant interaction between vocoder and sentence information [$F_{(1,10)} = .04$, $p = .85$, $\eta_g^2 < .01$]. Planned comparisons in the form of two separate pairwise t-tests showed a

significant difference between words and sentences both in the non-vocoded conditions [$t_{(10)}$ = –3.90, $p_{adj.}$ < .01, d = 1.17] and for the vocoded conditions [$t_{(10)}$ = -2.41, p < .05, d = .73].

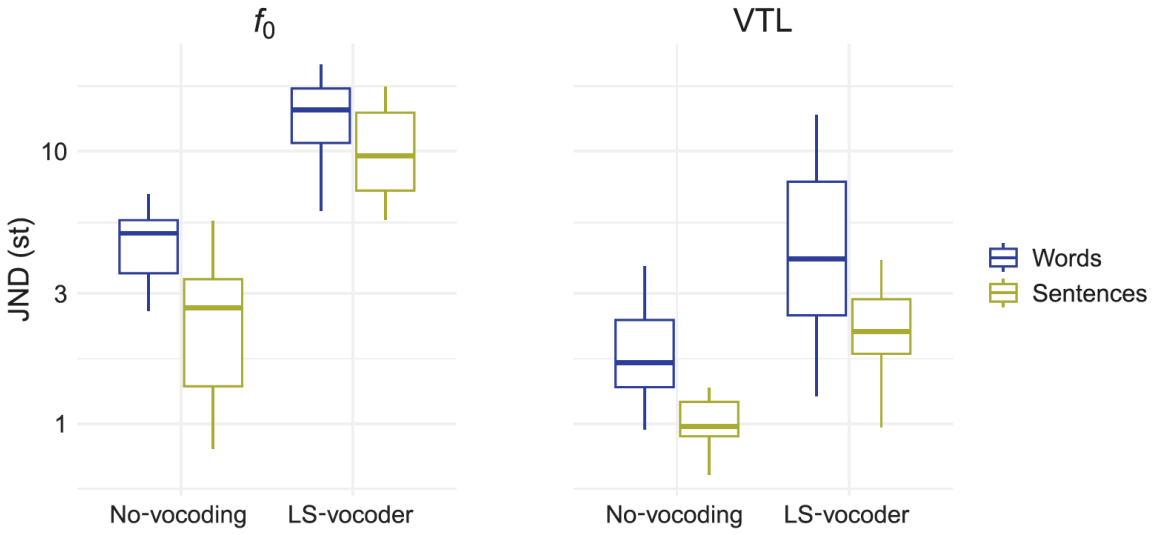

Figure 2 — Median JNDs on the y-axis (log-spaced) for each lexical content (words, sentences), vocoding (no, low spread = LS), and voice cue (F0, VTL) condition. Boxes extend from the lower to the upper quartile (the interquartile range, IQ), and the midline indicates the median. The whiskers indicate 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.

c. Discussion

The results of Experiment 2 suggest that more speech information by presenting sentences compared to words improved the discriminability of both F0 and VTL voice cues, with the exception of the F0 vocoded condition. More information available in sentences, in one form or another, means more variability in the signal. This might be reflected by more F0 fluctuations, and that probably means more formant fluctuations too, which would make the detection of mean differences between F0 or VTL voice cues more difficult. But this account assumes that the cues are simply accumulated over time without any effect of sentence information. Linguistic content would also provide interpretability or even predictability of the content. Meister et al. (2016) investigated the effect of both stimulus length and sentence information (semantic content) on voice gender categorization in NH and CI users, while systematically manipulating F0 and VTL cues. To independently manipulate the effect of

stimulus duration and amount of sentential context, a single word (0.4-0.5 s), a sequence of four unrelated words (quadruple, approximately 2 s), and simple sentences (mean length 1.7 s) were presented. The quadruple provided acoustic variation comparable to a sentence but without the predictability of pitch contour and intonation, without coarticulatory cues of naturally produced sentences, and without the semantic information and sentential context normally available in a sentence. Their results showed larger differences between the results for F0-VTL and F0 voice cue for sentences compared to quadruples, indicating effects of sentence information (content, context, and coarticulation), and again compared to single words, indicating effects of stimuli duration (number of words), on voice gender categorization. Hence, the results of the current study likely show the same combined effect on JNDs in F0 and VTL cues. Apart from a possible effect of stimulus length, presenting sentences within this task increases the amount of linguistic information available to the listener at phoneme, word, and sentence level. Findings from the literature suggest that this additional linguistic input can be used in a top-down manner to improve speech perception (e.g., Bashford & Warren, 1979; Başkent, 2012; Meister et al., 2009, 2016, p. 201; Verschuure & Brocaar, 1983; Warren, 1970), also for cognitive compensation of degraded speech (Başkent, Clarke, et al., 2016) as shown for VTL JNDs.

