The influence of perceived talker differences on the talker variability effect under cochlear implant simulation

Lissy Sijp
Rijksuniversiteit Groningen
l.g.sijp@student.rug.nl

ABSTRACT

Speech recognition is harder when communicating with multiple talkers. This "talker variability effect" has not been extensively examined for cochlear implant (CI) users, who have difficulty discriminating same-gender talkers but not different-gender talkers. A shadowing task was conducted with normal hearing listeners (N=19) under CI simulation. The test consisted of a single-talker condition (ST), and two multi-talker conditions with different male and female voices (MT-M) or only different female voices (MT-F). Response times were longer and accuracy was lower in MT-M compared to MT-F. Thus, the talker variability effect was observed only when talkers were perceptible under CI simulation.

Keywords

Speech perception, talker variability, cochlear implant simulation.

INTRODUCTION

In real-life communication, the listener has to adapt to new talkers' voices in order to process the nonlinguistic information conveyed by the talker's voice, in addition to the linguistic information. Processing this information can be a challenge when listening to multiple talkers. This phenomenon is called "talker variability effect", which is the slowing or decrease in accuracy of word recognition under conditions of talker change compared with conditions of talker stability. Because the listener has to adapt to each new voice, understanding spoken language takes more time and effort, resulting in slower and less accurate word recognition in multi-talker conditions compared to single-talker conditions (e.g., [2, 26, 27]).

In order to adjust to talker variability, listeners must adapt to different phonetic features of different speakers' voices, such as fundamental frequency, vowel length and formant frequencies (e.g. [9, 10]). For normal hearing (NH) listeners it is easy to distinguish different voices and to perceive subtle differences. However, hearing-impaired users of cochlear implants (CIs) have difficulty distinguishing talkers' voices, because the quality of the input signal is degraded compared to NH people [14].

CIs are auditory prosthetic devices that enable deaf people and people with profound hearing loss to hear. While CI users show good speech understanding in favorable conditions (e.g., quiet, single talker), some details of the speech signal are lost and important cues characterizing a talker's voice (e.g. f0 and vocal-tract length (VTL)) are not well conveyed [16]. The phonetic-relevance hypothesis states that sources of variability that affect phonetically-relevant acoustic speech features cause decrements in spoken word recognition [27]. Because of their poor perception of voice cues, CI users have difficulty perceiving differences between talkers, and demonstrate poor talker discrimination and identification [7]. Other common problems CI users have, are understanding speech in noise and music perception [11].

Findings on the talker variability effect in CI users have been inconclusive. On the one hand, studies suggest that talker variability does not influence speech perception in NH listeners when they do not sense a talker change [1, 25]. Since CI users show poor talker discrimination and identification [7], the talker variability effect may not play a role in speech perception in CI users. On the other hand, researchers have shown that talker variability indeed plays a role in speech perception in CI users (e.g., [21]). However, there is a large variance in speech perception between CI users. CI users with poor talker perception would probably have overall poorer speech understanding. The difference between single- and multi- talker conditions would be relatively less for them than for CI users with good perception. Thus, talker variability may affect not all users and only in certain conditions. Therefore, a new step in research into CI speech perception is to investigate the factors that influence the size of the talker variability effect.

In the current study, the following research question was investigated: Do different talker voices influence the talker variability effect under CI simulation? A shadowing task was conducted with 8-channel acoustic simulation of CI hearing, representative of the average CI user [12]. In this task, a participant listens to isolated words and then immediately repeats them. There were three conditions: one single-talker (ST) and two different multi-talker conditions. Because in the multi-talker conditions each stimulus was produced by another talker, there was trial to trial variability in talkers' voices. The perceived similarity of different voices was manipulated by including only female talkers (MT-F) or mixed female and male talkers (MT-M). Using CI simulations is also useful to exclude additional factors interacting with hearing performance common to CI users [23], including surgical, device, and demographic factors [3].

CI users are generally able to achieve some level of gender categorization [18, 22], but, perceiving differences between same-gender voices is difficult [22]. Similarly,

SRC 2018, November 9, 2018, The Netherlands.

under CI simulation, gender discrimination may be good but same-gender talker discrimination may be challenging. It is expected that CI simulations will result in problems perceiving variability in relevant phonetic properties of speech [6, 21, 27]. Subsequently, we hypothesized that listeners would experience more difficulty – slower and less accurate word recognition – in multi-talker conditions than in single-talker conditions. Further, we predicted that between the two multi-talker conditions, the talker variability effect will be stronger when there are both male and female talkers than with only female talkers.

