Talker variability in word recognition under cochlear implant simulation
Tamati, Terrin N; Sijp, Lissy; Başkent, Deniz

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/10.0001097

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Talker variability in word recognition under cochlear implant simulation: Does talker gender matter?
Terrin N. Tamati, Lissy Sijp, and Deniz Başkent

Citation: The Journal of the Acoustical Society of America 147, EL370 (2020); doi: 10.1121/10.0001097
View online: https://doi.org/10.1121/10.0001097
View Table of Contents: https://asa.scitation.org/toc/jas/147/4
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Acoustic factors affecting interaural level differences for cochlear-implant users
The Journal of the Acoustical Society of America 147, EL357 (2020); https://doi.org/10.1121/10.0001088

Production-perception relationship of Mandarin tones as revealed by critical perceptual cues
The Journal of the Acoustical Society of America 147, EL301 (2020); https://doi.org/10.1121/10.0000963

Visual analog of the acoustic amplitude envelope benefits speech perception in noise
The Journal of the Acoustical Society of America 147, EL246 (2020); https://doi.org/10.1121/10.0000737

Toddlers' fast-mapping from noise-vocoded speech
The Journal of the Acoustical Society of America 147, 2432 (2020); https://doi.org/10.1121/10.0001129

Training a U-Net based on a random mode-coupling matrix model to recover acoustic interference striations
The Journal of the Acoustical Society of America 147, EL363 (2020); https://doi.org/10.1121/10.0001125

Perception of musical melody and rhythm as influenced by native language experience
The Journal of the Acoustical Society of America 147, EL385 (2020); https://doi.org/10.1121/10.0001179
Talker variability in word recognition under cochlear implant simulation: Does talker gender matter?

Terrin N. Tamati,1,2,a) Lissy Sijp,3 and Deniz Başkent2

1Department of Otolaryngology—Head and Neck Surgery, The Ohio State University Wexner Medical Center, 915 Olentangy River Road, Suite 4000, Columbus, Ohio 43212, USA
2Department of Otorhinolaryngology/Head and Neck Surgery, University of Groningen, University Medical Center Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands
3Department of Linguistics, University of Groningen, Oude Kijk in ‘t Jatstraat 26, 9712EK, Groningen, The Netherlands
terrin.tamati@osumc.edu, l.g.sijp@student.rug.nl, d.baskent@umcg.nl

Abstract: Normal-hearing listeners are less accurate and slower to recognize words with trial-to-trial talker changes compared to a repeating talker. Cochlear implant (CI) users demonstrate poor discrimination of same-gender talkers and, to a lesser extent, different-gender talkers, which could affect word recognition. The effects of talker voice differences on word recognition were investigated using acoustic noise-vocoder simulations of CI hearing. Word recognition accuracy was lower for multiple female and male talkers, compared to multiple female talkers or a single talker. Results suggest that talker variability has a detrimental effect on word recognition accuracy under CI simulation, but only with different-gender talkers. © 2020 Acoustical Society of America

1. Introduction

Trial-to-trial changes in talkers’ voices have been shown to result in slower and less accurate spoken word recognition compared to single-talker conditions for normal-hearing (NH) listeners [e.g., Creelman (1957), Martin et al. (1989), and Mulennix et al. (1989)]. The effect of talker variability on spoken word recognition further interacts with the characteristics of the lexical items, with relatively greater effects of talker variability (i.e., poorer recognition) on more phonologically confusable lexically hard words (low frequency, high neighborhood density) than lexically easy words (high frequency, low neighborhood density) (Sommers et al., 1997; Sommers and Barcroft, 2006). Talker variability introduces changes to acoustic properties of speech relevant to phonetic perception that listeners must accommodate in order to achieve successful word recognition (Sommers and Barcroft, 2006). The talker variability effect on spoken word recognition can be attributed to the additional processing costs incurred from adjusting to these changes in talkers’ voices [e.g., Martin et al. (1989), Mulennix et al. (1989), Goldinger et al. (1991), and Nusbaum and Magnuson (1997)]. Additionally, the extent of the detrimental effects may rely upon the detection of talker change and the relative ease at which listeners can do so [e.g., Nusbaum and Morin (1992), Magnuson and Nusbaum (2007), and Barreda (2012)].

