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]

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

`{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.

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.
- For the blind children and teachers the sentences were played to the listeners via loudspeakers at a normal rate.
- 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.
- Figure 2 shows the comparison between all voices (natural and synthetic).
- To visualize the voices we performed Dynamic Time Warping (DTW) between the same prompts from different speakers.

Figure 1: Studio recordings of blind school children.

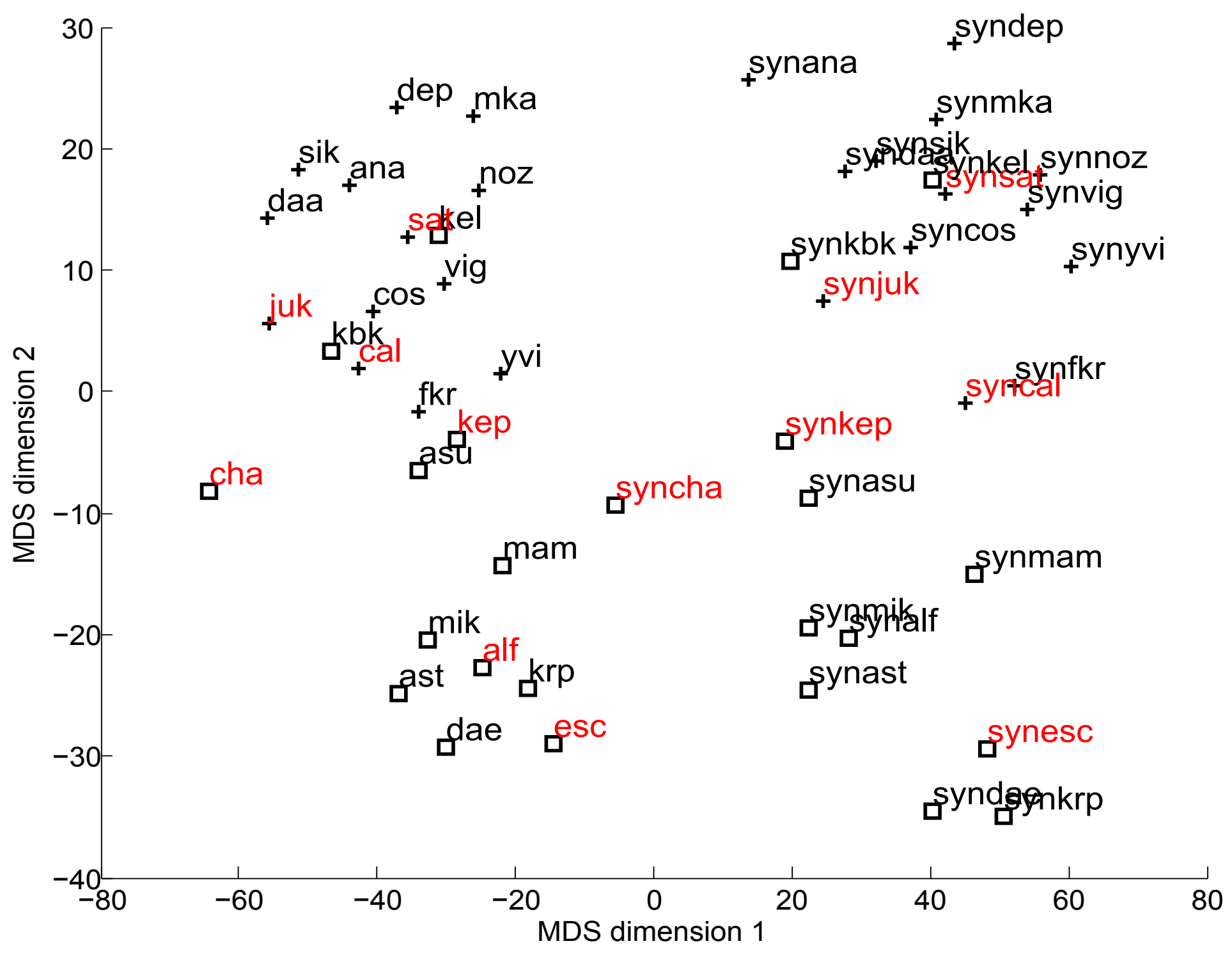

Figure 2: Comparison between synthetic ("syn" affixed) and natural voices. School children are marked in black, teachers in red. Female speakers with crosses, male speakers with squares.

Audio games

- The **labyrinth game** was used to measure engagement time.
- 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.
- The **memory game** was used to measure the performance of the player.
- Each round had a specific topic, e.g., musical instruments or animals.
- Upon key press, the synthetic voice pronounced the item associated with the field.

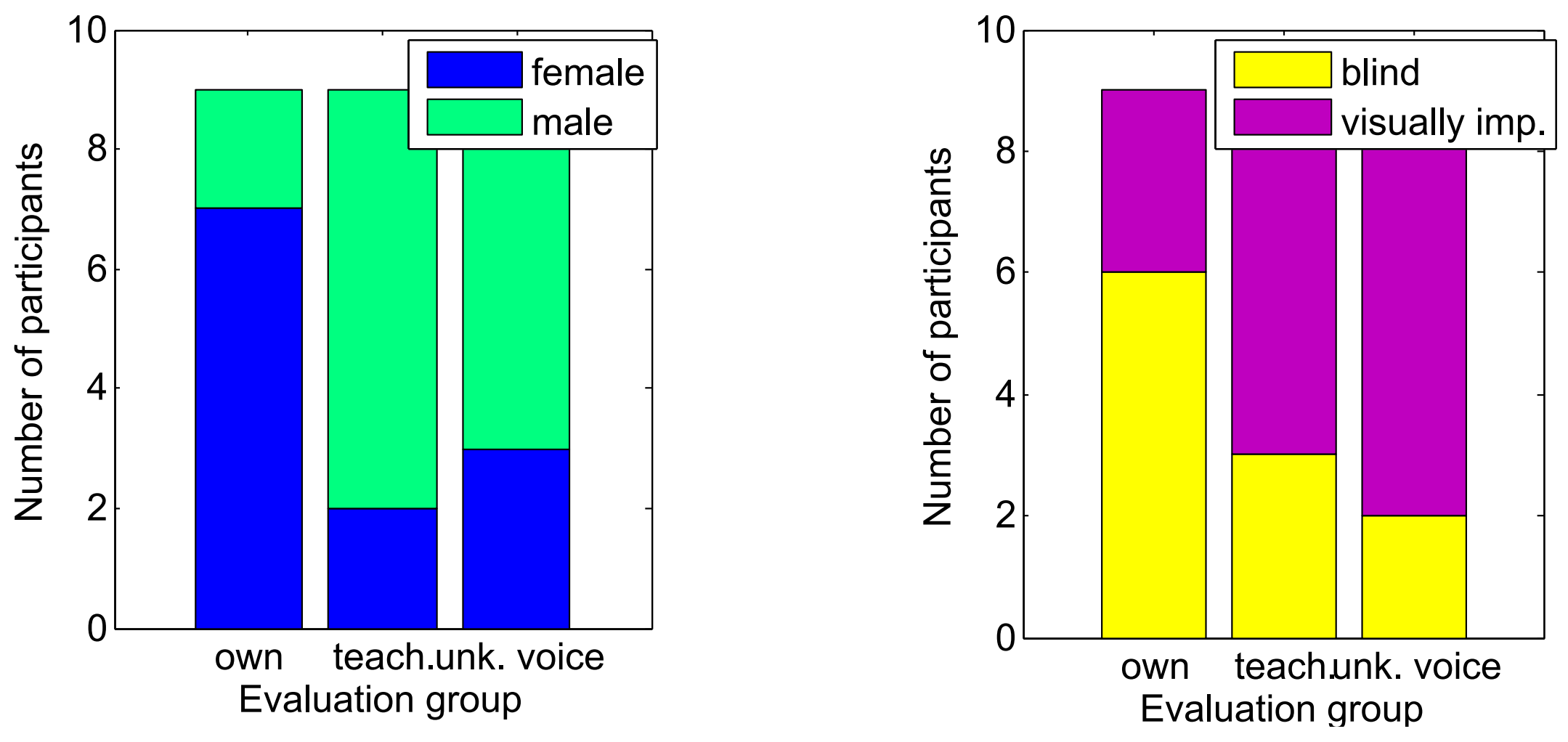

Figure 3: Participants characteristics within groups.

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.
- To measure engagement in the **labyrinth game** we used the time played overall and the number of games that were played.
- Figure 4 (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.
- In the experiments with the **memory games** the children had to play 8 mandatory rounds.
- We used the number of steps needed to solve all 8 rounds as performance variable.
- Figure 5 (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.
- Our results show that the use of one’s own voice increases the engagement time in audio games, which indicates a certain preference.
- 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.

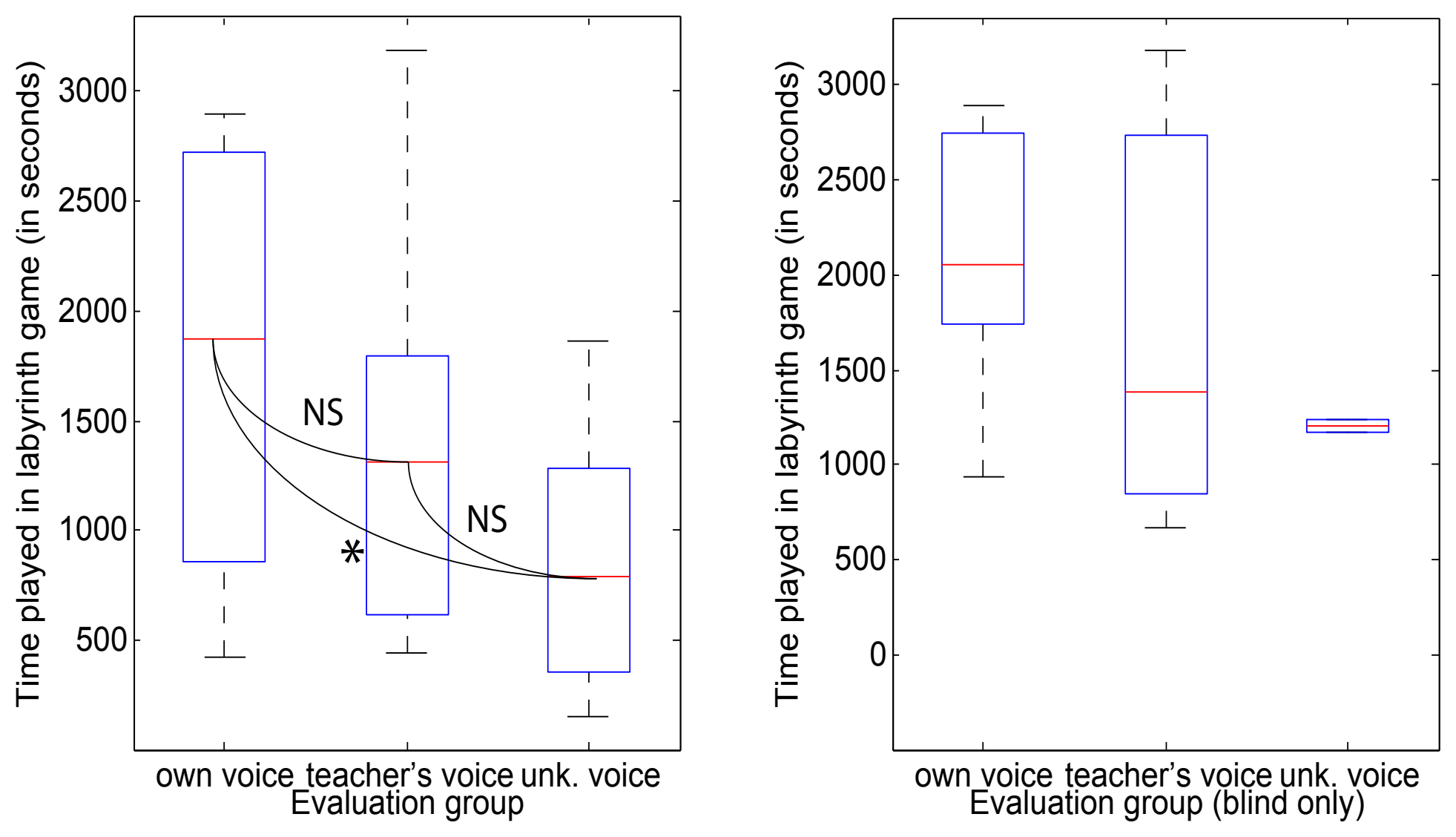

Figure 4: Time played per group in the labyrinth game for all participants (left) and blind-only participants (right).

