ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

Speech Communication: Paper ICA2016-738

A 3D computer game for testing perception of acoustic detail in speech

Daniel Duran[a], Natalie Lewandowski[b], Antje Schweitzer[c]

[a]Universität Stuttgart, Germany, Daniel.Duran@ims.uni-stuttgart.de
[b]Universität Stuttgart, Germany, Natalie.Lewandowski@ims.uni-stuttgart.de
[c]Universität Stuttgart, Germany, Antje.Schweitzer@ims.uni-stuttgart.de

Abstract

We present a novel experimental framework for perception studies and an application focusing on attention to fine phonetic detail in natural speech perception. Traditional psychological experiments in research on speech perception do not provide a natural testing scenario (notorious supervision and lack of naturalness). A solution to this problem is employing a computer game in which attention to fine phonetic detail comes natural. Computer games are increasingly used in psychology or in studying emotional speech production, where the communication in multiplayer games is recorded. Our novel framework implements a traditional psycholinguistic AB test paradigm within a computer game. Using a state-of-the-art game engine, we developed a first person shooter. This genre is ideally suited to implement a test scenario which requires the subjects to click on a specific point on the screen as fast as possible. The player moves around within a virtual 3D environment and reacts to stimuli presented by enemies which belong to two different categories, each of which is associated with one response key. The two categories are initially distinguished by visual and acoustic cues (e.g. different colors, and different sounds). Gradually, visual cues are removed. Thus, the subject has to attend to the acoustic cues and react accordingly. An additional important aspect of our framework is the high involvement in the game and motivation of the subjects to solve the task. In traditional psychological experiments, on the other hand, subjects may easily get tired or bored by the repetitive, unnatural task. We discuss practical and theoretical challenges encountered with the implementation of a psychological test within a computer game.

Keywords: speech perception; experimental framework; computer games

A 3D computer game for testing perception of acoustic detail in speech

1 Introduction

Computer game paradigms enjoy increasing popularity in psychological experiments [1, 2, 3]. In his review, Washburn [1] concludes that computer games, at least under certain conditions, "can actually increase the ecological validity or, at least, the lifelikeness of psychological research." Foreman [2] observes that psychological research has benefited from the development of virtual environments, and Kimball et al. [3] demonstrate the usefulness of computer games for research on speech sound learning.

1.1 Motivation

In her dissertation on phonetic convergence (the phenomenon of two people interacting with each other becoming more similar in their pronunciation), Lewandowski [4] found that phonetically talented subjects in a second language setting converged more than less talented subjects. This finding was explained by proposing that *attention* to fine phonetic detail is a prerequisite for its successful memory storage and subsequent retrieval in speech production. An individual's ability to pay attention to fine phonetic detail, in turn, was hypothesized to be a substrate of phonetic talent, which escapes conscious access and direct control and is located at the core of the convergence mechanism (alongside individual personality features which may influence adaptation). This hypothesis is supported by a post-hoc analysis of the convergence results in [4] involving data from a mental flexibility test, which revealed a positive correlation between the two dimensions. The mental flexibility test required subjects to quickly re-tune their attention to a changing scenario. The faster the subjects were in this test, the more phonetic convergence they displayed during the dialogs [5]. Segalowitz [6] proposes that two processes contribute to overall fluency in speech (in both decoding and producing), *access fluidity* (AF) and *attention control* (AC). Attention control was defined as the ability to focus and refocus attention on different semantic levels (local vs. global meaning relations). While Segalowitz focuses on shifting between local and global meaning access, we propose that attention control can also be involved in switching between various dimensions of the speech signal, for instance between detailed acoustic shape and meaning. AC is usually tested in an alternating runs paradigm [7], where participants make a series of judgments in two alternating differing tasks. Access fluidity, on the other hand, concerns the speed and/or automaticity of connecting words to their meaning. AF is usually measured by reaction times in (lexical or semantic) judgment tasks or tasks for automaticity in comprehension [6].

