Influence of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices

Michael Pucher[a,*], Bettina Zillinger[f], Markus Toman[b], Dietmar Schabus[c], Cassia Valentini-Botinhao[d], Junichi Yamagishi[d,e], Erich Schmid[g], Thomas Woltron[f]

[a]*Acoustics Research Institute (ARI), Austrian Academy of Sciences (OAW), Austria*
[b]*Vienna University of Technology (TUW), Austria*
[c]*Austrian Research Institute for Artificial Intelligence (OFAI)*
[d]*The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK*
[e]*National Institute of Informatics (NII), Japan*
[f]*University of Applied Sciences, Wiener Neustadt, Austria*
[g]*Federal Institute for the Blind, Vienna, Austria*

Abstract

In this paper we evaluate how speaker familiarity influences the engagement times and performance of blind children and young adults when playing audio games made with different synthetic voices. We also show how speaker familiarity influences speaker and synthetic speech recognition. For the first experiment we develop synthetic voices of school children, their teachers and of speakers that are unfamiliar to them and use each of these voices to create variants of two audio games: a memory game and a labyrinth game. Results show that pupils have significantly longer engagement times and better performance when playing games that use synthetic voices built with their own voices. These findings can be used to improve the design of audio games and lecture books for blind and visually impaired children and young adults. In the second experiment we show that blind children and young adults are better in recognising synthetic voices than their visually impaired companions. We also show that the average familiarity with a speaker and the similarity between a speaker's synthetic and natural voice are correlated to the speaker's synthetic voice recognition rate.

Keywords: speech perception, speech synthesis, audio games, blind individuals, child speech synthesis

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 (Pucher et al., 2010b; Yamagishi et al., 2012). 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 (Yamagishi and Kobayashi, 2007). This naturally leads to our research questions:

- How does a listener's perception of a synthetic voice depend on the listener's acquaintance with the speaker used to train the voice?
- How does a listener 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 young adults 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 her/his teacher [1].

To the best of our knowledge there are no existing studies on the perception of one's own synthetic voice. Synthetic voices of language learners have however been prosodically manipulated to adapt to a native model speaker in computer-assisted pronunciation training (Bissiri and Pfitzinger, 2009; Bonneau and Colotte, 2011).

Studies on the perception of one's own natural voice exist but are quite sparse and do not report on preference or intelligibility results (Fernyhough and Russell, 1997; Appel and Beerends, 2002; Rosa et al., 2008). Fernyhough and Russell (1997) investigates how children's private speech allows them to learn to distinguish between their own and other's voices. Appel and Beerends (2002) investigates the perception of one's own voice in a telephone setup where echo and distortion is introduced. Rosa et al. (2008) shows that there is a certain right-hemisphere advantage for self-compared to other-voice recognition similar to what was observed for self-face recognition. It is known, that

*Corresponding author
Email addresses: michael.pucher@oeaw.ac.at (Michael Pucher), bettina.zillinger@fhwn.ac.at (Bettina Zillinger), m.toman@neuratec.com (Markus Toman), dietmar.schabus@ofai.at (Dietmar Schabus), cvbotinh@inf.ed.ac.uk (Cassia Valentini-Botinhao), jyamagis@inf.ed.ac.uk (Junichi Yamagishi), erich.schmid@bbi.at (Erich Schmid), thomas.woltron@fhwn.ac.at (Thomas Woltron)

[1] Parts of the contents of this paper have been published in Pucher et al. (2015).

the so-called talker (own voice) and listener (ambient sounds) sidetone plays an important role in telephony if we want to achieve a natural phone conversation, since we normally also hear ourselves over the air channel (ITU-T, 1993; ETSI, 1996). The so-called *talker sidetone loss must lie within certain limits for a comfortable talking situation* (ETSI, 1996). If the loudness of the sidetone is however passing a certain threshold it is also a strange and annoying experience for the talker/listener. The part of our own voice that we hear over the bone channel is not necessary to model for telephony applications since it is produced during the conversation, but would need to be modelled for own voice synthesis. An estimation of the different components of air and bone-conducted sound was done by Pörschmann (2000). The use of a synthetic voice also allows us to modify all kinds of parameters like F0, duration, linguistic, and spectral parameters. This shows that there are several interesting open research questions concerning the perception of one's own natural and/or synthetic voice. With our study on the perception of one's own synthetic voice we aim to make a first step into this direction that also investigates preference and intelligibility.

There is however an extensive literature on the perception of familiar voices (Van Lancker et al., 1985; Lancker and Kreiman, 1987; Bőhm and Shattuck-Hufnagel, 2007; Nygaard et al., 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000; Newman and Evers, 2007; Souza et al., 2013). Most studies create familiarity by exposing their listeners to a certain voice, either in one or a few sessions across a certain time range (Nygaard et al., 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000). Such studies found that for both young adults (Nygaard et al., 1994; Nygaard and Pisoni, 1998) and older adults (Yonan and Sommers, 2000) prior exposure to a talker's voice facilitates understanding. In fact it is argued that this facilitation occurs because familiarity eases the effort for speaker normalisation, i.e. the mapping of an acoustic realisation produced by a certain speaker to a phonetic representation (Pisoni and Remez, 2008). Relatively few studies evaluated the impact of long-term familiarity, i.e., a voice you have been exposed to for weeks, months or years (Newman and Evers, 2007; Souza et al., 2013). Newman and Evers (Newman and Evers, 2007) 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 (Souza et al., 2013) 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 (Reynolds et al., 2000). It has been shown that increased exposure to synthetic speech improves its process in terms of reaction time (Reynolds et al., 2000). There are far fewer studies on the perception of synthetic speech which is similar to a particular person's voice or that has been synthesised with a particular voice (Nass and Lee, 2001; Wester and Karhila, 2011). (Nass and Lee, 2001) 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 (Wester and Karhila, 2011) 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 synthesiser based on the same voice, 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 (Hugdahl et al., 2004) and that this benefit is also observed for the intelligibility of synthetic speech (Papadopoulos et al., 2008; Pucher et al., 2010a) possibly due to the familiarity effect (Barouti et al., 2013) as blind individuals are exposed to the material more through the use of screen readers and audio books. It was also shown repeatedly that blind individuals show a much higher intelligibility for fast synthesised speech (Moos and Trouvain, 2007; Pucher et al., 2010a), an effect that can be found for a wide range of synthesisers: formant, diphone, unit selection and Hidden Markov Model (HMM) based (Syrdal et al., 2012). Bull et al. (1983) showed that blind individuals have a higher voice recognition accuracy than non-blind listeners, but they could not find a difference in voice recognition for different degrees of blindness by using natural speech samples.

