Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games

Michael Pucher[1], *Markus Toman*[1], *Dietmar Schabus*[1], *Cassia Valentini-Botinhao*[2]
Junichi Yamagishi[2,3], *Bettina Zillinger*[4], *Erich Schmid*[5]

[1] Telecommunications Research Center Vienna (FTW), Austria
[2] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[3] National Institute of Informatics, Japan
[4] University of Applied Sciences, Wiener Neustadt, Austria
[5] Federal Institute for the Blind, Vienna, Austria

{pucher,toman,schabus}@ftw.at, {cvbotinh,jyamagis}@inf.ed.ac.uk
bettina.zillinger@fhwn.ac.at, erich.schmid@bbi.at

Abstract

In this paper we evaluate how speaker familiarity influences the engagement times and performance of blind school children when playing audio games made with different synthetic voices. We developed synthetic voices of school children, their teachers and of speakers that were unfamiliar to them and used each of these voices to create variants of two audio games: a memory game and a labyrinth game. Results show that pupils had significantly longer engagement times and better performance when playing games that used synthetic voices built with their own voices. This result was observed even though the children reported not recognising the synthetic voice as their own after the experiment was over. These findings could be used to improve the design of audio games and lecture books for blind and visually impaired children.

Index Terms: speech perception, speech synthesis, audio games, blind individuals

1. Introduction

There is an ever increasing amount of applications that require customised speech synthesis that can reflect accent, speaking style and other features, particularly in the area of assistive technology [1, 2]. Current speech technology techniques make it possible to create synthetic voices that sound considerably similar to the original speaker using only a limited amount of training data [3]. This naturally leads to research questions regarding how a listener's perception of a synthetic voice depends on the listener's acquaintance with the speaker used to train the voice. Moreover how does one perceive a synthetic voice trained on one's own speech. These questions are particularly of interest when considering the design of audio lecture material for blind children and how learning may be improved by using familiar voices. One idea we are looking to exploit is the impact of using the child's own voice or that of their teacher.

To the best of our knowledge there are no existing studies on the perception of one's own synthetic voice. Studies on the perception of one's own natural voice exist but are quite sparse and do not report on preference or intelligibility results [4–6]. There is however an extensive literature on the perception of familiar voices [7–14]. Most studies create familiarity by exposing their listeners to a certain voice, either in one or a few sessions across a certain time range [10–12]. Such studies found that for both young adults [10,11] and older adults [12] prior exposure to a talker's voice facilitates understanding. In fact it's argued that this facilitation occurs because familiarity eases the effort for speaker normalization, i.e. the mapping of an acoustic realization produced by a certain speaker to a phonetic representation [15]. Relatively few studies evaluated the impact of long-term familiarity, i.e., a voice you have been exposed to for weeks, months or years [13, 14]. Newman and Evers [13] report an experiment of pupils shadowing a teacher's voice in the presence of a competing talker. Results show that pupils that were made aware that the target voice was their teacher's outperformed pupils that were unaware of this or that were unfamiliar with that particular teacher. Souza and colleagues [14] measured the long-term familiarity impact on speech perception by selecting spouses or pairs of friends and measuring how well they understand each other in noise. They found that speech perception was better when the talker was familiar regardless of whether the listeners were consciously aware of it or not.

There are also studies on the effect of familiarity of synthetic voices using a variety of synthesisers [16]. It has been shown that increased exposure to synthetic speech improves its process in terms of reaction time [16]. There are far fewer studies on the perception of synthetic speech which is similar to a particular person's voice or that has been synthesized with a particular voice [17, 18]. [17] showed that synthetic voices that are acoustically similar to one's own voice are generally not preferred over non-similar voices. A preference was however found for voices that showed the same personality as defined by duration, frequency, frequency range, and loudness of the voice. Another study [18] showed that it is more difficult for listeners to judge whether two sentences are spoken by the same person if one of the sentences is produced by a speech synthesizer and the other is natural speech as opposed to both being synthetic speech.