In currently on-going research on phonetic convergence in dialog situations [8], we apply various psychological tests aiming at participants' attention to phonetic detail. The typical tasks used in testing a subject's attention control are notoriously supervised, loaded with feedback control, and lacking in naturalness. A solution to this problem is employing a computer game paradigm in which AC and AF are a natural part of the game. Our hypothesis thus is that a

computer game will yield more natural data as paying attention arises as a necessity from the game scenario and requires a certain action in response to an event rather than an explicit judgment of any sort.

Wade & Holt [9] studied incidental perceptual learning of spectrally complex nonspeech sounds. They point out that the usual "categorization-with-feedback training" approach in studies of non-speech auditory categorization "is demonstrably quite unlike the processes by which humans are exposed to natural language sounds, and perhaps so fundamentally different as to preclude informative comparison." They thus propose a method which "captures several essential aspects of phonetic acquisition": a computer game incorporating game features of typical first person shooting (FPS) games. These are action games which are played from a first person perspective involving navigation through a 3D virtual environment. Lim & Holt [10] used the same computer game paradigm to train the adult native Japanese participants of their study to distinguish between English /r/–/l/ categories. Their results show that participants learned non-native phonetic categories without explicit categorization training. The result of 2.5 h of computer game-based training was comparable to 2–4 weeks of training with standard categorization training paradigms with explicit feedback.

In what follows, we describe our novel computer game-based framework for phonetic perception studies. In addition, we present an application of our game in an experimental study focusing on attention to fine phonetic detail in natural speech perception.

2 Outline of the ΨX732-framework

The ΨX732 computer game (pronounced ['saɪ ɛks] 732) is implemented using the Unity game engine and its editor [11]. This provides a state-of-the-art game engine for a high-quality 3D game which subjects experienced with modern computer games may find appealing. A first concept has been developed and tested in a perception study by Lange & Pfeiffer [12]. However, ΨX732 is a completely new implementation borrowing some ideas and concepts from that earlier work.

All game logic (like input handling, agent behavior, experiment control, logging, etc.) is implemented in C#. Experimental parameters are not hard-coded into the game but can be set through a plain text file which is loaded by the game at runtime. The structure of this configuration file corresponds to the Java *.properties* format with key–value pairs. These configurable parameters include, a.o., time limits, trial specifications and also the texts displayed on screen. The actual sound files (using wav format) are not compiled into the game, as well, but loaded at runtime from hard disk. This makes ΨX732 very flexible, providing a language-independent framework for various experimental scenarios.

2.1 Game stucture

The general setting is that of a first person shooter game. The player encounters *agents* (we avoid the usual term *enemy*) within a natural looking landscape and has to react to them. Agents belong to two categories (explained within the story as "alien invaders" vs. "human

civilians"). Once the player approaches a given agent, the agent becomes active and starts chasing the player. The trial sound stimulus is played once and a visual display next to the agent shows a blue color for aliens or a green color for humans along with a descriptive text label. The colors correspond to the colors of the rays emitted from the two "weapons" for the respective agent categories. A crosshair at the center of the screen aids in precise aiming.

The player is equipped with two tools (we avoid the usual term *weapons* within our game): one that freezes a hit agent in a block of blue ice and one which beams a hit agent up on board of a safe space ship within a bundle of green light rays. The tools are associated with the left and right mouse buttons. Both tools emit blue and green light rays which form a straight line from either the left or right side of the screen (corresponding to the left and right mouse buttons) to the hit object. If no agent was hit, a spherical blob of light appears at the hit target position. Both light ray and blobs disappear after a short time. This provides feedback for the player about current rotation and it makes aiming easier. The correspondences between mouse buttons, tools, colors and agent categories (*blue = freezer = alien*, *green = beamer = human*) are constant for the entire game.

The ΨX732 computer game consists of one introductory level and four experiment levels. A welcome screen which displays the title of the game and shows a rotating space station with a distant Earth in the background. On this first screen, the player has to enter their name and then fundamentals of the game (controls and story) are shown in text form.