Using an HMM-based text-to-speech synthesis system for Austrian German we built synthetic voices of 18 blind and visually impaired school children and seven teachers of the same school and an additional speaker who was not known to the children and young adults. The school is located in Vienna and is visited by children and young adults from elementary to high school and vocational school levels. We report in this article two separate experiments performed using those synthetic voices. In the first experiment we measured the engagement time and performance of a group of blind children and young adults playing audio games constructed with their own synthetic voice, their teacher's synthetic voice and the unknown synthetic voice. This experiment has been published in (Pucher et al., 2015). In the second experiment we measured the speaker recognition rates of the synthetic voices, the familiarity of the speaker, the similarity between synthetic and natural speech and the intelligibility of the synthetic voices.

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 and the results obtained in this experiment, followed by Section 4 where we present the experimental conditions and results of the second experiment. Finally, in Sections 5 and 6 we discuss our findings and conclude.

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

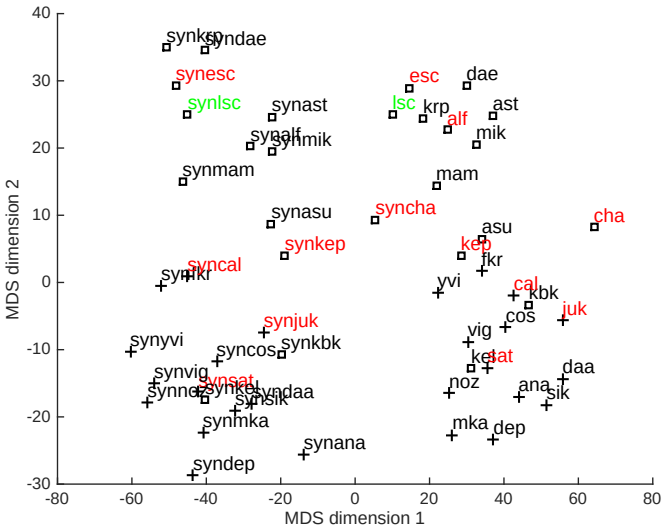

Figure 2: *Comparison between synthetic ("syn" affixed) and natural voices. School children are marked in black, teachers in red, the unfamiliar speaker in green. Female speakers with crosses, male speakers with squares.*

2. Speech databases and voices

In this section we present the databases used to train the synthetic voices used in the experiments reported in this paper, the technical details on how we built the voices and a visualisation of voice distances.

2.1. Speech databases

We recorded 223 phonetically balanced sentences spoken by 18 children and young adults and seven teachers [2]. The recordings were performed in an anechoic room with a professional microphone and recording equipment. Figure 1 shows the recording setup: the sentences were played to the listeners via loudspeakers at a normal rate and they had to repeat what they heard.

To built a synthetic voice of an unfamiliar speaker's voice we used the same 223 sentences recorded by a speaker of Regional Standard Austrian German (RSAG). The speaker was selected for one of our previous projects to record audio-visual speech data. An RSAG speaker was selected instead of a standard speaker to have a higher similarity with the other recorded speakers (school children and teachers) in terms of language variety. The unfamiliar speaker was male and 47 years old.

Ideally we would use a large corpus of 3.000 sentences for training, which was however not possible when developing voices for 25 school children and teachers. So we decided to use a small phonetically balanced recording script of 223 sentences for training. Since our experiments were on the influence of one's own voice, we used a speaker dependent training procedure instead of a speaker adapted one such that only data from the target speaker is used for training the voice.

2.2. Synthetic voices

Speaker-dependent synthetic voices were created using the training scripts provided with the HMM-based Speech Synthesis System (HTS) [3], which were slightly adapted for the Austrian German data set. For synthesis, the Speech Synthesis of Auditory Lecture Books for blind children (SALB) framework (Toman and Pucher, 2015) [4] was integrated into the audio games, using hts_engine [5] for waveform generation and an internal text analysis module for Austrian German. After the experiments, the participants were given packages containing the SALB framework together with their own voice to be installed as Microsoft Windows system voices.

When developing a synthetic voice for a speaker, we used 5-state left-to-right Hidden Semi-Markov Models (HSMM) as acoustic models. Models were trained from context-dependent data clusters using decision tree based clustering (Odell, 1995). We used explicit duration modelling (Levinson, 1986) instead of self-transitions. We trained separate models for fundamental frequency (F0), spectrum and duration for each speaker. To train these acoustic models, we extracted 40 Mel cepstral coefficients (Fukada et al., 1992) and F0 from the natural speech sampled at 44.1 kHz, using a frame length of 25 ms and frame shift of 5 ms. The observation vectors contained static, delta and delta-delta values. One stream was used for the spectrum and three for F0. The lexicon used by the front-end was from our released Austrian German open-source voice (Toman and Pucher, 2015) [6]. At generation time these parameters were again predicted from the context-dependent models at every 5 ms. The whole voice creation process for HMM-based synthesis is described in more detail in Zen et al. (2009) and Tokuda et al. (2013).