For CI users, this sentence benefit might be specifically beneficial, as they have a higher reliance on context information (Dingemanse & Goedegebure, 2019). Though other studies showed that how much these listeners profited varied greatly across different CI users, while often the processing effort increased (Wagner et al., 2019). Still, there are several aspects to which sentences are different from words. Sentences vary with respect to several linguistic aspects compared to words. Not only are they longer and contain more acoustic cues, but they also contain semantic information in the form of context, have higher lexical content, and show more phonetic aspects, such as intonation and coarticulations. Solely comparing sentences with words therefore leaves the question, which of the linguistic aspects add to which part of the measured benefit. Although these outcomes do not give us a definitive answer to this question, it shows the listener's ability to use this additional information to their advantage.

4. General Discussion

In the current study, we aimed to investigate the effect of top-down compensatory mechanisms related to lexical word content (lexical and phonemic) and sentence information (content, context, and coarticulations) on voice cue discriminability. In addition to previous findings (Koelewijn et al., 2021), the results of Experiment 1 show more specifically that semantic content available in words and phonological content available in both words and non-words improves VTL perception relative to time-reversed words. These results partly in line with Jebens et al. (2022), who did observe an effect of phonological content but not of lexical content on use of voice cues in voice gender categorization. The outcomes of Experiment 2 were as predicted by showing smaller F0 and VTL JNDs when sentences were presented compared to words, except for the F0 vocoded condition where no difference between JNDs was shown. This indicates that voice cue perception benefits from more speech (auditory) information, like sentential content (semantic), sentential context, and coarticulation between words, available in sentences compared to single words. This in addition to the linguistic content of the individual words that make up the sentence (Meister et al., 2016). Importantly, these results were shown with and without vocoding, which has implications for CI listening as discussed below.

a. Compensation for degradations caused by vocoding

As hypothesized, both experiments showed larger JNDs for the vocoded conditions compared to the non-vocoded conditions. In Experiment 1, both a low and high spread vocoder was used to mimic observed differences in spread of excitation by CI electric stimulation of the auditory nerve (see for a detailed explanation, Başkent, Gaudrain, et al., 2016). The results show a main effect of vocoding and no interaction with word status, which is in line with previous outcomes (Koelewijn et al., 2021). The benefit of lexical and phonemic information on VTL JNDs was shown even for the most degraded HS-vocoder condition. In Experiment 2, where only a low spread vocoder was implemented, a sentence information benefit was shown for the vocoded VTL voice cue condition but not for the vocoded F0 voice cue condition. Both experiments show a benefit on VTL voice cue perception were shown for the vocoder conditions. In line with our previous study (Koelewijn et al., 2021), this suggests that CI users

could resort to top-down mechanisms relying on linguistic content like phonemes, words, or full sentence context as a compensatory strategy (Başkent, Clarke, et al., 2016).

b. Overall implications for cochlear-implant users

The current results have some potential general implications for CI users. Until now, most research using syllable sequences (e.g., CV-triplets) or single words showed that CI listeners (El Boghdady et al., 2019, 2021; Gaudrain & Başkent, 2018; Nogueira et al., 2021), or NH listeners tested with vocoders (El Boghdady et al., 2018; Gaudrain & Başkent, 2015), have difficulty hearing VTL voice cues. Further, CI users or vocoder tested NH participants make voice gender categorization that differ from NH listeners, in how they are using VTL voice cues. In vocoder simulations, more random answers are given than a systematic misuse of F0 or VTL voice cues, while CI users show a systematic overreliance on F0 while making no use of VTL. The results from Experiment 2 show that when presenting sentences, either due to being a longer signal and presenting more acoustic cues related to voice both CI/vocoded and normal hearing listeners can make use of, or having rich linguistic and semantic content, this may help to better perceive VTL voice cues. In line with this, Meister et al., (2020) conclude that CI listers can use timbre cues (related to VTL) when presented with sentences instead of CVC or single word stimuli. A result also shown by Zaltz et al., (2018), although affected by age of implantation, using short three word sentences. This may mean that in real life CI listeners may be able to hear and make use of VTL voice cues, when listening to a conversation. Still, we need to keep in mind that these results are observed with vocoded speech presented to NH listeners. Although a vocoder provides some degradation implemented in a way that resembles some of the CI signal processing, and when vocoder parameters are carefully selected can provide a good overlap in results from simulations and CI users (Gaudrain & Başkent, 2015, 2018), there is still no guarantee that this degradation is a good representation of real CI listening for this specific study. The current results based on vocoded speech only implies that with sentences, NH listeners can compensate for some degradation of voice cues compared to syllables. In CI users, there are other elements that may play a role, such as damage to the auditory nerve, where the electrodes are located, and limitations of electric hearing, but also long-term adaptation to the CI transmitted sounds. Importantly, the study by Meister et al. (2016) does show an effect of using sentences compared to words in voice gender