The first level, then, is the introductory level within the space station. Here, the player is told about the background story which provides an engaging scenario for the player to immerse herself in. An essential part of the introduction is a *training* program where the player first has to navigate a path through a corridor and then test the equipment shooting test dummies of both categories. This training serves as a block of *practice trials*. It has been shown that differences between experienced gamers and individuals without 3D gaming experience can be alleviated after a short amount of training within the virtual environment [13]. This training also serves as a means to establish an individual baseline for each player by collecting statistics about navigation and targeting accuracy and reaction times in a relatively relaxed scenario prior to the actual game levels.

The remaining four levels of the game are actual experiment levels. They are all set in an open outdoor scenery on the surface of planet Earth: on an island group surrounded by open water, canyons or mountains (see figure 1). Within these levels, the player encounters a number of agents as specified by the experimental configuration. Due to the open layout of the levels, agents can often be easily seen from a long distance. They initially appear as glowing red, semi-transparent human figures (shown in Figure 1). This is explained to the player as an effect of a long-range detector which locates all targets. Once the player comes closer to a given agent, its figure is replaced by a female human character. This is a deliberate brake with usual conventions: agents are not hard to find in our game, allowing the player to prepare for them well in advance (depending on the current location within a level, where agents suddenly appearing behind a corner is not impossible). Inactive agents wander around aimlessly.

Immediate feedback is provided by a score shown on screen which may increase or decrease