To have an image of how each synthetic voice and each natural voice differ from each other we performed Multidimensional Scaling (MDS) over a distance matrix built using 29

[2] We use the terms "children and young adults" and "school children" to refer to our participants since they all visit the same school.

[3] HTS: `http://hts.sp.nitech.ac.jp/`

[4] `http://m-toman.github.io/SALB/`

[5] hts_engine: `http://hts-engine.sourceforge.net/`

[6] `https://sourceforge.net/projects/at-festival/`

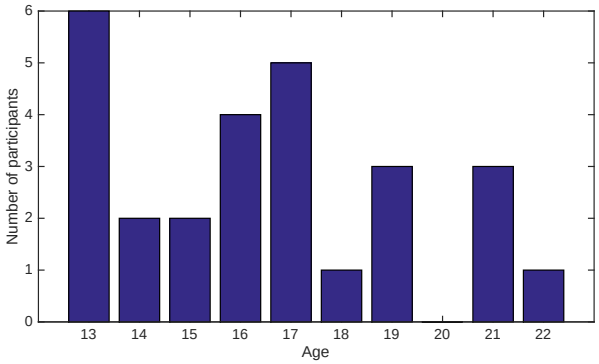

Figure 3: *Participants age distribution.*

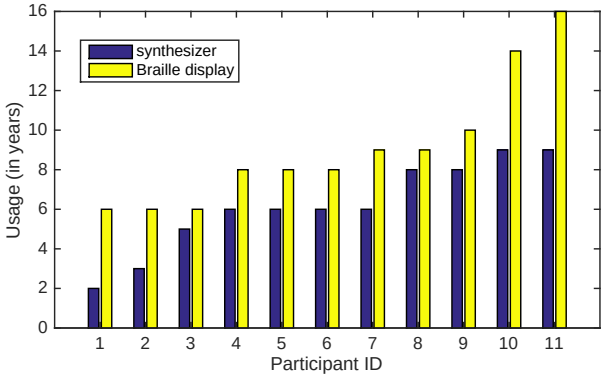

Figure 4: *Speech synthesis usage (blue bars) and Braille display usage (yellow bars) in years for the 11 blind participants.*

test sentences. With MDS we can project a distance matrix to a lower dimensional space, which then shows the dimensions that are most important for the distance. To calculate the distance between two voices we performed Dynamic Time Warping (DTW) between the acoustic features extracted from the same prompts recorded or generated by each voice. Each prompt of 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 the distance matrix we symmetrised the DTW scores. DTW was calculated using the L_2 norm as distance metric.

Figure 2 shows the reduced two-dimensional space using only the two most significant dimensions obtained after MDS. Along the horizontal axis we can see a separation into natural (left) and synthetic (right) voices. The vertical axis shows separation in terms of speaker. Interestingly we can see that a certain speaker is often 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 game experiments

In this section we describe the audio game experiment design, the design of the audio games, and the evaluation results concerning engagement time and performance.

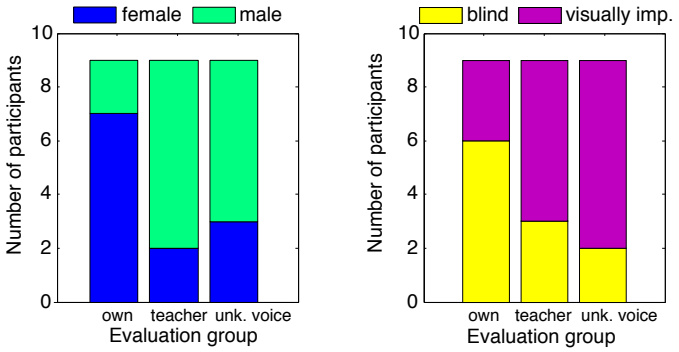

Figure 5: *Participants characteristics within groups.*

3.1. Experiment design

We recruited 27 children and young adults from the same school where we performed the recordings for the audio game experiments. Their age distribution is presented in Figure 3. Since there are many school levels at the blind school we have children from 13 to 18 years and also a few young adults from 19 to 22 years. The mean age of the children and young adults was 16.48 years. 16 of them are blind while the other 11 children are visually impaired. Figure 4 shows the number of years each of the blind children and young adults have been using speech synthesis technology and Braille displays. We can see that they start to use Braille displays much earlier than speech synthesis. Children and young adults were familiar with speech synthesis technology but not with HMM-based speech synthesis technology.

The participants of the perception test were organised into 3 groups. The children and young adults in each group performed the same tasks but using games constructed with different synthetic voices. One group played with audio games constructed with their own synthetic voices, one group listened to the teacher's voices, and one group listened to an unknown synthetic voice. For the children and young adults listening to the teacher's voice we made sure that they knew the teacher very well from the classroom. Availability of a synthetic voice, age, 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 due to the limited number of blind school children and their additional disabilities. We have then used the three most balanced groups that we could define, see Figure 5.

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 personalised version. They assumed that all of them were playing the same version of the game.

3.2. Audio games design

We developed two audio games for this experiment: the labyrinth game and the memory game. The labyrinth game was created to measure engagement time while the memory game allowed us to measure performance.