Cochlear implant (CI) users rely upon a speech signal that is highly degraded in spectro-temporal detail due to the limitations of the electrode-nerve interface and the relatively broad electric stimulation of the auditory nerve [for a review, see Başkent et al. (2016)]. CI users demonstrate a deficit in the perception of talker voice cues, such as F0 and vocal tract length (VTL) [e.g., Fuller et al. (2014), Meister et al. (2016), and Gaudrain and Başkent (2018)]. Relatedly, compared to NH listeners, CI users show poor discrimination of same-gender talkers [e.g., McDonald et al. (2003) and Cleary et al. (2005)], but are generally able to achieve some different-gender talker discrimination, although with different reliance on F0 and VTL cues compared to NH listeners [e.g., Fuller et al. (2014)]. Moreover, a large range of individual differences can be observed in sensitivity to talker voice cues (Gaudrain and Başkent, 2018).

The extent to which this limitation in talker perception influences CI word recognition is still largely unknown. Some previous studies suggest that the limitations of electric hearing may reduce the effects of talker variability, reporting no difference in the recognition accuracy between single- and multiple-talker conditions for CI users (Kirk et al., 2000). However, other studies examining word recognition in auditory only (Sommers et al., 1997; Kaiser et al., 2003)
and audiovisual conditions (Kaiser et al., 2003) and on vowel identification (Chang and Fu, 2006) have found poorer word or vowel recognition accuracy in multiple-talker compared to single-talker conditions in CI users. Acoustic simulations of electric hearing, simulating CI electric hearing only, have also been used to more systematically explore factors involved in word recognition with single and multiple talkers. Manipulating processing parameters, such as the carrier type and envelope cutoff frequency, alter the acoustic cues available, resulting in less/more availability of important talker cues, such as F0, that can be used to discriminate talkers or talker gender from the simulated speech [e.g., Gaudrain and Başkent (2015)]. While a cutoff frequency of 300 Hz represents the upper limit of temporal pitch perception in CIs [e.g., Zeng (2002)], in simulations higher cutoff frequencies and sine-wave carriers typically result in stronger voice pitch perception [e.g., Gaudrain and Başkent (2015)]. Varying carrier type and envelope cutoff frequency, Chang and Fu (2006) found no effect of talker variability for NH listeners with noise vocoded speech with a 160 Hz temporal envelope cutoff frequency but did find an effect of talker variability with sine-wave vocoded speech with 20 and 160 Hz cutoff frequencies. Given previous findings demonstrating good voice gender discrimination with sine-wave vocoded speech with 160 Hz cutoff but not with 20 Hz (Fu et al., 2004), findings from Chang and Fu (2006) suggest a limited association between talker variability effects and the detection of talker change in CI users.

Consistent with NH accounts, these varied findings may be at least partially due to differences in the similarity of talkers’ voices used in the multiple-talker conditions as well as individual differences in CI users’ sensitivity to talker differences. Yet, no study has directly investigated the potential factors underlying single- and multiple-talker word recognition in CI users. The current study investigated the effects of talker voice differences on multiple-talker word recognition, while partially controlling for individual listener differences. In particular, using acoustic noise-vocoder simulations of CI hearing with NH listeners, the study aimed to establish, first, whether a detrimental effect of talker variability can be observed with degradations of acoustic noise-vocoder simulations of CI hearing (single vs multiple talkers) and, second, whether the extent of the effect depends on the type of talker differences (same-gender vs different-gender). Given previous studies demonstrating greater talker variability effects for lexically hard words (Sommers et al., 1997; Sommers and Barcroft, 2006), we examined talker variability effects on the recognition of both lexically easy and lexically hard words under CI simulation.

2. Methods

2.1 Listeners

Forty native Dutch speakers (33 female, ages 18–34 years, M = 25.4 years, SD = 4.6 years) participated in the current study. Participants had normal hearing, with pure tone thresholds ≤25 dB hearing level (HL) at frequencies between 250 and 8000 Hz. Listeners received 10 euros for their participation.

2.2 Materials

Target words were 40 Dutch CVC words, originally included in the lists of words of NVA corpus (Bosman, 1989). Words were selected based on lexical characteristics; 20 words were “easy” (high frequency, low density) and 20 were “hard” (low frequency, high density). Lexical frequency and density properties of all words in the NVA corpus were calculated using the CLEARPOND database (Marian et al., 2012). The mean word frequency for the easy and hard words were 534.3 and 5.3 per million words, respectively. The mean number of phonologically similar words (neighborhood density) was 11.0 for the easy words and 20.8 for the hard words.