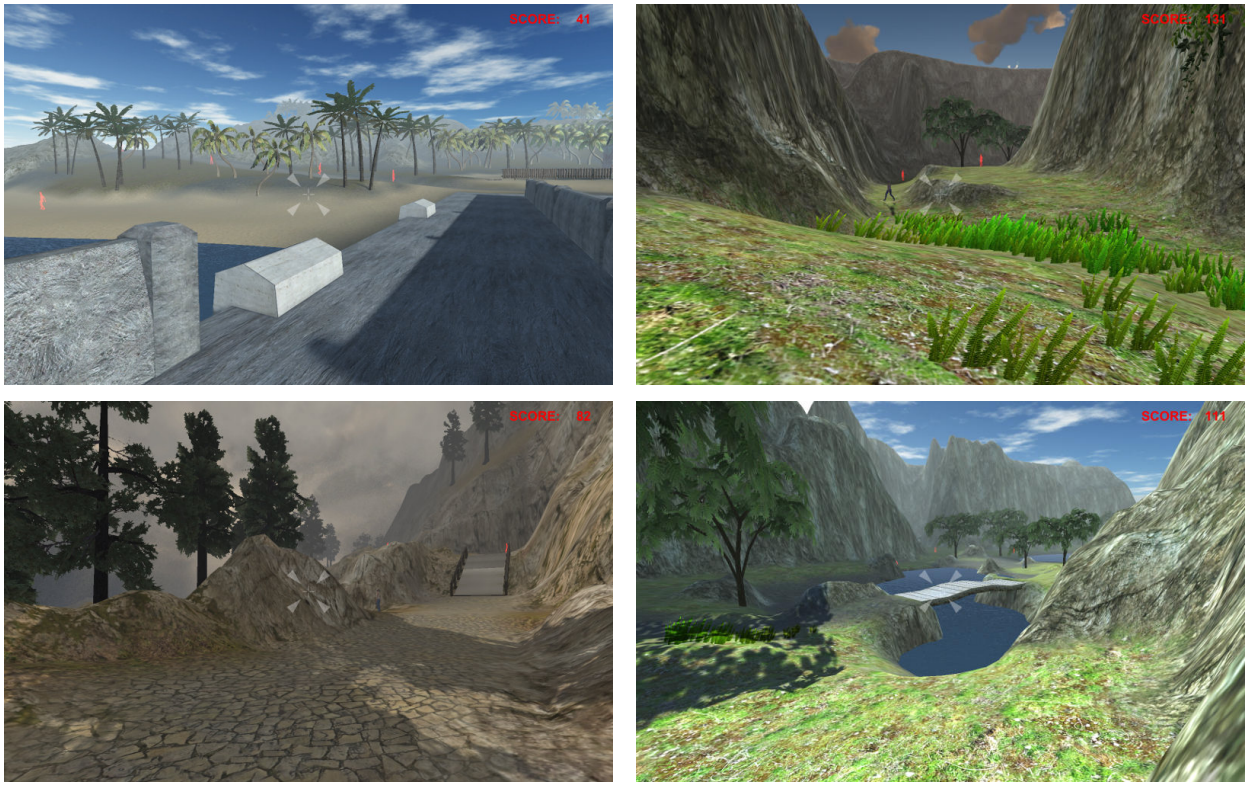

Figure 1: Screenshots showing views of the four experiment levels.

according to the player's actions. A slight bias towards positive score increments in the game leads to a steady increase of the score. This provides additional motivation for the player. The game ends with a final screen which shows a highscore table in front of the space station from the introductory screen. All scores are stored in a file and later used to display the highscore table (the player's name is not stored and the table uses some placeholder names for all scores except for the current player). The last screen also provides statistics about correct and wrong hits, failed shots etc. for all four experiment levels.

Player actions are logged and stored in a text file for post-processing and evaluation. In addition to Unity's own time measurements, time stamps of relevant events are logged using the C# class `System.Diagnostics.Stopwatch`, which allows for fine grained time measurements. In order to elicit fast reactions bonus scores are given as an incentive for fast, correct clicks. All mouse clicks are recorded along with the player's position within world space and the corresponding target agent's position in world space and screen space. This allows for post-analysis of clicks with the correct mouse button which missed the target. The rough direction of such clicks can be taken into account allowing for the evaluation of correct reactions despite only nearly missing the target in the game.

2.2 Experimental trials

Experimental trials within the game are defined by the following events: (1) The trial starts with the playback of an acoustic stimulus (a sound file as specified by the game's configuration). At this moment, one agent becomes active. (2) The trial ends when the player hits the agent with one tool or when the player passes the currently active agent.

Due to the game nature of the framework, it is possible that a player does not react to an active agent but continues on his path through the level. Once a pre-defined distance threshold is reached between the player and the currently active agent, the agent is removed and counted as a missed trial. It is also possible to shoot an agent before it is active, i.e. before the start of a trial. This is counted as a distance hit.

Attention to acoustic/phonetic detail is assessed by gradually removing the visual cues (both text and color) from the agents. Eventually, the player can only decide upon the category of the agents (alien vs. human) by relying on the acoustic stimuli they emit.

Despite the efforts of creating a modern-looking, high-quality computer game with a state-of-the-art game engine, some usual game conventions had to be broken in order to implement a reliable psychological test paradigm: One such deviation from usual game design patterns is that only one agent at a time can become active. Once a trial is started, all other agents remain in their inactive state even if the player comes close within activation distance. Another deviation from usual game designs is that agents become only active if they are within a narrow window in front of the player. On the one hand, this allows the player to prepare for the attack by aiming at the agent in advance. Thus faster reactions can be achieved by reducing the aiming effort. On the other hand, this also reduces the amounts of in-place rotations of the player which may be difficult for inexperienced players or even cause discomfort (see discussion below).

3 Experiment

We apply the ΨX732-framework in a study to test perception of acoustic detail in speech. The experiments were conducted at the Institute for Natural Language Processing of the University of Stuttgart. The computer game was employed as a testing framework within a larger, on-going study on the role of individuals' attention to fine phonetic detail in dialog situations [8]. Only the application of the ΨX732 computer game is reported here.

3.0.1 Material

A female native speaker of German was recorded for this study. Various utterances, ranging from single words to entire sentences, covering different phonetic phenomena were recorded in a sound attenuated booth.

Four phonetic features have been manipulated using Praat [14]: the range of the fundamental frequency (f0), the height of the second formant of vowels (F2), an increased voice onset time of plosives (VOT) and removed lower frequencies in the spectrum of fricative sounds

(FRIC) using a band-pass filter. The manipulated stimuli are always associated with the "alien" category and the original utterances with the "human" category. The four different phonetic cue categories correspond to four experiment levels of the game.