When starting the labyrinth game, instructions are presented to the player by the game voice. After the instructions, the

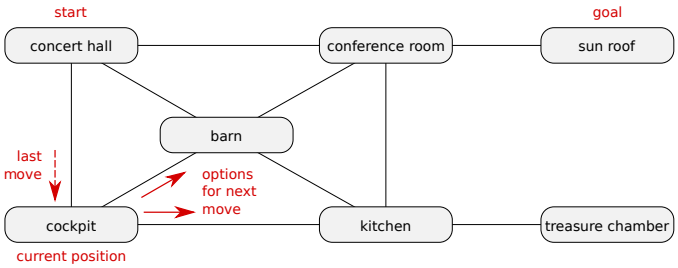

Figure 6: *Illustration of the labyrinth game with a small labyrinth (seven rooms). The player has just moved from the "concert hall" to the "cockpit". The next possible moves are going to the "barn" or to the "kitchen", or to go back to the "concert hall". The game is won when the player enters the goal room "sun roof". The "treasure chamber" is a dead end. The participants were informed that there is a final room that ends the walk through the labyrinth, but they were allowed to play the game as long as they liked.*

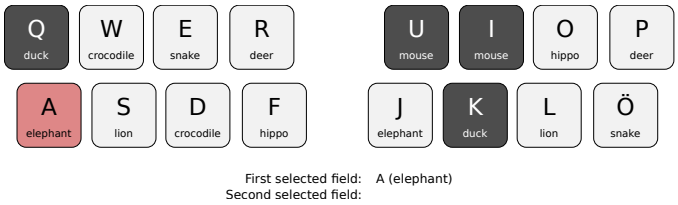

Figure 7: *Illustration of the memory game with a large field (16 keys). The two pairs (Q, K) – "duck" and (U, I) – "mouse" have already been found and the corresponding keys cannot be selected anymore. The player has just selected A ("elephant") as the first field and is trying to remember the position of the second "elephant" (J).*

player can choose between different labyrinth sizes: small (seven rooms), medium (15 rooms), large (50 rooms) and huge (100 rooms). The goal of the game is to find the exit of the labyrinth with as few steps as possible by remembering already visited rooms and the labyrinth structure. Only when entering each room the room name (e.g., "kitchen", "barn") is 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."). Apart from the synthesised speech, non-disruptive ambient sounds were used as well as foot step sounds when moving through the labyrinth. Figure 6 illustrates the labyrinth game.

Keyboard cursor keys were used to navigate through the labyrinth, the space bar replayed 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 labyrinths were internally represented by randomly generated graphs with all nodes having a degree of four or less, a defined start and end point and a defined number of additionally attached dead ends. While the graphs were randomly generated, the same random seed was used for all players to ensure the experience would be the same for each player for each labyrinth size.

As with the labyrinth game, when starting the memory game instructions are 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

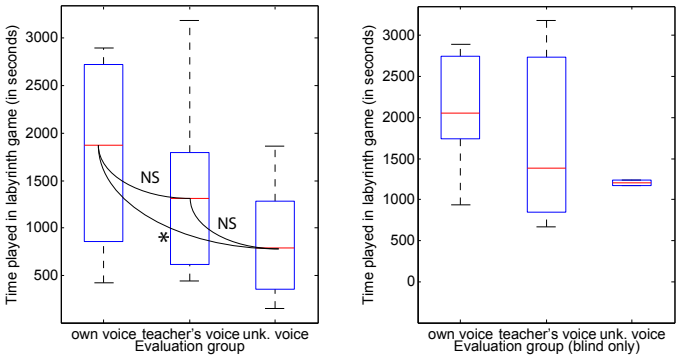

Figure 8: *Engagement measure: time spent playing the labyrinth game calculated per group for all participants (left) and blind-only participants (right).*

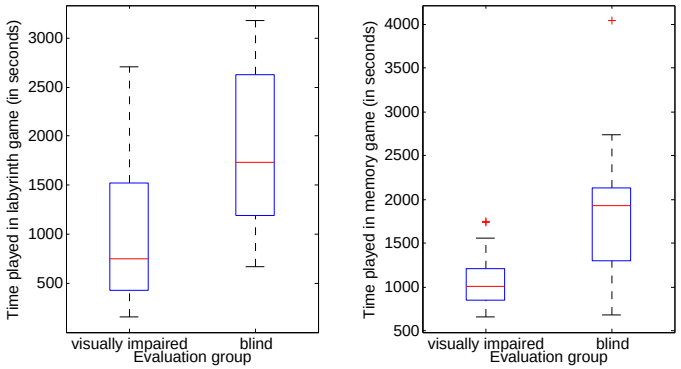

Figure 9: *Engagement measure: time spent playing the labyrinth game (left) and memory game (right) for blind vs. visually impaired players.*

fields and 4 items (16 and 8 for the large option). Each item is associated with two fields (e.g. the item "elephant" is 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 synthesised 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. Figure 7 illustrates the memory game. The games were produced for the Windows operating system and the version that uses the synthetic voice built from the voice talent voice is available for download here: `https://github.com/m-toman/Audio-Games/`.

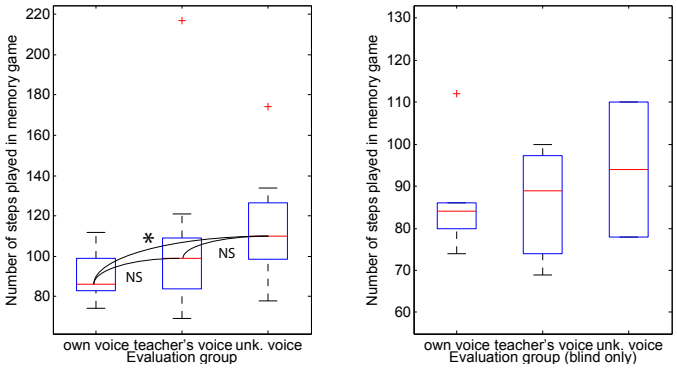

Figure 10: *Performance measure: number of steps taken in the memory game calculated per group for all participants (left) and blind-only participants (right).*

3.3. Results

3.3.1. Engagement

To measure engagement in the labyrinth game we used the time played overall and the number of games that were played. School 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 8 shows the time spent playing the labyrinth game, measured per evaluation group for all participants (left) and for the blind participants only (right). We can see that the participants that were using 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 8 (right), but they are not significant. We did not find any significant gender differences for the labyrinth game. The time that blind users using an unknown voice (Figure 8 rightmost bar) played in the labyrinth does not show much variance, since this group only consisted of 2 participants as shown in Figure 5. The time difference between them was only 69 seconds.