For the current study, target words were produced by 10 female speakers (ages 19.2–21.6 years, M = 20.8 years, SD = 0.9 years) and 5 male speakers (ages 22.1–33.8 years, M = 25.4 years, SD = 4.8 years). All speakers were native speakers of Dutch and received 40 euros for their participation in a larger recording session. For the recordings, talkers were asked to read aloud the words (presented visually one-by-one on a computer screen) in a natural speaking style. Participants wore a Shure head-mounted microphone (SM10A), positioned approximately 2 cm from the left corner of the mouth. The microphone output was fed to an Applied Research Technology microphone tube pre-amplifier connected to a MOTU MicroBooc IIc, which digitized the signal and transmitted it via USB ports to the laptop, where each utterance was recorded in a WAV 16-bit digital sound file at a sampling rate of 44.1 kHz using Audacity. Overall, each speaker participated in two sessions of two hours, recording words, sentences, and paragraphs. For the current study, target stimuli were saved into individual files and equated in level using PRAAT (Boersma and Weenink, 2017).
Word recordings were processed through a noise-vocoder with eight spectral channels, representative of recognition accuracy by typical CI users [e.g., Friesen et al. (2001)], distributed within the frequency range of 150 to 7000 Hz (Greenwood, 1990), using scripts maintained by Deniz Başkent’s Speech Perception Laboratory (dB SPL) at the University Medical Center Groningen and implemented in MATLAB [e.g., Gaudrain and Başkent (2015)]. Stimuli and white noise were filtered using eighth order, zero-phase, bandpass filters, producing analysis and synthesis (carrier) bands, respectively. The temporal envelope in each analysis band was extracted using half-wave rectification and low-pass filtering at 300 Hz, using a zero-phase, fourth order Butterworth filter. The cutoff frequency of 300 Hz was selected to represent the upper limit of temporal pitch perception in CIs [e.g., Zeng (2002)]. The noise carriers in each channel were modulated with the corresponding extracted envelope and the modulated noise bands from all vocoder channels were added together to construct the final stimuli.

2.3 Procedure

Participants were seated at a distance of 1 m from the speaker in a soundproof room. All stimuli were presented via a loudspeaker at approximately 68 dB SPL. To familiarize the participants with CI simulations, an eight-channel noise vocoded version of “The North Wind and the Sun” in Dutch (International Phonetic Association, 1999) was played prior to the first experimental block. For each trial, a tone (250 Hz) was played for 100 ms, followed by a 1000 ms silence and then the stimulus item. Each participant was asked to repeat the target word as fast as possible without compromising accuracy.

Each participant completed the word recognition task with three conditions: single talker (ST), multiple female talkers (MT-Female), and multiple male and female talkers (MT-Mixed). The ST condition consisted of the 40 target words produced by the same talker. The MT-Female conditions included the 40 target words produced by the 10 different female talkers (4 words per talker). The MT-Mixed condition consisted of the 40 target words produced by 5 male and 5 female talkers (4 words per talker). A subset of 7 female talkers (out of the aforementioned 10 female talkers) appeared in the MT-Mixed and ST conditions. All 5 male talkers appeared in the ST and MT-Mixed condition. For the ST condition, listeners were randomly assigned one of the talkers included in the MT-Mixed condition. Across all listeners, these talkers were presented in the ST condition 4 times across listeners.

A pilot experiment with eight NH native Dutch listeners was carried out to evaluate talker discrimination for same-gender and different-gender talker pairs from the current study. Participants listened to pairs of words from the target stimuli (described above) and indicated whether the words were produced by the same talker or different talkers for three different talker pair conditions: same-talker; same-gender, different-talker; different-gender, different-talker. Proportions of same-talker or different-talker responses are provided in Table 1. Listeners reported hearing the same talker more often for same-gender, different-talker pairs, only achieving 41% (SD = 13%) correct discrimination, than for different-gender, different-talker pairs, achieving 65% (SD = 12%) correct discrimination. Same-gender, same-talker pairs were at 66% (SD = 13%) correct. Thus, although performance was low overall, the NH listeners tended to show better discrimination of different-gender talkers compared to same-gender talkers for the target words under CI simulation, suggesting that the talkers in the different-gender condition of the main experiment would be more likely to be perceived as different talkers, compared to the same-gender condition.