3.1 Participants

Thirty adult, native speakers of German (15 female) were recruited on campus of the University of Stuttgart. All participants reported no known hearing impairments. In order to counter any effects of experiment order, the participants have been grouped into four groups as shown in Table 1.

Table 1: Participant groups

Group	Exp.1	Exp.2	Exp.3	Exp.4	f	m
1	f0	FRIC	VOT	F2	4	4
2	F2	VOT	FRIC	f0	3	4
3	FRIC	F2	f0	VOT	3	5
4	VOT	f0	F2	FRIC	3	3

3.2 Method

Participants were seated in a quiet, window-less room in front of a computer monitor. A standard office keyboard and a mouse were used as input devices. Participants were wearing Sennheiser headphones. The game was run on a Windows platform. The display resolution of the game used for this study was 1680×1050 pixels. At the beginning of the session, participants were instructed to adjust the sound volume to a comfortable level, such that they can hear well. It was not pointed out to the participants that they had to rely on the speech stimuli in order to distinguish the two agent categories. The time limit for each experiment level was set to 12 minutes. A session with the game thus lasted for approximately one hour in total. The trials within each experiment level have been randomized, without repetitions of stimuli, for each participant. During the entire session, the player is alone in the room and the experimenter waits in the next room with a closed door in between.

3.3 Preliminary results

The experiments were not finished at the time of writing this paper. We present here preliminary results for the first subjects who played the game in Table 2 (note that data for some players is missing).

In this table, the number of clicks are shown for each player in the four experiments (which are in different orders according to the players' group as shown in Table 1). Column c shows the total number of correct hits, i.e. cases where the player clicked the correct mouse button and hit the target agent. Column w shows the total number of wrong hits, i.e. clicks with the wrong mouse button that hit the target. Column d shows the number of hits from a long distance, i.e. before the hit agent had been activated and associated with a trial. Column n, finally, shows

the number of clicks that did not hit the target agent. As the number of stimuli is balanced for both categories, the ratio c/w should be above 50% in order for the number of correct hits to be above chance level. Most players reported after the game that they had no clue how to identify the agents once the visual cues were gone. Nevertheless, in most cases it can be observed that $c > w$. d is very low for most players which shows that the participants really waited for the agents to start their chasing and produce the sound stimulus. Moreover, though easily possible within the game, no player of the current study ever missed a single trial, i.e. ignored an active agent.

The peculiar results of player 4 can be explained by the specific strategy that player chose: he reported after the game that he chose to shoot as fast as possible as he could not detect any cue to categorize the agents. The instructions within the game may have been misleading promising a bonus for fast reactions. Only one participant (player 17) had to abort the game prematurely after 3 levels due to slight dizziness. This might be a case of cybersickness.

The raw numbers in this table do not take into account correct reactions (i.e. correct selections of either the left or right mouse button) which did not hit the target agent. These clicks are contained in the *no hit* (n) columns. They do also not take into account the fact that at the beginning of each level, players could rely on visual cues. A more detailed analysis is required, splitting all player actions into three blocks per level: (1) the beginning phase during the first 5 trials when visual cues were present, (2) the transition phase when visual cues fade away and (3) the final phase during which no visual cues are present.

Table 2: Results: c = correct hit; w = wrong hit; d = premature distance hit; n = no hit.