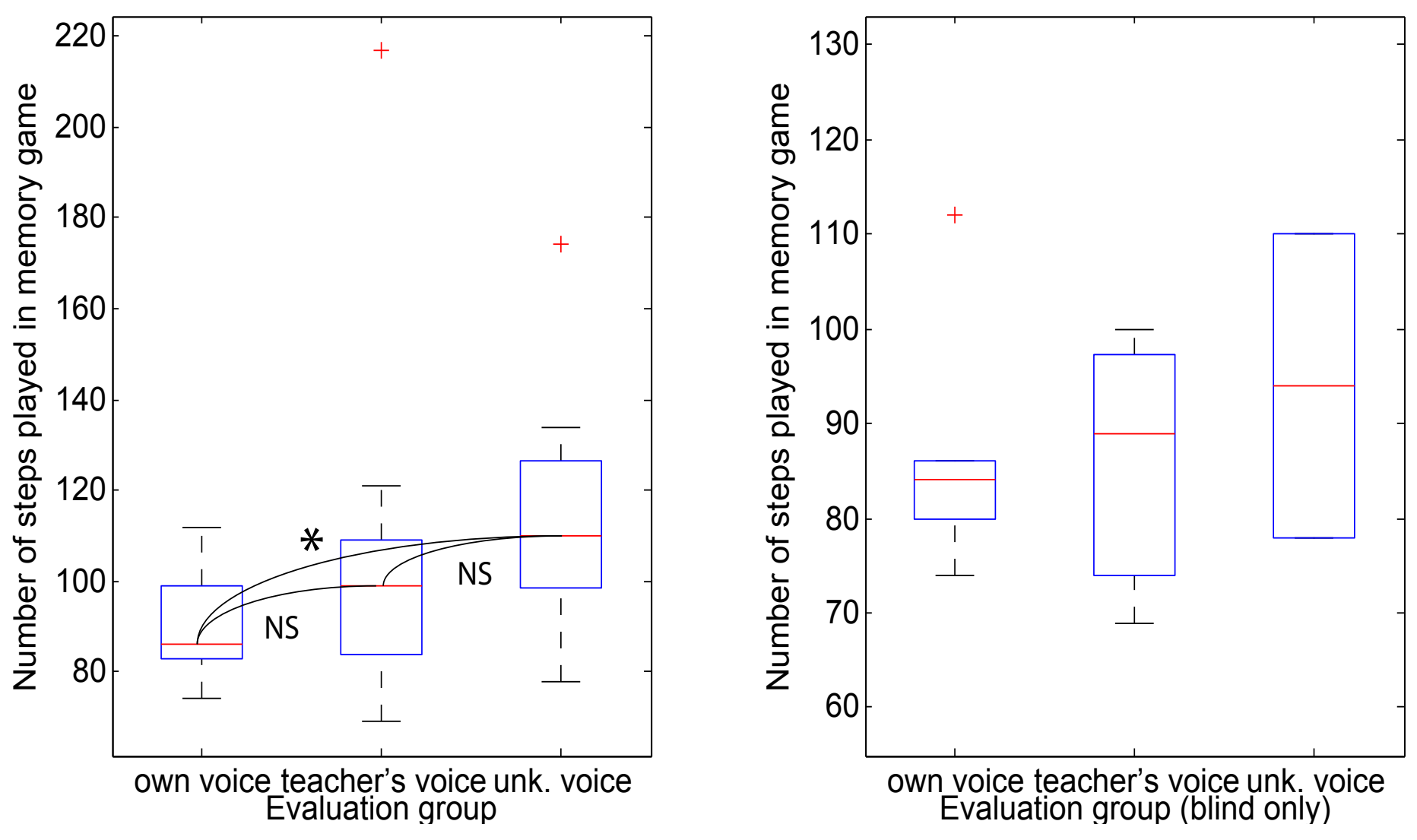

Figure 5: Number of steps per group in the memory game for all participants (left) and blind-only participants (right).