categorization with CI users, which support the idea that the current results might be replicable by a future study with actual CI users.

c. Conclusions

The current study extends our previous findings (Koelewijn et al., 2021) by similarly showing an interaction between voice perception and the acoustic and linguistic content of the stimuli. But further than the previous study, the results show that the interaction between lexical content and voice perception is specifically related to phonemes available in words and non-words compared to time-reversed words. Word meaning does not show a significant benefit on VTL perception. In addition, results suggest that linguistic relationships provided by sentences improve voice discrimination. Interestingly, both the phonological and sentence content advantage were resilient to signal degradation by means of vocoding. These outcomes suggest that top-down mechanisms depending on lexical and phonemic information available in words and sentence information, likely a combination of semantic content, context and acoustic cues like coarticulations between words (Meister et al., 2016), could potentially be utilized as a compensatory strategy by CI listeners (Başkent, Clarke, et al., 2016).

Acknowledgement

The authors thank Jennifer Breetveld and Tessa Peijzel for their help with facilitating this project. The study was conducted as part of the internships of third and fourth authors at the UMCG, within the programs of the European Master's in Clinical Linguistics (EMCL+) and the EU programme for education, training, youth and sport (ERASMUS+), respectively. 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 the last author, 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.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available in the DataverseNL repository, [https://doi.org/10.34894/JPGDMN].

References

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

Amichetti, N. M., Atagi, E., Kong, Y.-Y., & Wingfield, A. (2018). Linguistic Context Versus Semantic Competition in Word Recognition by Younger and Older Adults With Cochlear Implants: *Ear and Hearing*, *39*(1), 101–109. https://doi.org/10.1097/AUD.0000000000000469

Arts, F., Başkent, D., & Tamati, T. N. (2021). Development and structure of the VariaNTS corpus: A spoken Dutch corpus containing talker and linguistic variability. *Speech Communication*, *127*, 64–72. https://doi.org/10.1016/j.specom.2020.12.006

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384. https://doi.org/10.3758/BF03192707

Bashford, J. A., & Warren, R. M. (1979). Perceptual synthesis of deleted phonemes. *The Journal of the Acoustical Society of America*, *65*(S1), S112.

Başkent, D. (2012). Effect of Speech Degradation on Top-Down Repair: Phonemic Restoration with Simulations of Cochlear Implants and Combined Electric–Acoustic Stimulation. *JARO: Journal of the Association for Research in Otolaryngology*, *13*(5), 683–692. https://doi.org/10.1007/s10162-012-0334-3

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., & Gaudrain, E. (2016). Cognitive Compensation of Speech Perception With Hearing Impairment, Cochlear Implants, and Aging: How and to What Degree Can It Be Achieved? *Trends in Hearing*, *20*, 233121651667027. https://doi.org/10.1177/2331216516670279

Başkent, D., Gaudrain, E., Tamati, T. N., & Wagner, A. (2016). Perception and Psychoacoustics of Speech in Cochlear Implant Users. In A. T. Cacace, E. de Kleine, A. G. Holt, & P. van Dijk (Eds.), *Scientific foundations of audiology: Perspectives from physics, biology, modeling, and medicine*. Plural Publishing, Inc.