Listeners were also randomly assigned one of four possible condition orders prior to testing (1: ST, MT-Female, MT-Mixed; 2: ST, MT-Mixed, MT-Female; 3: MT-Female, MT-Mixed, ST; 4: MT-Mixed, MT-Female, ST). Crucially, all conditions appeared in the first experimental block to account for any learning effects. A microphone standing approximately 30 cm away

<table>
<thead>
<tr>
<th>Target talker pair</th>
<th>Same talker</th>
<th>Different talker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same talker</td>
<td>66.3</td>
<td>33.7</td>
</tr>
<tr>
<td>Different talker</td>
<td>59.6</td>
<td>40.4</td>
</tr>
<tr>
<td>Different gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Different talker</td>
<td>34.8</td>
<td>65.2</td>
</tr>
</tbody>
</table>
from the participant recorded both stimuli and responses. Accuracy and response time (RT) were collected; only RTs for correct answers were included in the analysis.

3. Results

Mean accuracy scores across conditions are displayed in Fig. 1. A repeated measures analysis of variance (ANOVA) on accuracy scores, with within-subject Talker Condition (ST, MT-Female, MT-Mixed) and Lexical Difficulty (Easy, Hard), revealed a main effect of Talker Condition \[F(2,78) = 31.3, \ p < 0.001, \ \eta_p^2 = 0.205\] and Lexical Difficulty \[F(1,39) = 38.5, \ p < 0.001, \ \eta_p^2 = 0.154\], and a significant Talker Condition × Lexical Difficulty interaction \[F(2,78) = 3.1, \ p = 0.006, \ \eta_p^2 = 0.020\]. Post hoc Tukey tests demonstrated that accuracy was significantly lower in the MT-Mixed condition than both the MT-Female and ST conditions (all \(p\)'s < 0.001). Accuracy was also significantly lower in the ST condition than the MT-Female condition (\(p = 0.003\)). Accuracy was significantly lower for Hard words than Easy words overall (\(p < 0.001\)). Examining the accuracy across Talker Condition for Easy and Hard words, accuracy varied less by Talker Condition for Easy words (ST—MT-Female: n.s.; ST—MT-Mixed: \(p = 0.012\); MT-Female—MT-Mixed: \(p < 0.001\)) than for Hard words (ST—MT-Female: \(p = 0.002\); ST—MT-Mixed: \(p = 0.053\); MT-Female—MT-Mixed: \(p < 0.001\)), although this appears to be mostly driven by the drop in performance for Hard words in the ST condition.

Mean RTs across conditions are displayed in Fig. 2. A repeated measures ANOVA on RT, with Talker Condition (ST, MT-Female, MT-Mixed) and Lexical Difficulty (Easy, Hard), revealed a significant main effect of Lexical Difficulty on RT \[F(1,39) = 16.6, \ p < 0.001, \ \eta_p^2 = 0.015\] and a significant Talker Condition × Lexical Difficulty interaction \[F(2,78) = 4.5, \ p = 0.014, \ \eta_p^2 = 0.004\]. Although RT was slightly slower for Hard words, as demonstrated in Fig. 2, post hoc Tukey tests revealed no significant differences in RT between Easy and Hard words. Exploring the interaction by examining accuracy across Talker Condition for Easy and Hard words, none of the Talker Condition comparisons reached significance for either Easy or Hard words.

To further examine talker variability effects, given that not all female talkers appeared in all Talker Conditions, additional analyses were carried out on the accuracy scores for only the five female talkers who appeared in all three Talker Conditions (ST, MT-Female, MT-Mixed). Scores were calculated from mean accuracy across all participants, as shown in Fig. 1. Paired-comparison \(t\)-tests showed that accuracy in MT-Mixed condition was significantly lower than in the ST condition \([t(18) = 2.49, \ p = 0.023, \ d = 0.420]\) and the MT-Female condition \([t(39) = 2.11, \ p = 0.041, \ d = 0.751]\), but the ST and MT-Female conditions were not significantly different. Again, mean RTs, shown in Fig. 2, for the same five talkers was not significantly different across Talker Conditions.

![Fig. 1. (Color online) Average word recognition accuracy (percent correct) by Talker Condition (ST, MT-Female, MT-Mixed) and Lexical Difficulty (Easy, Hard). The boxes extend from the lower to the upper quartile (the interquartile range, IQ), the solid midline indicates the median, and the bold X indicates the mean. Red diamonds indicate the mean scores for the five female talkers appearing in all Talker Conditions. The whiskers indicate the highest and lowest values no greater than 1.5 times the IQ, and the dots indicate the outliers, which are defined as data points larger than 1.5 times the IQ.](https://doi.org/10.1121/10.0001097)
4. Discussion

