Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

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

Q1 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), Austria*
[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*

Received 29 April 2016; received in revised form 19 January 2017; accepted 28 May 2017

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 recognizing 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;

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author.
E-mail address: michael.pucher@oeaw.ac.at (M. Pucher), bettina.zillinger@fhwn.ac.at (B. Zillinger), m.toman@neuratec.com (M. Toman), dietmar.schabus@ofai.at (D. Schabus), cvbotinh@inf.ed.ac.uk (C. Valentini-Botinhao), jyamagis@inf.ed.ac.uk (J. Yamagishi), erich.schmid@bbi.at (E. Schmid), thomas.woltron@fhwn.ac.at (T. Woltron).

http://dx.doi.org/10.1016/j.csl.2017.05.010
0885-2308/

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) investigate how children's private speech allows them to learn to distinguish between their own and other's voices. Appel and Beerends (2002) investigate the perception of one's own voice in a telephone setup where echo and distortion is introduced. Rosa et al. (2008) show 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 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 modeled 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 normalization, i.e., the mapping of an acoustic realization 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 (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 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 synthesizers (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

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

Please cite this article as: M. Pucher et al., Influence of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.05.010