Binder, J. R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., & Possing, E.T. (2000). Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Bosman, A. J., & Smoorenburg, G. F. (1995). Intelligibility of Dutch CVC Syllables and Sentences for Listeners with Normal Hearing and with Three Types of Hearing Impairment. *International Journal of Audiology*, *34*(5), 260–284. https://doi.org/10.3109/00206099509071918

Colby, S., & Orena, A. J. (2022). Recognizing Voices Through a Cochlear Implant: A Systematic Review of Voice Perception, Talker Discrimination, and Talker

Identification. *Journal of Speech, Language, and Hearing Research*, *65*(8), 3165–3194. https://doi.org/10.1044/2022_JSLHR-21-00209

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913. https://doi.org/10.1121/1.1616924

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dingemanse, J. G., & Goedegebure, A. (2019). The Important Role of Contextual Information in Speech Perception in Cochlear Implant Users and Its Consequences in Speech Tests. *Trends in Hearing*, *23*, 233121651983867. https://doi.org/10.1177/2331216519838672

Donaldson, G. S., Rogers, C. L., Johnson, L. B., & Oh, S. H. (2015). Vowel identification by cochlear implant users: Contributions of duration cues and dynamic spectral cues. *The Journal of the Acoustical Society of America*, *138*(1), 65–73. https://doi.org/10.1121/1.4922173

El Boghdady, N., Başkent, D., & Gaudrain, E. (2018). Effect of frequency mismatch and band partitioning on vocal tract length perception in vocoder simulations of cochlear implant processing. *The Journal of the Acoustical Society of America*, *143*(6), 3505–3519. https://doi.org/10.1121/1.5041261

El Boghdady, N., Gaudrain, E., & Başkent, D. (2019). Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users? *The Journal of the Acoustical Society of America*, *145*(1), 417–439. https://doi.org/10.1121/1.5087693

El Boghdady, N., Langner, F., Gaudrain, E., Başkent, D., & Nogueira, W. (2021). Effect of Spectral Contrast Enhancement on Speech-on-Speech Intelligibility and Voice Cue Sensitivity in Cochlear Implant Users. *Ear & Hearing*, *Publish Ahead of Print*. https://doi.org/10.1097/AUD.0000000000000936

Everhardt, M. K., Sarampalis, A., Coler, M., Başkent, D., & Lowie, W. (2020). Meta-Analysis on the Identification of Linguistic and Emotional Prosody in Cochlear Implant Users and Vocoder Simulations. *Ear & Hearing*, *41*(5), 1092–1102. https://doi.org/10.1097/AUD.0000000000000863

Fielden, C. A., Kluk, K., & Boyle, P. J. (2015). The perception of complex pitch in cochlear implants: A comparison of monopolar and tripolar stimulation. *J. Acoust. Soc. Am.*, 14.

Fu, Q.-J., Chinchilla, S., Nogaki, G., & Galvin, J. J. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *The Journal of the Acoustical Society of America*, *118*(3), 1711–1718. https://doi.org/10.1121/1.1985024

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *104*(6), 3586–3596. https://doi.org/10.1121/1.423941

Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., & Başkent, D. (2014). Gender Categorization Is Abnormal in Cochlear Implant Users. *Journal of the Association for Research in Otolaryngology*, *15*(6), 1037–1048. https://doi.org/10.1007/s10162-014-0483-7

Gaudrain, E., & Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *137*(3), 1298–1308. https://doi.org/10.1121/1.4908235

Gaudrain, E., & Başkent, D. (2018). Discrimination of Voice Pitch and Vocal-Tract Length in Cochlear Implant Users: *Ear and Hearing*, *39*(2), 226–237. https://doi.org/10.1097/AUD.0000000000000480

Gianakas, S. P., & Winn, M. B. (2019). Lexical bias in word recognition by cochlear implant listeners. *The Journal of the Acoustical Society of America*, *146*(5), 3373–3383. https://doi.org/10.1121/1.5132938

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605. https://doi.org/10.1121/1.399052

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Jebens, A., Başkent, D., & Rachman, L. (2022). Phonological effects on the perceptual weighting of voice cues for voice gender categorization. *JASA Express Letters*, *2*(12), 125202. https://doi.org/10.1121/10.0016601

Jiam, N. T., Caldwell, M., Deroche, M. L., Chatterjee, M., & Limb, C. J. (2017). Voice emotion perception and production in cochlear implant users. *Hearing Research*, *352*, 30–39. https://doi.org/10.1016/j.heares.2017.01.006

Koelewijn, T., Gaudrain, E., Tamati, T., & Baskent, D. (2021). The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception. *J. Acoust. Soc. Am.*, 15.

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments* (4.4-0). https://CRAN.R-project.org/package=ez

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Massida, Z., Marx, M., Belin, P., James, C., Fraysse, B., Barone, P., & Deguine, O. (2013). Gender Categorization in Cochlear Implant Users. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1389–1401. https://doi.org/10.1044/1092-4388(2013/12-0132)

Meister, H., Fuersen, K., Streicher, B., Lang-Roth, R., & Walger, M. (2020). Letter to the Editor Concerning Skuk et al., "Parameter-Specific Morphing Reveals Contributions of Timbre and Fundamental Frequency Cues to the Perception of Voice Gender and Age in Cochlear Implant Users." *Journal of Speech, Language, and Hearing Research*, *63*(12), 4325–4326. https://doi.org/10.1044/2020_JSLHR-20-00563

Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., & Walger, M. (2016). The Use of Voice Cues for Speaker Gender Recognition in Cochlear Implant Recipients. *Journal of Speech, Language, and Hearing Research*, *59*(3), 546–556. https://doi.org/10.1044/2015_JSLHR-H-15-0128

Meister, H., Landwehr, M., Pyschny, V., Walger, M., & von Wedel, H. (2009). The perception of prosody and speaker gender in normal-hearing listeners and cochlear implant

recipients. *International Journal of Audiology*, *48*(1), 38–48. https://doi.org/10.1080/14992020802293539

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, *E99.D*(7), 1877–1884. https://doi.org/10.1587/transinf.2015EDP7457

Nagels, L., Bastiaanse, R., Başkent, D., & Wagner, A. (2020). Individual Differences in Lexical Access Among Cochlear Implant Users. *Journal of Speech, Language, and Hearing Research*, *63*(1), 286–304. https://doi.org/10.1044/2019_JSLHR-19-00192

Nogueira, W., Boghdady, N. E., Langner, F., Gaudrain, E., & Başkent, D. (2021). Effect of Channel Interaction on Vocal Cue Perception in Cochlear Implant Users. *Trends in Hearing*, *25*, 233121652110301. https://doi.org/10.1177/23312165211030166

Nygaard, L.C. (2008). Perceptual integration of linguistic and nonlinguistic properties of speech. In Pisoni, D.B. & Remez, R.E. (Eds.), *The Handbook of Speech Perception* (pp. 390–413). Blackwell Publishing.

Pisoni, D.B. (1997). Some thoughts on "normalization" in speech perception. In Johnson, K. & Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing*. Academic Press.

Ptacek, P. H., & Sander, E. K. (1966). Age Recognition from Voice. *Journal of Speech and Hearing Research*, *9*(2), 273–277. https://doi.org/10.1044/jshr.0902.273

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of Fundamental Frequency, Formant Frequencies, Aperiodicity, and Spectrum Level on the Perception of Voice Gender. *Journal of Speech, Language, and Hearing Research*, *57*(1), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Tamati, T. N., & Moberly, A. C. (2022). Processing of linguistic and indexical information in adult cochlear implant users. *The Journal of the Acoustical Society of America*, *152*(4), A90–A90. https://doi.org/10.1121/10.0015644

Verschuure, J., & Brocaar, M. P. (1983). Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. *Perception & Psychophysics*, *33*(3), 232–240. https://doi.org/10.3758/BF03202859

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America*, *107*(3), 1671–1684. https://doi.org/10.1121/1.428451

Vestergaard, M. D., Fyson, N. R. C., & Patterson, R. D. (2011). The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening. *The Journal of the Acoustical Society of America*, *130*(1), 429–439. https://doi.org/10.1121/1.3596462

Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., & Başkent, D. (2019). Individual Variations in Effort: Assessing Pupillometry for the Hearing Impaired. *Trends in Hearing*, *23*, 233121651984559. https://doi.org/10.1177/2331216519845596

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392–393. https://doi.org/10.1126/science.167.3917.392

Winn, M. B., & 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 in Hearing*, *22*, 233121651880896. https://doi.org/10.1177/2331216518808962

Zaltz, Y., Goldsworthy, R. L., Kishon-Rabin, L., & Eisenberg, L. S. (2018). Voice Discrimination by Adults with Cochlear Implants: The Benefits of Early Implantation for Vocal-Tract Length Perception. *Journal of the Association for Research in Otolaryngology*, *19*(2), 193–209. https://doi.org/10.1007/s10162-017-0653-5