player	sex	group	f0				F2				VOT				FRIC			
			c	w	d	n	c	w	d	n	c	w	d	n	c	w	d	n
1	f	1	26	13	0	1	43	27	2	7	37	27	2	6	38	26	2	4
2	m	2	45	23	0	3	23	18	0	7	37	23	1	4	47	28	0	1
3	m	4	50	28	0	5	56	40	0	1	39	17	0	1	72	10	0	12
4	m	3	34	22	45	14	31	12	32	17	36	21	25	22	34	22	0	1
5	m	3	57	32	0	3	42	36	2	2	50	33	2	4	36	24	0	0
6	f	3	47	46	0	1	37	30	0	3	29	9	0	0	35	18	2	3
7	m	3	56	47	2	10	46	27	0	11	53	45	0	5	46	27	1	10
8	m	1	46	23	0	2	49	26	0	1	54	44	0	0	43	32	0	1
9	f	4	38	28	0	0	35	19	0	2	35	26	0	1	24	15	0	2
10	m	2	60	28	0	2	34	18	0	0	54	23	0	2	67	15	0	1
11	f	4	32	25	0	2	44	30	0	3	46	21	1	2	39	39	1	3
12	m	4	47	30	1	2	53	47	4	1	33	30	0	1	46	36	3	1
13	f	2	19	6	0	2	24	23	0	4	39	27	0	1	36	22	1	0
14	m	3	46	47	0	1	44	29	1	1	43	28	1	3	28	14	1	3
15	m	2	49	31	3	5	38	21	0	0	43	37	0	3	47	35	0	0
16	f	1	30	19	0	1	41	27	0	0	45	36	1	0	23	6	0	0
17	f	2	NA	NA	NA	NA	32	23	2	5	46	12	0	0	31	32	0	0
18	f	3	48	24	1	3	35	13	0	5	37	27	0	0	34	16	0	1
19	f	4	56	23	0	9	54	37	0	11	27	25	0	3	59	11	1	18
21	m	1	24	11	2	2	41	37	0	0	42	21	1	2	33	23	0	0

4 Discussion

One potential problem of computer games as phonetic research tools is the phenomenon of *cybersickness*. Frey et al. [13], for example, report symptoms of cybersickness (nausea, headaches or dizziness) in nine out of 85 participants. They estimate the highest risk of experiencing cybersickness for female participants aged over 31 years with no or only little experience with first-person shooter games. They also point out that the risk of cybersickness may be reduced, for example by having a stable horizon or by avoiding narrow pathways. In conclusion, Frey et al. [13] emphasize that "game-based virtual 3D environments can be used for experiments even when participants have no prior experience with this type of environment."

5 Conclusions

Our preliminary results indicate that the ΨX732-framework is a suitable tool for testing perception of acoustic detail in speech. Its configurability makes it easy to address various different research questions by just modifying the configuration file and exchanging the sound files.

In general, computer games offer a promising alternative to traditional psychological tests with a more natural and motivating environment for test subjects.

The computer game will be made available to the research community.

Acknowledgements

This work is funded by the German Research Foundation (DFG) within the Collaborative Research Center SFB 732 / A4

References

[1] David A. Washburn. The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers*, 35(2):185–193, May 2003.

[2] Nigel Foreman. Virtual Reality in Psychology. *Themes in Science and Technology Education*, 2(1-2):225–252, 2009.

[3] Garrett Kimball, Rodrigo Cano, Jingyi Feng, Lei Feng, Erica Hampson, Evan Li, Michael G. Christel, Lori L. Holt, Sung-joo Lim, Ran Liu, and Matthew Lehet. Supporting research into sound and speech learning through a configurable computer game. In *IEEE International Games Innovation Conference (IGIC)*, pages 110–113, Sept 2013.

[4] Natalie Lewandowski. *Talent in nonnative phonetic convergence*. Doctoral dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2012.

[5] Natalie Lewandowski. Phonetic convergence and individual differences in non-native dialogs. Montréal, Canada, 2013. Abstract presented at New Sounds.

[6] Norman Segalowitz. Access Fluidity, Attention Control, and the Acquisition of Fluency in a Second Language. *TESOL Quarterly*, 41(1):181–186, 2007.

[7] Robert D. Rogers and Stephen Monsell. Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2):207–231, 1995.

[8] Antje Schweitzer, Natalie Lewandowski, and Daniel Duran. Attention, please! Expanding the GECO database. In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, 2015. Paper number 620.

[9] Travis Wade and Lori L. Holt. Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, 118(4):2618–2633, 2005.

[10] Sung-joo Lim and Lori L. Holt. Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science*, 35(7):1390–1405, September 2011.

[11] Unity Technologies. Unity. