INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Can Better Perception Become a Disadvantage? Synthetic Speech Perception in Congenitally Blind Users

Gerda Ana Melnik-Leroy[1], *Gediminas Navickas*[1]

[1]Institute of Data Science and Digital Technologies, Vilnius University, Lithuania
gerda.melnik@mif.vu.lt, gediminas.navickas@mif.vu.lt

Abstract

Modern Text-To-Speech systems are rarely tested on non-standard user groups, such as people with impairments. Nevertheless, evidence suggests that some of these groups might perceive synthetic speech differently (better or worse) than regular users. The current study investigated for the first time how synthetic speech is perceived by blind vs. sighted users. For this purpose, we used a speeded AX discrimination task and tested how sighted and blind listeners perceive synthetic speech of different qualities. Results show that blind participants had significantly better discrimination on this task, and both groups performed worse when the perceptual differences in the synthetic speech were smaller. This suggests that blind participants were indeed more sensitive to the acoustic characteristics of synthetic speech compared to their sighted peers. We discuss implications for speech perception and the development of modern speech technologies.

Index Terms: speech synthesis, Text-To-Speech, blind users, speech perception, behavioral methods, synthesized speech evaluation

1. Introduction

Although contemporary Text-To-Speech (TTS) systems made great advances and can produce high-quality speech, listeners have still much more difficulty perceiving it compared to natural speech [1], [2]. Moreover, some groups of listeners, such as non-native speakers, children, elderly adults or people with impairments have even more difficulty when dealing with this type of speech [3]–[5]. Although such non-standard users are hardly ever tested, with the rising prevalence of modern speech technologies, the study of synthetic speech perception across various groups becomes indispensable. This is especially true for blind listeners, who are dependent on text-to-speech technologies in their daily activities [6], [7].

There is substantial evidence that due to a neuronal reorganization, blind listeners have more effective perceptual processing of auditory information [8], [9]. Specifically, studies in neuroscience and neurophysiology have provided evidence that cross-modal plasticity occurs in the brain of congenitally blind listeners, and the visual cortex takes over some of the functions of auditory processing [10]. Over the last decade, there has been an increase in research showing that blind people perceive linguistic information more accurately than sighted people under a variety of conditions [8], [11], [12]. However, to our knowledge, no study examined whether this perceptual advantage in blind participants also affects their perception of synthetic speech. As synthetic speech is perceived differently than natural speech in sighted participants [2], this advantage in blind participants could as well disappear when processing synthetized speech. If, however, this perceptual advantage persists in synthetic speech processing, blind target users could be much more sensitive to signal distortions than their sighted peers. As TTS technologies are usually being built to suit sighted listeners only, blind participants might be in this way disadvantaged compared to their sighted counterparts.

In this study, we address the question of synthetic speech perception in congenitally blind participants, using a controlled experimental paradigm. The study has a twofold objective: first, to compare the perception of two groups of listeners, namely sighted and congenitally blind participants. Second, to evaluate the perceptual differences between objectively different synthesized speech qualities. Following recent calls to innovate methods of synthetic speech evaluation [13]–[15], we propose that more accurate and sensitive behavioral measures are necessary for the evaluation of speech perception rather than the traditionally used scales, such as MOS (mean opinion score) or its modifications [16]. In this study we therefore used a well-known experimental method from psycholinguistics, i.e. a speeded AX discrimination task [17]. The advantage of this paradigm lies in its simplicity and the possibility to record participants' perception (or lack of it) without having to rely on subjective measures that are difficult to quantify, such as the 'naturalness' etc. We built an experiment, which allowed to test sighted and blind listeners' perception of different synthetic speech qualities obtained by using different data set sizes for the acoustic model training. Our study is a first attempt to bring a better understanding of synthetic speech perception in congenitally blind vs. sighted participants, thus providing insights for the field of speech perception and practical implications for the development of modern speech technologies.

2. Methods

2.1. Participants

Sixteen blind and thirty-three sighted control participants took part in the study. The blind participants were recruited and tested at the Lithuanian Audiosensory Library, while the sighted control participants were recruited and tested at Vilnius University. The four blind participants who lost their sight at age 10 or later had to be removed, as only congenitally blind participants could be tested for comparability reasons (as neural plasticity reduces later on in life, the characteristics of speech processing could differ in such participants). In addition to this, four participants were removed from the sighted group, as their native language was other than Lithuanian. Thus, data from twelve blind participants (3 females and 9 males) and twenty-nine controls (16 females and 13 males) was included in the data analyses. They were all native speakers of Lithuanian.

10.21437/Interspeech.2023-2013

None of the participants reported a history of hearing or speech problems.

2.2. Stimuli

92 words or short sequences of words in Lithuanian were used for the experiment, with additional 4 items for the participant training phase. In order to test the possible differences in perception in both groups we chose to create synthetic stimuli of three different qualities using different amounts of training data (more details below). The rationale behind this choice was that smaller training data sets contain less acoustic information and thus are less representative of natural speech. This in turn introduces certain distortions in the signal at the segmental level or in the transitions between segments. The stimuli set contained 23 items synthesized both in low and high quality (for the different pairs in the Easy condition); 23 items in low and average quality (for the different pairs in the Difficult condition). The stimuli for the same trials consisted of 12 items synthesized in good quality; 12 in average quality and 22 in low quality. This resulted in a total of 138 synthesized items. Two professional phoneticians checked that the stimuli for the different trials (i.e. those, comparing items of two different qualities) contained a perceptible acoustic difference and validated the stimuli set.

Stimuli for the tasks were synthesized using the Merlin Toolkit for Deep Neural Network models adapted for Lithuanian language [18]. The Lithuanian language corpus LIEPA created in Vilnius University was used to train the model [19]. Three qualities were obtained for the synthetic stimuli by using different amounts of training data: low quality (training data consisted of 400 sentences), medium quality (training data consisted of 800 sentences) and high quality (training data consisted of 1600 sentences). All stimuli for the experiment consisted of Lithuanian words or short sequences of words (1-3 words) spoken by a young female voice, recorded at 16000 Hz.

Note that the Merlin Toolkit can be used as an end-to-end speech synthesis system for English using an external front-end, such as Ossian or Festival which is doing the text processing, and one of the two supported vocoders (WORLD and STRAIGHT) which generate the waveform. However, as none of the front-ends work for Lithuanian language, they could not be applied. Instead, we used already processed text as the input for the Merlin Toolkit. The text was processed and formatted as HTS-style labels (also called lab files) with state-level alignment. Lab files contained linguistic features (phonemes, their position, syllable count, durations etc.). The toolkit converted such labels into vectors of binary and continuous features for neural network input (NN). Merlin allows to use different neural networks for the training of the acoustic model. A simple architecture of the feedforward neural network (DNN) with default hyper parameters was used for the creation of the acoustic model. The architecture consisted of 5 hidden layers and 1024 nodes in each layer. Hyperbolic tangent activation function was used for hidden layers. The open source vocoder WORLD was used for the waveform generation.

2.3. Procedure

The experimental paradigm was an AX (same-different) speeded discrimination task, in which participants heard two items in each trial and had to answer as quickly as possible whether they sounded the same or different. Half of the trials (same trials) contained identical recordings (the same word of the same quality was played twice), half of them (different trials) were of different speech qualities (the same word was played in two different qualities). Stimuli were presented binaurally over closed-ear headphones at a comfortable listening level. Participants provided their response by pressing the right vs. the left arrow keys on the keyboard. Both groups of participants completed two condition blocks: in the Easy condition, the items in different trials were of low vs. high quality synthetic speech. As the acoustic difference between those two qualities is large, the discrimination task was expected to be easier in this condition. Conversely, in the Difficult condition, participants heard medium vs. high quality synthetic speech stimuli, which acoustically differed less. The trials were presented in a pseudo-random order, such that no more than three trials of the same type (same or different) would appear in a row.

A speeded AX paradigm with an interstimulus interval (ISI) of 600 ms was chosen in order to ensure that the task taps into processing at a phonetic-phonological stage (as compared to an acoustic mode of perception with short ISI of ~100ms). This relies on evidence that speech processing in humans consists of several stages: in order to decode the incoming acoustic signal into meaningful words the listener has to succeed in accurately performing throughout stages, starting from auditory processing, phonetic and phonological analysis, to word recognition and lexical access [20]. Previous research has shown that variable memory demands and cognitive load (e.g. different ISI length) in the task triggers different processing levels [21]. Hence, the performance on perceptual tasks can vary depending on the level of processing being tested as even difficult sounds or small acoustic details can be perceived at the low acoustic level [22], [23]. As this study focused on synthetic speech perception of full words or sequences of words and thus involved higher short-term memory demands, we aimed at triggering the phonetic-phonological stages of processing and chose a relatively high ISI of 600 ms.

Each condition block started with a practice phase of four trials, during which participants received feedback as to whether their responses were correct. In the case of an incorrect response or no response within 2500 ms of the stimulus offset, the trial was repeated until the correct response was given. During the test phase, participants received no feedback and if they did not respond within 2500 ms the next trial was presented. A silent interval of 1000 ms separated the participant’s response or the time-out from the presentation of the next stimulus. Participants could take a short break in between condition blocks.

3. Results

Prior to analysis, we inspected the performance of both groups of participants to detect outlier items. Items were discarded if their rate of correct responses was three standard deviations below the mean of all items in the same condition. Two words were thus discarded from the same trials in the Easy condition. Accuracy scores for blind and sighted participants in both conditions are shown in Table 1 and Figure 1.

Table 1: *Mean accuracy scores for blind and sighted participants on both conditions (standard errors in parentheses).*

	Easy condition	**Difficult condition**
Sighted	0.65 (0.09)	0.44 (0.09)

Blind	0.79 (0.12)	0.65 (0.14)

We analyzed the datasets using generalized mixed effects regression modeling for binomial distribution [24]. We constructed a model with Accuracy on the different trials as the dependent variable and fixed factors Condition (easy vs. difficult) and Group (sighted vs. blind), as well as an interaction between them. Both categorical independent variables were contrast-coded. The model included random intercepts for Participants and Items. P-values were obtained by likelihood ratio tests of the full model against the model without the effect or interaction in question. We found significant effects of Condition ($\beta = -0.91$, SE = 0.17, $\chi2(1) = 23.70$, $p < .001$), with participants being more accurate in the Easy condition ($\text{mean}_{\text{easy}} = 0.69$) than in the Difficult one ($\text{mean}_{\text{difficult}} = 0.50$). There was also a significant main effect of Group ($\beta = 0.83$, SE = 0.23, $\chi2(1) = 10.99$, $p < .001$), as blind participants performed better than their sighted counterparts in both conditions ($\text{mean}_{\text{blind}} = 0.72$ vs. $\text{mean}_{\text{sighted}} = 0.54$). The interaction between Condition and Group, though, was not significant.

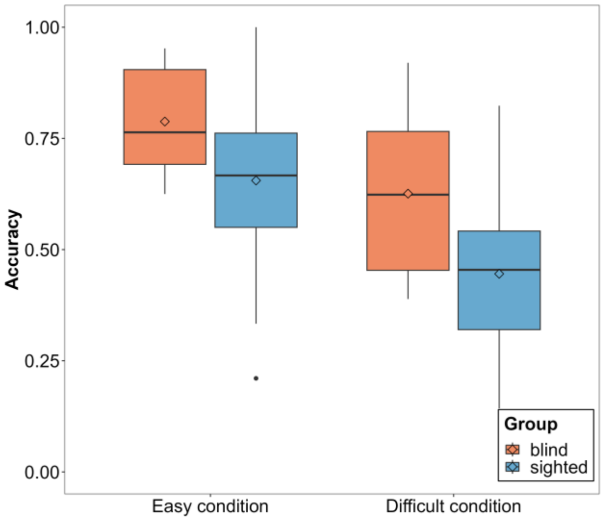

Figure 1: *Boxplots showing participants' accuracy in both conditions. The diamond-shaped marks indicate means.*

4. Discussion and conclusions

The current study investigated for the first time whether blind and sighted listeners perceive synthetic speech of different qualities in the same way. The results showed that both groups of participants are sensitive to synthetic speech quality differences, induced by using training data sets of different sizes. Specifically, both groups performed better in the Easy condition than in the Difficult one. Crucially, we found that blind participants were more accurate in both conditions than the sighted participants. In particular, although the accuracy in the Difficult condition dropped significantly in both groups, the blind group remained well above chance, while the sighted group fell below it. This suggests that the phonetic-phonological processing was stronger affected by the acoustic characteristics of synthetic speech in blind listeners than in their sighted counterparts. This evidence adds to the existing literature on the auditory processing advantage that congenitally blind listeners exhibit when tested on natural speech [12].

These findings have direct implications for the use and further development of TTS systems. First, the auditory processing advantage of the blind listeners over their sighted peers might paradoxically become a disadvantage when it comes to noticing synthetic speech quality imperfections. Specifically, blind users can be much more disturbed by these imperfections than regular users would be. Importantly, the perception of distortions, noise or other kinds of imperfections does not only reduce the perceived 'naturalness' or 'fluency' of the speech, but it can also impact the general processing of the speech and its understanding [25]. Namely, such noticeable distortions engage additional cognitive resources as the listener's brain has to engage compensatory mechanisms to efficiently process the speech signal [3]. As individuals with visual impairments often rely on TTS-based tools to manage their daily life activities, this can negatively impact their quality of life [6]. Hence, in order to answer the needs of truly diverse target users, TTS quality evaluation should include tests on congenitally blind listeners. The current study was a first attempt to test the perception at the phonetic-phonological level of processing. Further research should examine whether the observed effects also pertain to later stages of processing, such as word recognition and semantic processing. This would shed more light on whether the increased auditory sensitivity in congenitally blind listeners has a (negative) bottom-up influence on the general comprehension of the synthetic speech.

Another issue lies in the more general definition of what is the optimal quality of synthetic speech, and hence the size of data sets used to train the acoustic models. The TTS systems developed for English and several other large languages have undoubtedly reached great accuracy by using huge amounts of training data [26]. The situation is, however, very different for the reminder of the world's languages, especially, for under-resourced ones [27], [28]. The current study showed that sighted listeners seem to discriminate fewer acoustic details in the signal. In particular, their performance was below chance in the Difficult condition, suggesting that they overall did not perceive the difference between the low and average quality synthetic speech. This would suggest that a model trained with a smaller training set might generate synthetic speech of sufficient quality for sighted listeners. As the availability of training data and the cost of computing power is still an issue for many languages, further studies should continue examining the question of training set optimization according to specific user groups.

Finally, this points to the need to rethink the methods used for synthetic speech evaluation. The traditional subjective evaluation methods are not sufficiently precise to examine speech perception. As the field of psycholinguistics offers a wide range of experimental designs, these paradigms could be potentially applied to enhance synthetic speech evaluation. These behavioral methods have the advantage of being well-tested, their precision in evaluating speech perception has long been established, they are methodologically rigorous and require proper statistical analyses to interpret the collected data [17]. Such interdisciplinary methodological enhancements could not only improve the quality of TTS systems, but also ensure that diverse groups of users can fully benefit from them.

5. Acknowledgements

We would like to thank the Lithuanian Audiosensory Library for helping to recruit and test blind participants, as well as all the participants for taking part in the study.

6. References

[1] S. Ronanki, G. E. Henter, Z. Wu, and S. King, "A template-based approach for speech synthesis intonation generation using LSTMs," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-Sept, pp. 2463–2467.

[2] M. I. Lehet, K. M. Fenn, and H. C. Nusbaum, "Shaping perceptual learning of synthetic speech through feedback," *Psychon. Bull. Rev.*, vol. 27, no. 5, pp. 1043–1051, 2020.

[3] S. Winters and D. Pisoni, "Perception and comprehension of synthetic speech," *Res. Spok. Lang. Process.*, vol. 26, no. 26, pp. 95–138, 2004, [Online]. Available: http://www.cs.utoronto.ca/~gpenn/csc2518/winters-pisoni04.pdf.

[4] L. E. Humes, K. J. Nelson, and D. B. Pisoni, "Recognition of synthetic speech by hearing-impaired elderly listeners," *J. Speech Hear. Res.*, vol. 34, no. 5, pp. 1180–1184, 1991.

[5] R. KOUL, "Synthetic Speech Perception in Individuals With and Without Disabilities," *Augment. Altern. Commun.*, vol. 19, no. 1, pp. 49–58, Jan. 2003.

[6] D. Freitas and G. Kouroupetroglou, "Speech technologies for blind and low vision persons," *Technol. Disabil.*, vol. 20, no. 2, pp. 135–156, 2008.

[7] K. Papadopoulos and A. Koutsoklenis, "Reading media used by higher-education students and graduates with visual impairments in Greece," *J. Vis. Impair. Blind.*, vol. 103, no. 11, pp. 772–777, 2009.

[8] V. Delvaux, K. Huet, M. Piccaluga, and B. Harmegnies, "The perception of anticipatory labial coarticulation by blind listeners in noise: A comparison with sighted listeners in audio-only, visual-only and audiovisual conditions," *J. Phon.*, vol. 67, pp. 65–77, 2018.

[9] I. Hertrich, S. Dietrich, A. Moos, J. Trouvain, and H. Ackermann, "Enhanced speech perception capabilities in a blind listener are associated with activation of fusiform gyrus and primary visual cortex," *Neurocase*, vol. 15, no. 2, pp. 163–170, 2009.

[10] V. Occelli, C. Spence, and M. Zampini, "Auditory, tactile, and audiotactile information processing following visual deprivation.," *Psychol. Bull.*, vol. 139, no. 1, pp. 189–212, Jan. 2013.

[11] U. Schild and C. K. Friedrich, "What determines the speed of speech recognition? Evidence from congenitally blind adults," *Neuropsychologia*, vol. 112, no. October 2017, pp. 116–124, 2018.

[12] L. Arnaud, V. Gracco, and L. Ménard, "Enhanced perception of pitch changes in speech and music in early blind adults," *Neuropsychologia*, vol. 117, no. June 2017, pp. 261–270, 2018.

[13] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0," *Proc. Int. Conf. Speech Prosody*, vol. 2020-May, no. May, pp. 965–969, 2020.

[14] P. Wagner *et al.*, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*, Sep. 2019, pp. 105–110.

[15] A. Peiró-Lilja, G. Cámbara, M. Farrús, and J. Luque, "Naturalness and Intelligibility Monitoring for Text-to-Speech Evaluation," *Speech Prosody 2022*, no. May, pp. 445–449, 2022.

[16] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimed. Syst.*, vol. 22, no. 2, pp. 213–227, 2016.

[17] A. M. B. de Groot and P. Hagoort, *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*. Wiley-Blackwell, 2017.

[18] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," *9th ISCA Speech Synth. Work. SSW 2016*, pp. 202–207, 2016.

[19] S. Laurinciukaite, L. Telksnys, P. Kasparaitis, R. Kliukiene, and V. Paukštyte, "Lithuanian speech corpus liepa for development of human-computer interfaces working in voice recognition and synthesis mode," *Inform.*, vol. 29, no. 3, pp. 487–498, 2018.

[20] D. B. Pisoni and P. A. Luce, "Acoustic-phonetic representations in word recognition," *Cognition*, vol. 25, no. 1–2, pp. 21–52, Mar. 1987.

[21] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, no. 1, pp. 35–44, 1985.

[22] B. Díaz, H. Mitterer, M. Broersma, and N. Sebastián-Gallés, "Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access," *Learn. Individ. Differ.*, vol. 22, no. 6, pp. 680–689, 2012.

[23] G. A. Melnik and S. Peperkamp, "High-Variability Phonetic Training enhances second language lexical processing: evidence from online training of French learners of English," *Biling. Lang. Cogn.*, vol. 24, no. 3, pp. 497–506, May 2021.

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015.

[25] Z. Malisz, G. E. Henter, C. Valentini-botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in *Proceedings of ICPhS*, 2019, pp. 487–491.

[26] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: Disadvantages and limitations," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 234–239, 2018.

[27] G. A. Melnik-Leroy, J. Bernatavičienė, G. Korvel, G. Navickas, G. Tamulevičius, and P. Treigys, "An Overview of Lithuanian Intonation: A Linguistic and Modelling Perspective," *Informatica*, vol. 0, no. 0, pp. 1–38, Dec. 2022.

[28] J. Vít, Z. Hanzlíček, and J. Matoušek, "Czech Speech Synthesis with Generative Neural Vocoder," in *Text, Speech, and Dialogue*, Springer International Publishing, 2019, pp. 307–315.