METHODS

Listeners

Listeners consisted of 19 native speakers of Dutch (2 male, 17 female), ages 18.2-24.5 yr. (M = 21.4 yr.; SD = 1.7 yr). Listeners had pure tone thresholds better than 25 dB at frequencies between 250 and 8,000 Hz. The study was approved by the Medical Ethical Committee of the University Medical Center Groningen. Listeners were provided with detailed information about the study, and written informed consent was obtained. Listeners received 10 euros for their participation.

Materials

For the stimuli, target words were selected from the NVA corpus [5], which consists of monosyllabic words and is often used in clinical setting for hearing tests. Half of the selected words were easy, based on lexical frequency and density characteristics generally associated with easier recognition, and the other half were hard (low frequency, high density) [24]. Five male speakers, ages 22.1-33.8 yr. (M = 25.4 yr.; SD = 4.8 yr.) and ten female speakers, ages 19.2-21.6 yr. (M = 20.8 yr.; SD = 0.9 yr.) produced the words. All speakers were native speakers of Dutch and received 40 euros for their participation in a larger recording session. For the recordings, speakers were asked to read aloud the words in a natural speaking style. The words were presented visually one-by-one on a MacBook laptop using PsyScope X Build 77 [8]. Speakers wore a Shure head-mounted microphone (SM10A), positioned approximately two centimeters from the left corner of the mouth. The microphone output was fed to an Applied Research Technology microphone tube pre-amplifier. The output of the microphone pre-amplifier was 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 recording sessions of two hours, involving words, sentences, and paragraphs. Audio files were edited using PRAAT [4], creating separate audio files for each stimulus.

Stimulus words were processed through an 8-channel noise-band vocoder implemented in Matlab. To achieve this, the original signal was filtered into 8 bands between 150 and 7000 Hz, using 8th order, zero-phase Butterworth filters, based on previous studies (e.g., [15]). The bands were partitioned based on Greenwood's frequency-to-place mapping function, simulating evenly spaced regions of the cochlea [17]. The same cutoff frequencies were used for both the analysis and synthesis filters. The temporal envelope from each frequency band was extracted by half-wave rectification and low-pass filtering at 300 Hz, using a zero-phase 4th order Butterworth filter. Filtering white noise into spectral bands was done to generate noise-band carriers independently for each channel. For this, the same 12th order Butterworth bandpass filters were used. To construct the final stimuli, 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.

In the ST condition, 40 different words were spoken by the same speaker. In the MT-F condition, ten different female talkers produced the same 40 words (4/talker). The MT-M condition was almost the same, except that there were five different female talkers and five different male talkers. Across listeners, all individual speakers appeared in the ST condition twice (two listeners), so that an effect of condition could not have been caused by the baseline intelligibility of individual talkers. The multi-talker conditions were the same for each listener.

Procedure

Before testing, listeners filled out an online questionnaire, with questions about the listeners' residential history, language background, and hearing history. The listener underwent a hearing screening the day of testing. The listeners were seated at a distance of 1 meter from the speaker in a soundproof room. All stimuli were presented via a loudspeaker at 65 dB. To familiarize the listeners with CI simulations, an 8-channel noise vocoded version of "The North Wind and the Sun" in Dutch [20] was played prior to the first experimental block.

On each trial, a tone (250 Hz) was played for 100 milliseconds, followed by a 1000 milliseconds silence and then the stimulus item (via PsyScope X Build 77 [8]). The listener then repeated what he/she has heard as fast as possible without compromising accuracy. A microphone standing approximately 30 centimeters away from the listener recorded both stimuli and responses. In PRAAT [4], accuracy (right or wrong) and response time for right answers were annotated. Every listener started with the single-talker condition, but the order of the multi-talker blocks was balanced over listeners.

RESULTS

The dependent variables were accuracy and response time. Talker condition was the independent variable.

Accuracy

Mean accuracy across conditions is displayed in Figure 1. Accuracy in the ST condition was lower (M = 0.704, SD = 1.131) than in the MT-F condition (M = 0.807, SD = 0.074). MT-F was higher than accuracy in the MT-M condition (M = 0.658, SD = 0.077).

A repeated measures ANOVA, with condition (ST, MT-F and MT-M) as a within-subjects factor revealed a main effect of condition ($F(2, 36) = 21{,}584$, $p < 0.001$). Post-hoc Bonferroni tests showed that accuracy in the ST condition was significantly lower than in the MT-F condition ($p = 0.002$) and accuracy in the MT-F condition was significantly higher than in the MT-M condition ($p < 0.001$). However, the ST condition did not differ significantly from the MT-M condition ($p = 0.212$).

Response time

Response times in the ST (single-talker) condition were slower (M = 2461.95, SD = 226.794) than in the MT-F (multi-talker female only condition (M = 2353.80, SD =

188.762). Response times in the MT-F condition were faster than in the MT-M (multi-talker mixed) condition (M = 2443.90, SD = 238.197).

Again, a repeated measures ANOVA, with condition (ST, MT-F and MT-M) as a within-subjects factor revealed a main effect for condition ($F(2, 36) = 9.874$, $p < 0.001$). Post-hoc Bonferroni tests showed that response time in the ST condition was significantly slower than in the MT-F condition ($p = 0.002$) and response times in the MT-F condition were significantly slower than in the MT-M condition ($p = 0.005$). However, the ST condition did not differ significantly from the MT-M condition ($p = 1.000$).

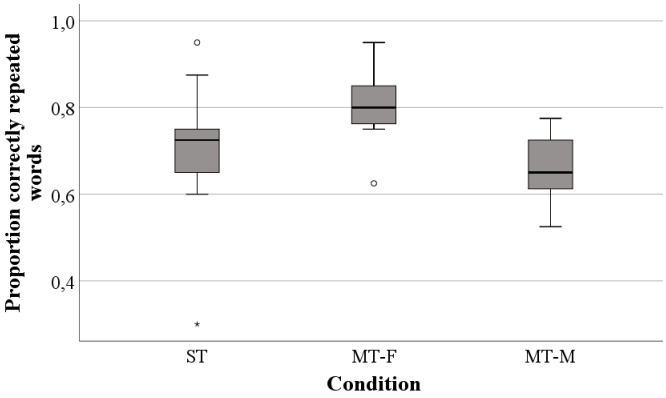

Figure 1. Accuracy in all three conditions.

CONCLUSION

The current study investigated whether different talker voices influence the talker variability effect under CI simulation. Contrary to our initial prediction and to previous studies (e.g., [25, 26]), listeners responded faster and more accurately in the MT-F condition compared to the ST condition. In order to understand this unexpected result, scores on the ST condition were further examined (see Table 1). Adaptation to vocoded speech occurs very quickly within about 20 sentences [19] but may be harder when listening to words. During the last ten trials of the ST condition, the response time was shorter and almost equal to the response time in the MT-F condition. Therefore, one could assume that listening to the vocoded paragraph was not enough to fully adapt to the degradation, and listeners needed additional exposure.

Table 1
Detailed results for ST and MT-F.

	Accuracy	**Response time**
ST (entire block)	0.704	2461.95
ST (last 10 trials)	0.758	2365.35
MT-F (entire block)	0.807	2353.80

The interesting finding from the current study is the comparison of the two multi-talker conditions. Consistent with the second hypothesis, understanding words was harder when there were speakers from different genders. Since CI simulation approximates CI hearing, we expected same-gender voice differences to be more difficult to perceive than different-gender voice differences. Listeners responded more slowly and less accurately in the MT-M condition compared to the MT-F condition, consistent with studies which showed that CI users are able to identify speakers of different genders [18, 22]. However, the talker variability effect does not play a role in the MT-F condition, which could maybe be explained by a study which states that same-gender voice differences are not perceived under CI simulation [22]. Thus, the talker variability effect relies upon good perception of talker differences, which CI users do not have.

While all listeners started with the ST condition, the sequence of the multi-talker conditions was balanced, half completed ST – MT-F – MT-M, the other half completed ST – MT-M – MT-F. Future research would require balancing the presentation of all three conditions across listeners. As such, it is expected that accuracy and response time would be similar between ST and MT-F conditions. Another way to prevent this learning effect, is to add a training session before starting with the actual experiment, so adapting to CI simulation would be finished before starting with the actual experiment. However, to prevent any kind of learning effect, applying counterbalancing as well as a training session would be recommended.

The present results could be helpful to improve the hearing training CI patients receive after being implanted. Optimization of speech perception depends on passive learning (daily listening), and active learning, via auditory training programs in the clinic [13]. In the clinic, CI users are often tested with single-talker tests that reduce real-life sources of degradation. Most CI users score well on these speech perception tests. However, outside the clinic, patients often have problems with understanding speech in noise [11] or dealing with other sources of variability, such as different speaking styles [29] or regional accents [28]. Based on the current research, one could say that clinical hearing training could maybe improve by making use of different male and female voices, to see if this helps CI users getting better used to daily hearing situations.

To conclude, the size of the talker variability effect under CI simulation depends on whether differences in talkers' voices are perceptible under CI simulation. If listeners are unable to detect differences between talkers' voices, talker variability does not influence word recognition. However, when listeners do detect differences, listeners respond more slowly and less accurately. Future research should include counter balancing over all three conditions and a training session to get listeners used to the sound of a CI, in order to prevent a learning effect and to investigate the size of the talker variability effect relative to single-talker conditions. In addition, research should be carried out directly with CI users to investigate whether the talker variability effect further varies based on individual listener's auditory functioning and perceptual skills. The implication for actual CI users, who vary greatly in their ability to hear differences both between talkers of the same gender and between different genders (e.g., [14 & 16]), would be that some implant users may be more susceptible to talker variability in real-life speech communication and may struggle to understand speech outside the clinic. Results of the current study may implicate that hearing training for CI patients could be improved by using multiple talkers in speech recognition assessments and/or training protocols.

ROLE OF THE STUDENT

Lissy Sijp was an undergraduate student when carrying out this research in 2018 and worked under the supervision of dr. Terrin Tamati and dr. Simone Sprenger. The topic was

proposed by the supervisor. The test was developed together with the supervisor. Processing of the data and writing were done by the student.

ACKNOWLEDGMENTS

I would like to thank Terrin Tamati and Simone Sprenger for their helpful feedback and the participants for their time.

REFERENCES

1. Barreda, S. (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *J Acoust Soc Am, 132*(5), 3453-3464.
2. Bent, T., Holt, R.F. (2013). The influence of talker and foreign-accent variability on spoken word identification. *J Acoust Soc Am, 133*(3), 1677-1686.
3. Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., . . . Lazard, D. (2013). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: An update with 2251 patients. *Audiology and Neuro-Otology, 18*(1), 36-47.
4. Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36, retrieved 28 November 2017 from http://www.praat.org/
5. Bosman, A.J. (1989). Speech perception by the hearing impaired. Proefschrift Universiteit Utrecht.
6. Chang, Y., & Fu, Q. (2006). Effects of talker variability on vowel recognition in cochlear implants. *Journal of Speech, Language, and Hearing Research, 49*(6), 1331-1341.
7. Cleary, M., Pisoni, D.B., Kirk, K.I. (2005). Talker discrimination in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research, 48*, 204-223.
8. Cohen, J.D., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 25*, 257–271.
9. Coleman, R. O. (1971). Male and Female Voice Quality and Its Relationship to Vowel Formant Frequencies. *Journal of Speech, Language, and Hearing Research, 14*, 565-577.
10. Coleman, R. O. (1976). A Comparison of the Contributions of Two Voice Quality Characteristics to the Perception of Maleness and Femaleness in the Voice. *Journal of Speech, Language, and Hearing Research, 19*, 168-180.
11. Faulkner, K.F., & Pisoni, D.B. (2013). Some observations about cochlear implants: challenges and future directions. *Neuroscience Discovery, 1*(1), 1-10.
12. Friesen, L., Shannon, R., Başkent, D., & Wang, Y. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am, 110*(2), 1150-1163.
13. Fu, Q., & Galvin, J. (2007). Perceptual learning and auditory training in cochlear implant recipients. *Trends in Amplification, 11*(3), 193-205.
14. 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.
15. Gaudrain, E., & Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *J Acoust Soc Am, 137*(3), 1298–1308.
16. Gaudrain, E., & Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear & Hearing, 39*(2), 226-237.
17. Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J Acoust Soc Am, 87*, 2592 – 2605.
18. Hazrati, O., Ali, H., Hansen, J.H.L. & Tobey, E. (2015). Evaluation and analysis of whispered speech for cochlear implant users: Gender identification and intelligibility. *J Acoust Soc Am, 138*(1), 74-79.
19. Huyck, J.J., Smith, R.H., Hawkins, S. & Johnsrude, I.S. (2017). Generalization of Perceptual Learning of Degraded Speech Across Talkers. *Journal of Speech, Language and Hearing Research, 60*, 3334-3341.
20. International Phonetic Association, 1999. Handbook of the International Phonetic Association. Cambridge: Cambridge University Press.
21. Kirk, K.I., Pisoni, D.B., & Miyamoto, R.C. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. Journal of Speech, Language, and Hearing Research: *Journal of Speech, Language, and Hearing Research*, *40*(6), 1395-405.
22. Loebach, J.L., Bent, T., & Pisoni, D.B. (2008). Multiple routes to the perceptual learning of speech. *J Acoust Soc Am, 124*(1), 552-561.
23. Loizou, P.C. (1998). Mimicking the human ear. *IEEE Signal Processing Magazine, 15*, 101-130.
24. Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing, 19*(1), 1-36.
25. Magnuson, J.S. & Nusbaum, H.C. (2007). Acoustic Differences, Listener Expectations, and the Perceptual Accomodation of Talker Variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391-409.
26. Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *J Acoust Soc Am, 85*, 365-378.
27. Sommers, M.S. & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J Acoust Soc Am, 119*(4), 2406-2416.
28. Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2014). Influence of Early Linguistic Experience on Regional Dialect Categorization by an Adult Cochlear Implant User: A Case Study. *Ear and Hearing, 35*, 383-386.
29. Tamati, T., Janse, E., & Başkent, D. (in press). Perceptual Discrimination of Speaking Style Under Cochlear Implant Simulation. *Ear and Hearing, 00*. doi:10.1097/AUD.00000000