We have also measured engagement time when playing the memory game. Figure 9 presents the time spent playing the labyrinth game (left) and the memory game (right) of visually impaired and blind children and young adults. We found that 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.

3.3.2. Performance

In the experiments with the memory game the children and young adults had to play 8 mandatory rounds. As the conditions were the same for all participants in this case, the first 6 rounds were on a normal game board, the next 2 on a large board. All participants 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 10 (left) shows that the children and young adults needed significantly fewer 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. Recognition experiments

The results in the previous experiment show that the use of one's own voice increases the engagement time in audio games, which indicates a certain preference. We were then interested to see whether this increase was due to a facilitation in understanding a familiar voice and to measure more precisely how familiar the voices truly were for those children and young adults.

In order to do so we performed a second experiment where we measured word/sentence recognition error rate and the speaker identification error rate of the synthetic voices, as well as self-reported familiarity with the person whose voice we built synthetic voices.

4.1. Experiment design

For this experiment we recruited 30 children and young adults from the same school. There were 16 male and 14 female subjects, 11 blind and 19 visually impaired, aged between 14 and 24 years. 9 participants out of these 30 were also recorded previously to built synthetic voices and were also part of the first experiment.

All participants performed the same tasks using the same material. There were four tasks:

- a voice recognition task,
- a familiarity task,
- a similarity task
- and an intelligibility task.

In the first task participants had to identify the speaker of a range of synthetic voices based on one synthetic speech sample. They were not given a list of possible speakers. They knew however that we have built voices of teachers and school children, so they knew that the list of possible speakers was constrained.

After the speaker recognition task, participants were asked if they knew the speakers, and had to give their acquaintance a value between 1 and 5, 1 meaning good acquaintance and 5 means that they don't know the person at all.

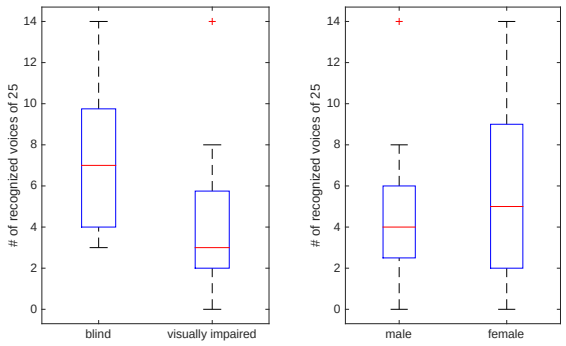

Figure 11: *Voice recognition for blind/visually impaired and male/female listeners.*

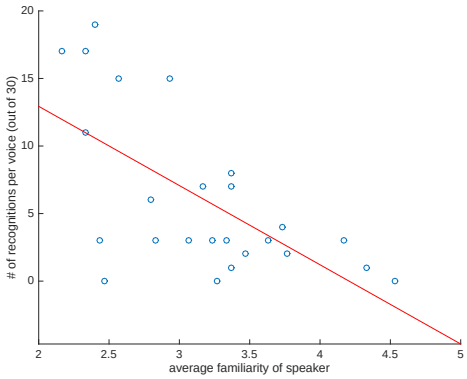

Figure 12: *Scatter plot of speaker familiarity against # of recognitions per voice (out of 30) (Pearson correlation coefficient R* -0.63, $p < 0.005$*; Maximal Information Coefficient (MIC)* 0.36*; MIC-*R^2 -0.03*).*

The third task participants undertook was to judge the similarity of synthesised and natural voices. Voice similarity was again given a value between 1 to 5 from very similar to very different. Two speech samples were played to the listeners, one speaker's natural voice and one synthesised voice that was trained from the same speaker, and they had to rate the speaker similarity of the samples on a scale from 1 to 5.

In the word recognition experiments the word error rate was evaluated by using 25 sentences where each sentence was associated to one of the 25 synthetic voices. Each listener then had to listen to each sentence once and had to transcribe it. Then the word-error-rate was computed as it is done in speech recognition as the minimum number of substitutions, deletions, and insertions that are necessary for transforming between transcription and correct transcript. Before computing the word-error-rate we corrected orthographic errors and potential typos. The same procedure was used for all participants, which could introduce age related differences due to different writing skills.

4.2. Results

4.2.1. Speaker recognition, familiarity and similarity

Figure 11 shows the number of voices that were recognised by visually impaired and blind participants (left) and by male and female participants (right). Overall the speaker recognition rate was quite low. Only 25% of the speakers that were

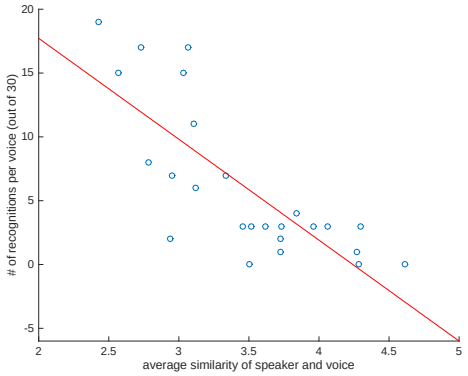

Figure 13: *Scatter plot of speaker similarity against # of recognitions per voice (out of 30) (Pearson correlation coefficient R* -0.78, $p < 0.005$*; MIC* 0.71*; MIC-*R^2 -0.10*).*

known to a listener could be identified by their synthetic voices. Blind participants performed significantly better in voice recognition ($p < 0.05$) than visually impaired participants. Concerning gender there are no significant differences in voice recognition. 9 out of the 30 participants had their synthetic voices in the experiment so were in fact tasked with identifying their own voice. This number is quite low although we have 25 different synthetic voices because only 9 participants were still at the school when the experiments were made due to the time difference between the recordings, the voice development and the experiments. From these nine school children, five were able to identify their own voices (55%) whereas overall only 25% = 153/601 were able to identify speakers that are known by them correctly. The children and young adults that listened to their own voices had however the advantage that they also had heard their synthetic voices before in the first experiment, and were also able to use their own synthesisers on their computers at home.

The general familiarity between listeners and speakers is high, which is natural for a school context, 84% of blind and 73% of visually impaired listeners know the speakers that were used to develop synthetic voices. The acquaintance with a speaker of a synthetic voice was judged on a 1 to 5 scale with 1 meaning that the speaker is known very well, and 5 meaning that the speaker is not known at all. We defined that a listener knew a speaker if the familiarity rating was ≤ 4.

Figure 12 shows the scatter plot of number of times a synthetic voice was correctly identified against speaker familiarity associated with that voice and averaged across all participants. Not surprisingly, familiar voices were recognised more often than the unfamiliar ones. There is a negative correlation between average familiarity and number of recognitions with a Pearson correlation coefficient R of -0.63 ($p < 0.005$) and a Maximal Information Coefficient (MIC) of 0.36. The MIC is a value between $[0, 1]$ and can be interpreted as a correlation measure with 1 meaning perfect and 0 no correlation (Reshef et al., 2011). The difference between MIC and R^2 shows if there is a linear or non-linear correlation within the data, since MIC gives a high value also to non-linear correlations. The low

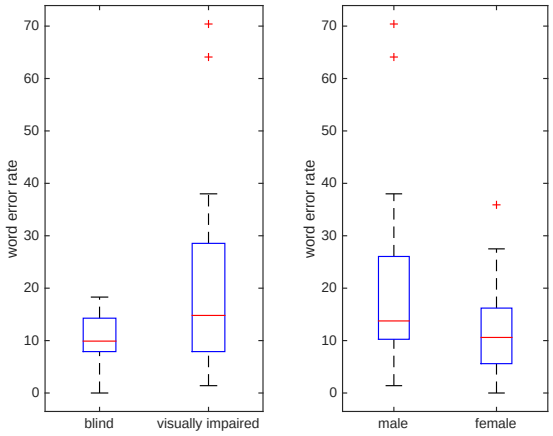

Figure 14: *Word error rates for blind vs. visually impaired (left) and male vs. female (right) listeners.*

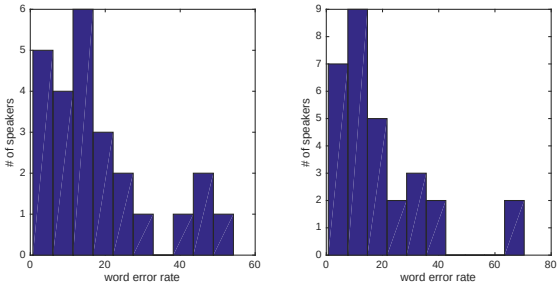

Figure 15: *Word error rates distribution per voices (left) and listeners (right).*

value of MIC-R^2 = −0.03 in this case shows that there is a linear correlation within the data.

There were however two speakers with a high average familiarity of ≈ 2.5 and ≈ 3.25 that were never recognised from their synthetic voices. This is related to quality issues with the synthetic voices, which enable us to retain speaker similarity of synthetic voices better for some speakers than others.

Figure 13 presents a scatter plot of number of times a synthetic voice was correctly identified against the similarity score of the voice averaged across all participants. As shown in the figure, synthetic voices that where judged to be similar to the real voices were most often correctly identified, which is exactly what one would expect. There is a negative correlation between average similarity and number of recognitions with a Pearson coefficient R of −0.78 ($p < 0.005$) and a MIC of 0.71. The difference MIC-R^2 of −0.10 also shows a linear correlation for this case.

4.2.2. Intelligibility

Figure 14 shows the word error rate obtained by blind and visually impaired participants (left) and male and female participants (right). We can see that blind individuals where slightly but not significantly better in terms of word recognition compared to visually impaired school children. There were also no significant differences for word error rates in terms of gender.

Figure 15 presents the distribution of word error rate per synthetic voice and listeners. We can observe some outliers: two

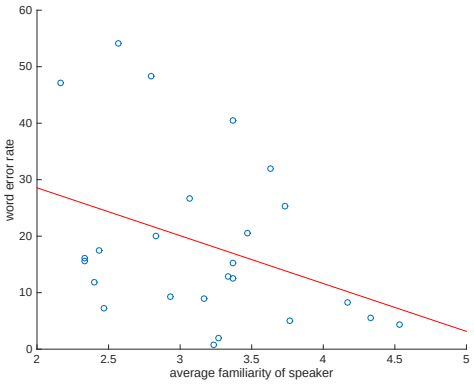

Figure 16: *Scatter plot of speaker familiarity against word error rate (Pearson correlation coefficient R* −0.36*; MIC* 0.32*; MIC-R^2* 0.18*).*

listeners with a word error rate around 70% and four synthetic voices had a word error rate above 40%. Interestingly these two distributions look quite similar, although they are generated by completely different processes. Word-error-rate per voice (left) indicates that we had some synthetic voices with quality issues or prompts that were difficult to understand. Word-error-rate per listener (right) indicates that there were some listeners that were particularly bad at understanding the synthetic voices.

The word error rate per speaker ranges from 0.7% to 54.2% but since it is evaluated using only one sentence per speaker there are also effects that depend on the intrinsic complexity of the specific sentence. We had to use such an experimental design to be able to evaluate all 25 synthetic voices.

Figure 16 shows a scatter plot of the word error rate against the averaged familiarity of the speaker. There is a slight negative correlation between the two values with a Pearson correlation coefficient R of −0.36 (not significant) and a MIC of 0.32. The low MIC-R^2 value of 0.18 indicates that there is also no non-linear correlation within the data.

Figure 17 presents the scatter plot of the word error rate against the number of times a voice was correctly recognised. The Pearson correlation coefficient is 0.38 (not significant) and the MIC is 0.42 giving an MIC-R^2 value of 0.27, which also shows no non-linear correlation within the data.

Finally Figure 18 shows the scatter plot of the word error rate against the speaker similarity averaged across participants. The negative Pearson correlation coefficient was found to be very low at −0.29 and the MIC at 0.40, resulting in a MIC-R^2 value of 0.31, again indicating also no non-linear correlation within the data.

5. Discussion

The results obtained in the first experiment showed 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 Nass and Lee (2001) 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

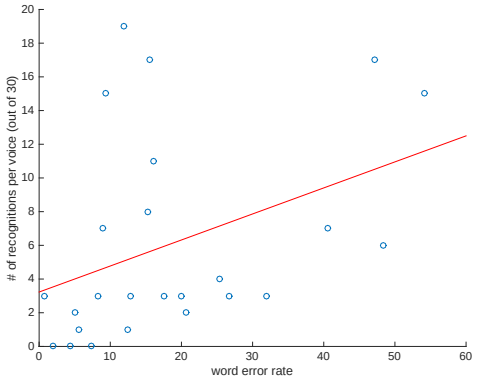

Figure 17: *Scatter plot of word error rate against # of recognitions per voice (out of 30) (Pearson correlation coefficient R* 0.38*; MIC* 0.42*; MIC-R^2* 0.27*).*

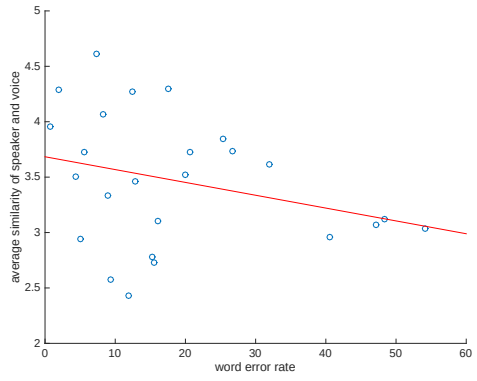

Figure 18: *Scatter plot of word error rate against speaker similarity (Pearson correlation coefficient R* −0.29*, MIC* 0.40*; MIC-R^2* 0.31*).*

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. To increase the identifiability of one's own voice and thereby also the preference it would be interesting to develop synthetic voices that sound like a listeners perception of their own voice. For this one could use a bone conduction microphone and develop a mixed synthesiser on recording from standard and bone conduction microphones following a similar approach as in Tamiya and Shimamura (2004).

In the recognition experiments that followed we showed that overall voice recognition is quite low (20%), but that blind children and young adults outperformed visually impaired children and young adults on this task. One reason could be that identifying a speaker from his/her voice is more important for blind than for visually impaired children and young adults. Bull et al. (1983) already showed that blind individuals have a higher voice recognition accuracy than non-blind listeners, but they could not find a difference in voice recognition for different degrees of blindness by using natural speech samples. In terms of recognising one's own voice we saw a trend that the school children were better in recognising their own voices (55% correct) than recognising voices of others (25%), which were largely known to them, but we would need a larger experiment for a definitive proof of this fact.

We also showed that the average familiarity with a speaker and the similarity between a speaker's synthetic and natural voice are correlated to the speaker's synthetic voice recognition rate. Interestingly we observed however that the intelligibility of a synthetic voice, measured by word error rate, was not correlated with any of the speaker similarity measures we extracted.

It would have been interesting to investigate the relationship between the performance and engagement times obtained in the first experiment and the speaker familiarity, similarity, and word error rate obtained in the second evaluation. This was however not possible because we do not have enough performance and engagement data for any particular synthetic voice. To investigate the relationship between these two sets of variables we would need more listeners for the first type of evaluation, such that many listeners would listen to the same synthetic voice and all 25 synthetic voices were covered. Still we believe that the results obtained by the two experiments agree to a certain extent: participants were more engaged when playing audio games with familiar voices and voices judged to be more familiar were also easier to identify, which could explain the increase in engagement time and performance. Voice intelligibility however did not seem to correlate with the speaker recognition rates.

6. Conclusion

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

We think that since our results hold for this mixed group of children and young adults, where the majority are children, it is likely that the results on the perception of one's own voice can be extended to blind individuals in general, and maybe also to all types of listeners.

We also showed that blind listeners engage longer with the audio games than visually impaired listeners. We hypothesise that blind listeners are more accustomed to listening to synthetic speech and it is easier for them to process synthetic speech. The ease of processing is also shown by their ability to process fast synthetic speech.

In the second experiment we showed that blind children and young adults were also better in recognising synthetic voices than their visually impaired companions.

We also saw that there was a clear correlation between speaker similarity and speaker recognition, as well as between average familiarity with a speaker and speaker recognition. We found no correlation between intelligibility and familiarity, similarity, or speaker recognition.

For blind users that are using speech synthesis on a regular basis there is a need to make their synthesiser experience

more engaging and pleasurable, which can be accomplished by using their own or familiar voice in the synthesiser. In voice user interface design we should use the adaptive capabilities of state-of-the-art speech synthesis technology and support such an adaptation to individual user needs.

For the future it would be interesting to investigate the interplay between engagement time, familiarity, speaker recognition, similarity, and word error rate by combining the two types of experiments. For such an evaluation we would however need a large number of users to evaluate all 25 voices. It would also be interesting to synthesise speech that is as close as possible to the speech sound that a user hears for his/her own speech, and evaluate if this further increases the engagement and/or performance.

7. Acknowledgement

This work was supported by the BMWF - Sparkling Science project *Speech synthesis of auditory lecture books for blind children* (SALB) and by the Austrian Science Fund (FWF) project *Acoustic modelling and transformation of varieties for speech synthesis* (P23821-N23).

References

Appel, R., Beerends, J.G., 2002. On the quality of hearing one's own voice. Journal of the Audio Engineering Society 50, 237–248.

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

Bissiri, M., Pfitzinger, H., 2009. Italian speakers learn lexical stress of german morphologically complex words. Speech Communication, Special Issue on Spoken Language Technology for Education 51(10), 933–947.

Bőhm, T., Shattuck-Hufnagel, S., 2007. Utterance-final glottalization as a cue for familiar speaker recognition, in: Proc. Interspeech, Antwerp. pp. 2657–2660.

Bonneau, A., Colotte, V., 2011. Automatic Feedback for L2 Prosody Learning, in: Ipsic, I. (Ed.), Speech and Language Technologies. Intech, pp. 55–70. URL: `https://hal.inria.fr/inria-00579255`.

Bull, R., Rathborn, H., Clifford, B.R., 1983. The voice-recognition accuracy of blind listeners. Perception 12, 223–226.

ETSI, 1996. Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks. Technical Report. ETSI TR 250.

Fernyhough, C., Russell, J., 1997. Distinguishing one's own voice from those of others: A function for private speech? International Journal of Behavioral Development 20, 651–665.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for Mel-cepstral analysis of speech, in: Proc. ICASSP, San Francisco, USA. pp. 137–140.

Hugdahl, K., Ek, M., Takio, F., Rintee, T., Tuomainen, J., Haarala, C., Hmlinen, H., 2004. Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds. Cognitive Brain Research 19, 28 – 32.

ITU-T, 1993. Telephone transmission quality - Measurements related to speech loudness - Some effect of sidetone. Technical Report. ITU-T Supplement 11, Series P.

Lancker, D.V., Kreiman, J., 1987. Voice discrimination and recognition are separate abilities. Neuropsychologia 25, 829–834.

Levinson, S., 1986. Continuously variable duration hidden Markov models for automatic speech recognition. Computer Speech & Language 1, 29–45.

Moos, A., Trouvain, J., 2007. Comprehension of ultra-fast speech – blind vs. 'normally hearing' persons, in: Proc. Int. Congress of Phonetic Sciences, Saarbrücken, Germany. pp. 677–680.

Nass, C., Lee, K.M., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied 7, 171.

Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. J. of Phonetics 35, 85 – 103.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Perception & Psychophysics 60, 355–376.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech perception as a talker-contingent process. Psychological Science 5, 42–46.

Odell, J.J., 1995. The Use of Context in Large Vocabulary Speech Recognition. Ph.D. thesis. University of Cambridge. Cambridge, UK.

Papadopoulos, K., Argyropoulos, V.S., Kouroupetroglou, G., 2008. Discrimination and comprehension of synthetic speech by students with visual impairments: The case of similar acoustic patterns. Journal of Visual Impairment & Blindness 102, 420–429.

Pisoni, D., Remez, R., 2008. The Handbook of Speech Perception. John Wiley & Sons.

Pörschmann, C., 2000. Influences of bone conduction and air conduction on the sound of one's own voice. Acta Acustica united with Acustica 86, 1038–1045.

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

Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V., 2010b. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. Speech Comm. 52, 164–179.

Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., Schmid, E., 2015. Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games, in: Proc. Interspeech, Dresden, Germany. pp. 1625–1629.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., Sabeti, P., 2011. Detecting novel associations in large datasets. Science 223, 1518–1524.

Reynolds, M., Isaacs-Duvall, C., Sheward, B., Rotter, M., 2000. Examination of the effects of listening practice on synthesized speech comprehension. Augmentative and Alternative Communication 16, 250–259.

Rosa, C., Lassonde, M., Pinard, C., Keenan, J.P., Belin, P., 2008. Investigations of hemispheric specialization of self-voice recognition. Brain and Cognition 68, 204–214.

Souza, P., Gehani, N., Wright, R., McCloy, D., 2013. The advantage of knowing the talker. Journal of the American Academy of Audiology 24, 689.

Syrdal, A.K., Bunnell, H.T., Hertz, S.R., Mishra, T., Spiegel, M.F., Bickley, C., Rekart, D., Makashay, M.J., 2012. Text-to-speech intelligibility across speech rates., in: Proc. Interspeech, Portland, USA. pp. 623–626.

Tamiya, T., Shimamura, T., 2004. Reconstruction filter design for bone-conducted speech., in: Proc. Interspeech, Jeju Island, Korea. pp. 1085–1088.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K., 2013. Speech synthesis based on hidden markov models. Proceedings of the IEEE 101, 1234–1252.

Toman, M., Pucher, M., 2015. An open source speech synthesis frontend for hts, in: Proceedings of the 18th International Conference of Text, Speech and Dialogue (TSD), PlzeÅ, Czech Republic. pp. 291–298.

Van Lancker, D., Kreiman, J., Emmorey, K., 1985. Familiar voice recognition: Patterns and parameters. part I: Recognition of backward voices. J. of Phonetics 13, 19–38.

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

Yamagishi, J., Kobayashi, T., 2007. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE Trans. Inf. & Syst. E90-D, 533–543.

Yamagishi, J., Veaux, C., King, S., Renals, S., 2012. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. Acoustical Science and Technology 33, 1–5.

Yonan, C.A., Sommers, M.S., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychology and Aging 15, 88.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Communication 51, 1039–1064. doi:`10.1016/j.specom.2009.04.004`.