It has been shown that blind individuals obtain higher intelligibility scores when compared to sighted individuals [19] and that this benefit is also observed for the intelligibility of synthetic speech [20, 21] possibly due to the familiarity effect [22] as blind individuals are exposed to the material more through the use of screen readers and audio books.

In the context of a research project together with a school for blind children we evaluated the use of different synthetic voices in audio games. Assuming that synthetic voices still

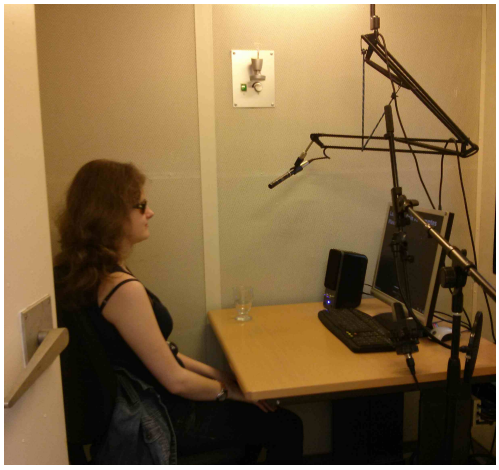

Figure 1: *Studio recordings of blind school children.*

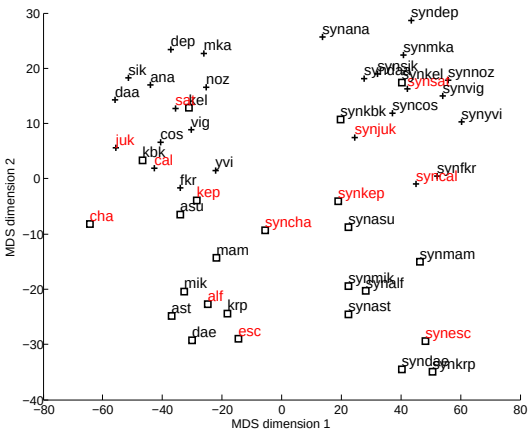

Figure 2: *Comparison between synthetic ("syn" affixed) and natural voices. School children are marked in black, teachers in red. Female speakers with crosses, male speakers with squares.*

benefit from the familiarity effect and that one's own synthetic voice is in a certain way a familiar voice, we evaluate the engagement time and game performance of a group of blind children playing audio games incorporating their own synthetic voice, their teacher's synthetic voice and an unknown synthetic voice. Using a HMM-based speech synthesis system for German we built voices of 18 school children and 7 teachers of the same school and an additional speaker who was not known to the children.

This paper is organised as follows: in Section 2, we describe the natural speech database used to train the voices and how they were created. In Section 3, we explain the design of the games, how to play them and measure their performance followed by Section 4 where we present experimental conditions and results. Finally, in Sections 5 and 6 we discuss our findings and conclude.

2. Speech databases and voices

To develop synthetic voices for the 18 children and 7 teachers of the school we recorded 200 phonetically balanced sentences for each speaker. The recordings were performed in an anechoic room with a professional microphone and recording equipment. Figure 2 shows the recording setup. For the blind children and teachers the sentences were played to the listeners via loudspeakers at a normal rate. We also recorded speech at fast and slow speaking rates from the same speakers. However these were not used in the current experiments. For the unfamiliar speaker's voice we used the same 200 sentences to develop a synthetic voice of the same quality as the children and teacher's.

When developing a synthetic voice for a speaker, we train a separate model for F0, spectrum, and duration for that speaker. These parameters are predicted for each speech unit by taking a large context into account. This leads to a more similar voice than only modifying certain speech parameters like overall duration, F0, frequency range, and loudness.

Figure 2 shows the comparison between all voices (natural and synthetic).To visualize the voices in a two-dimensional space we performed Dynamic Time Warping (DTW) between the same prompts from different speakers. For each of the 50 speakers (natural and synthetic) we had 29 different test prompts that were not used for voice training. Each prompt from a certain speaker was compared to the same prompt from all other speakers and the score was added to the respective speaker-speaker score. To obtain a similarity matrix for Multidimensional Scaling (MDS) we symmetrized the DTW scores. DTW uses the L_2 norm as distance metric.

Figure 2 shows the reduced two-dimensional space using only the two most significant dimensions. Along the horizontal axis we can see a speech type separation into a natural (left) and a synthetic (right) class. The vertical axis shows separation in terms of speaker. On this axis we can see that a certain speaker is closest to his/her respective synthetic voice. Furthermore, the y-axis shows a separation between female (crosses) and male (squares) speakers. Finally, there is no visible clustering according to age in this comparison as teachers (red font) are distributed across the space.

3. Audio games

To keep the children engaged with the experiment for a whole day, we integrated our experiments in audio-only games using speech synthesis. We developed two audio games to measure the impact of the chosen voice on game performance and engagement time.

3.1. Labyrinth game

The labyrinth game was used to measure engagement time. When starting up, instructions were presented to the player by the game voice. After the instructions, the player could choose between different labyrinth sizes: small with 7 rooms, medium with 15 rooms, large with 50 rooms and huge with 100 rooms. Keyboard cursor keys were used to navigate through the labyrinth, space bar allowed to replay the last spoken instruction, F1 presented help information to the user and F2 and F3 could be used to change the speaking rate of the game voice. The goal for the player was to find the exit of the labyrinth with as few steps as possible by remembering already visited rooms and labyrinth structure. The labyrinths were internally represented by randomly generated graphs with all nodes having a degree smaller than 4, a defined start and end point and a defined number of additionally attached dead ends. While the trees were randomly generated, the random seed used was the same for all players to ensure the experience would be the same for each player for each labyrinth size. Each node was randomly

assigned a room name (e.g., “kitchen”, “barn”) which was read to the player as well as the possible movement options (e.g., “You are now in the cockpit. Press left to go to the barn, press right to go to the kitchen.”) along the edges. Apart from the synthesized speech, non-disruptive ambient sounds were used as well as foot step sounds when moving through the labyrinth.

3.2. Memory game

The memory game was used to measure the performance of the player. As with the labyrinth game, when starting up, instructions were presented to the player by the game voice. Each round had a specific topic, e.g., musical instruments or animals. The game then constructed a non-visual, board with 8 (large: 16) fields and 4 (large: 8) items with each item associated with two fields (e.g. the item “elephant” was associated with the field belonging to keys a and j). A single key on a keyboard with German layout was associated with each field: a, s, d, f, j, k, l, ö for the normal field. For the large field, additional keys were added: q, w, e, r, u, i, o, p. Each turn consisted of the player being asked to press a key for the first field. Upon key press, the synthetic voice pronounced the item associated with the field. The player was then asked to pick a second field by pressing a key. Again upon selection, the synthetic voice pronounced the item associated with the field. If both fields were associated with the same item, the fields were removed from the current round. This was repeated until all duplicate items were found and all fields removed. Apart from the synthesized speech giving feedback on the player choices, sound effects were used for success or failure or pressing an invalid or already selected/removed key. At the end of each round, the player was told how many guesses he/she had needed to clear the board.

4. Experiments

For the experiments, 27 children played the two audio-only games. The children were grouped into 3 groups, where one group listened to their own synthetic voices in the games, one group listened to the teacher’s voices, and one group heard an unknown synthetic voice. For the children listening to the teacher’s voice we made sure that they knew the teacher very well from the classroom. Availability of a voice model, age (see Figure 5), gender and degree of visual impairment were the factors used to balance the groups. Note that it is, however, impossible to perfectly balance all the factors because of the limited number of blind children and their additional disabilities and hence we have used the three most balanced groups that we could define (see Figure 3).

The experiment was conducted in two computer rooms in school with the groups evenly split between the rooms. The games were deployed to the computers so that each child got a personalized version. They assumed that all of them were playing the same version of the game.

Figure 3 shows the descriptions of the users that participated in the experiment. We had 27 school children that participated in the evaluation. The users of speech synthesis and Braille displays were identical to the blind participants. Speakers were familiar with speech synthesis but not with HMM-based speech synthesis. We had slightly more female and blind participants in the first group.

Figure 4 shows the number of years blind users have been using speech synthesis technology and Braille displays. We can see that the blind children start to use Braille displays much earlier than speech synthesis.

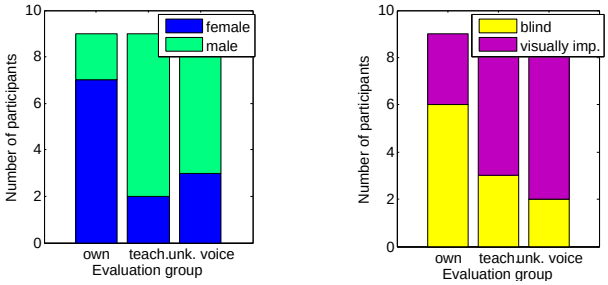

Figure 3: *Participants characteristics within groups.*

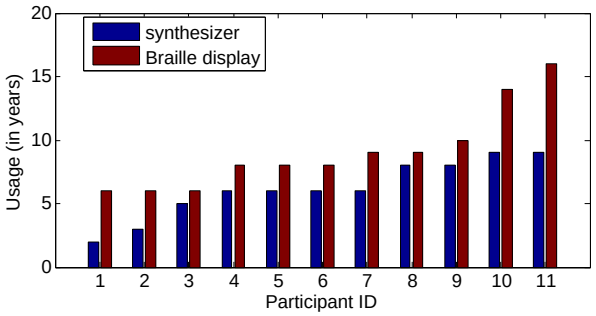

Figure 4: *Speech synthesis usage (blue bars) and Braille display usage (red bars) in years for the 11 blind participants.*

4.1. Labyrinth game

To measure engagement in the labyrinth game we used the time played overall and the number of games that were played. Children could choose how many games they wanted to play, and they could also choose the labyrinth size. The labyrinth game has a goal, namely finding the exit of the labyrinth, but it can also be played in an exploratory style where the players explore the rooms of the labyrinth.

Figure 6 (left) shows that participants hearing their own synthetic voice played significantly longer than users listening to an unknown synthetic voice ($p < 0.05$) according to a Wilcoxon rank sum test for equal medians. Differences between the teacher’s voice and unknown as well as own voices were not significant. The same trends are seen for groups with blind-only participants as shown in Figure 6 (right), but they are not significant. We did not find any significant gender differences for the labyrinth game.

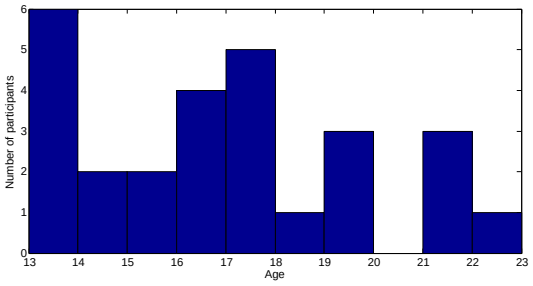

Figure 5: *Participants age distribution.*

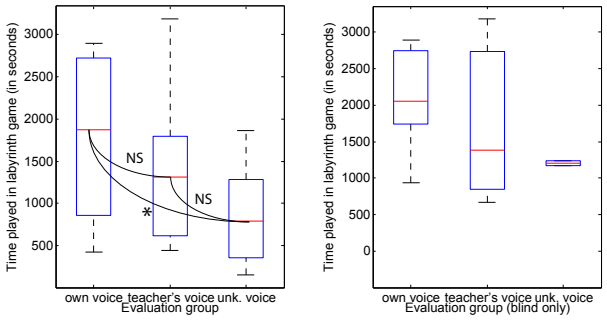

Figure 6: *Time played per group in the labyrinth game for all participants (left) and blind-only participants (right).*

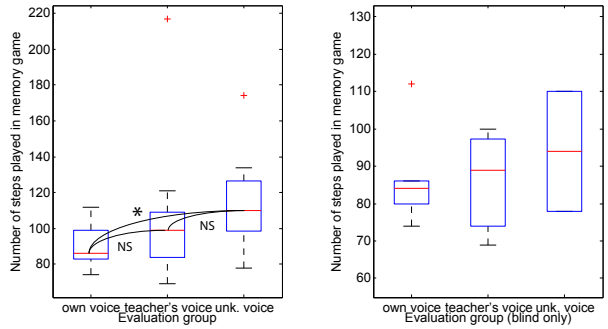

Figure 7: *Number of steps per group in the memory game for all participants (left) and blind-only participants (right).*

4.2. Memory game

In the experiments with the memory games the children had to play 8 mandatory rounds. As the conditions were the same for all children in this case, the first 6 rounds were on a normal game board, the next 2 on a large board. All children had the same topics for each round and the same assignments of items to fields. After playing the 8 rounds they could continue playing as long as they liked and freely choose the board size. To analyse the performance we only considered the 8 mandatory rounds. We used the number of steps needed to solve all 8 rounds as performance variable.

Figure 7 (left) shows that the children needed significantly less steps ($p < 0.05$) for finishing the memory game when using their own synthetic voice compared to an unknown synthetic voice. Differences between the teacher's voice and unknown as well as own voices were not significant. Again we can see the same trends also for groups with blind-only participants, but they are not significant. No significant gender differences were found for the memory game.

4.3. Blind vs. visually impaired users

As Figure 8 shows, blind participants played significantly longer ($p < 0.05$) than visually impaired participants. This is true for the labyrinth as well as for the memory game. The stronger engagement of blind users in playing is also true for other performance variables. We think that blind users are more sensitive to the auditive modality and can thereby gain more pleasure in playing audio-only games.

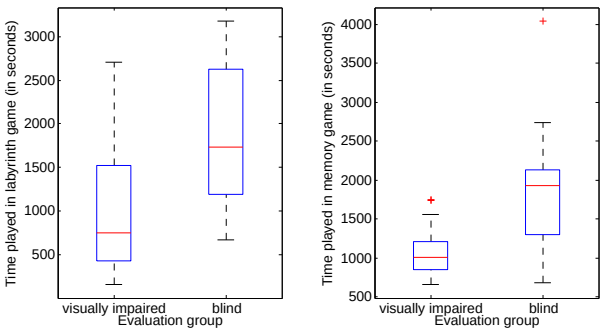

Figure 8: *Time spent playing the labyrinth game (left) and memory game (right) for blind vs. visually impaired players.*

5. Discussion

Our results show that the use of one's own voice increases the engagement time in audio games, which indicates a certain preference. To align our results with the results in [17] one's own voice can also be considered as the extreme case of a voice from a speaker with the same personality as oneself. Results for listeners of teacher's voices, although not significant, show a trend that reflects the special role of familiarity when a voice of a speaker to which the listener has a special social relation (teacher) is concerned.

The children in our study prefer known voices although they did not recognise the speakers (neither themselves nor the teachers). This indicates a certain type of cognitive processing where speech recognition and speaker recognition are independent but features of familiar speakers can be used in the recognition process. This ease of recognition of familiar speakers could be one explanation for the longer engagement times.

6. Conclusion

In this paper, we have shown that listening to one's own synthetic voice increases engagement and performance of blind school children in audio games significantly. For the evaluation we developed an audio-only labyrinth game to measure engagement time and a memory game to measure performance. Familiar voices like teacher's voices show a trend of increased engagement and performance, but more experiments are needed for verifying this hypothesis.

We also showed that blind listeners engage longer with the audio games than visually impaired listeners. We hypothesize that blind listeners are more accustomed to listening to synthetic speech and it is easier for them to process synthetic speech.

For blind users that are using speech synthesis on a regular basis there is a need to make their synthesizer experience more engaging and pleasurable, which can be accomplished by using their own or familiar voice in the synthesizer.

7. Acknowledgement

This work was supported by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB) and by the Austrian Science Fund (FWF): P23821-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWFJ, and the City of Vienna. The COMET program is managed by the FFG.

8. References

[1] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, 2010.

[2] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[3] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[4] C. Fernyhough and J. Russell, "Distinguishing one's own voice from those of others: A function for private speech?" *International Journal of Behavioral Development*, vol. 20, no. 4, pp. 651–665, 1997.

[5] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, 2002.

[6] C. Rosa, M. Lassonde, C. Pinard, J. P. Keenan, and P. Belin, "Investigations of hemispheric specialization of self-voice recognition," *Brain and cognition*, vol. 68, no. 2, pp. 204–214, 2008.

[7] D. Van Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters. part i: Recognition of backward voices." *Journal of phonetics*, vol. 13, pp. 19–38, 1985.

[8] D. V. Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[9] T. Bőhm and S. Shattuck-Hufnagel, "Utterance-final glottalization as a cue for familiar speaker recognition," in *Proc. Interspeech, Antwerp*, 2007, pp. 2657–2660.

[10] L. C. Nygaard, M. S. Sommers, and D. B. Pisoni, "Speech perception as a talker-contingent process," *Psychological Science*, vol. 5, no. 1, pp. 42–46, 1994.

[11] L. C. Nygaard and D. B. Pisoni, "Talker-specific learning in speech perception," *Perception & psychophysics*, vol. 60, no. 3, pp. 355–376, 1998.

[12] C. A. Yonan and M. S. Sommers, "The effects of talker familiarity on spoken word identification in younger and older listeners." *Psychology and aging*, vol. 15, no. 1, p. 88, 2000.

[13] R. S. Newman and S. Evers, "The effect of talker familiarity on stream segregation," *Journal of Phonetics*, vol. 35, no. 1, pp. 85 – 103, 2007.

[14] P. Souza, N. Gehani, R. Wright, and D. McCloy, "The advantage of knowing the talker," *Journal of the American Academy of Audiology*, vol. 24, no. 8, p. 689, 2013.

[15] D. Pisoni and R. Remez, *The Handbook of Speech Perception*. John Wiley & Sons, 2008.

[16] M. Reynolds, C. Isaacs-Duvall, B. Sheward, and M. Rotter, "Examination of the effects of listening practice on synthesized speech comprehension," *Augmentative and Alternative Communication*, vol. 16, no. 4, pp. 250–259, 2000.

[17] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

[18] M. Wester and R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5372–5375.

[19] K. Hugdahl, M. Ek, F. Takio, T. Rintee, J. Tuomainen, C. Haarala, and H. Hmlinen, "Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds," *Cognitive Brain Research*, vol. 19, no. 1, pp. 28 – 32, 2004.

[20] K. Papadopoulos, V. S. Argyropoulos, and G. Kouroupetroglou, "Discrimination and comprehension of synthetic speech by students with visual impairments: The case of similar acoustic patterns," *Journal of Visual Impairment & Blindness*, vol. 102, no. 7, pp. 420–429, 2008.

[21] M. Pucher, D. Schabus, and J. Yamagishi, "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners," in *Proc. Interspeech*, Chiba, Japan, Sept. 2010, pp. 2186–2189.

[22] M. Barouti, K. Papadopoulos, and G. Kouroupetroglou, "Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate," in *European AAATE Conference*, Portugal, 2013, pp. 695–699.