The current study investigated talker variability effects with NH listeners tested under acoustic CI simulation. Consistent with previous findings with NH listeners with unprocessed stimuli (Kaiser et al., 2003; Chang and Fu, 2006), a detrimental effect of talker variability was observed in the current study. However, here, the talker variability effect was observed only in the MT-Mixed condition. Word recognition accuracy was significantly lower when listeners were presented with mixed male and female talkers (MT-Mixed) compared to when they were presented with multiple female talkers (MT-Female) or a single talker (ST). This trend was observed in the larger analysis with all talkers, and in the separate analysis with only the female talkers who appeared in all three conditions. Across all talkers, word recognition accuracy was also lower in the ST than the MT-Female condition. However, this result may be due to the different talkers appearing in the ST and MT-Female conditions; only five talkers appeared in the ST and MT-Mixed conditions, but all ten female talkers appeared in the MT-Female condition. When examining the five female talkers who appeared in all three conditions, accuracy was not significantly different in the MT-Female compared to the ST condition, suggesting that the additional five female talkers in the MT-Female condition may have been slightly more intelligible than the five recurring female talkers, potentially resulting in higher accuracy in the MT-Female condition.

These findings suggest that while an effect of talker variability may be observed with spectro-temporally degraded speech, this effect appears to depend on whether talker voice differences are distinct enough to be perceptible even with this degradation. Findings are largely consistent with previous studies with CI users and NH listeners under CI simulation demonstrating poor same-gender talker discrimination but relatively good gender discrimination [e.g., McDonald et al. (2003) and Fuller et al. (2014)]. Talker voice cues crucial for same-gender talker discrimination are poorly conveyed in CI simulations, and accordingly same-gender talker variability appears to have little impact on word recognition accuracy. In contrast, between-gender variability seems distinct enough to be at least partially conveyed in CI simulations, resulting in detectable talker voice differences that lead to a cost in accuracy. Whether this effect relies upon a change in talker gender is unclear; it may also be the case that same-gender talkers with drastically different voice and speaker characteristics may also yield a talker variability effect if also detectable under CI simulation. Further, the extent to which these results may be specific to female talkers is unclear, since a multiple-talker condition with male talkers was not included for comparison. It is possible that that the male talkers’ voices would be better discriminated, leading to differing word recognition performance with multiple male talkers. A future study should include same-gender, multiple-talker conditions with female talkers and with male talkers. Further, each talker should appear in both the single- and multiple-talker conditions, either balanced across listeners or randomly.

While differences in talkers’ voices are potentially discriminable, it is unclear from the current data whether the listeners were actually detecting and accommodating the variability during the word recognition task. In the current study, while accuracy was lower in the MT-Mixed condition, RT was consistent across conditions. The contextual tuning hypothesis of talker
normalization suggests that the accommodation of talker variability in multiple-talkers conditions may depend on the detection of talker changes [e.g., Nusbaum and Morin (1992), Magnuson and Nusbaum (2007), and Barreda (2012)]. According to this account, listeners will use extrinsic talker information in normalization when no talker change is detected, but will rely on intrinsic talker information when a talker change is detected. Consistent with this account, independently manipulating F0 and VTL, Barreda (2012) found that for conditions that likely did not result in the detection of a talker change (voices with differences in VTL, but similar F0), listeners showed lower vowel perception accuracy but relatively fast RTs. However, for conditions that likely resulted in the detection of a talker change (voices with F0 and VTL differences), listeners showed relatively better accuracy but slower RTs.

The findings in the current study may be consistent with those reported by Barreda (2012). Here, listeners may not have changed their perceptual strategy to accommodate the additional talker variability in the MT-Mixed condition, even if the talker differences were potentially detectable, resulting in poorer recognition accuracy but not slower RT. In the word recognition task, NH listeners under CI simulation may not have been devoting resources to accommodate the talker variability. Prior studies have similarly found limited effects of talker familiarity or change on sentence recognition after brief exposure under CI simulation [e.g., Huyck et al. (2017) and Kapolowicz et al. (2018)], when talker specificity effects are typically observed in NH listeners with unprocessed speech. While these findings with simulations, taken together, may suggest a limited role of talker variability in word recognition in CI users, long-term exposure and adaptation to spectrally degraded speech may improve the detection of between-gender talker changes and use of talker information to inform spoken word recognition in CI users.

Acknowledgments
We would like to thank Juliette Vertregt, Anne Nijman, and Rose van Doorn for their assistance with this project. The study was primarily supported by a VENI grant (Grant No. 275-89-035) from the Netherlands Organization for Scientific Research (NWO) to T.N.T., a VICI grant (Grant No. 918-17-603) from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw) to D.B., and funds from the Heinsus Houbolt Foundation.

References and links


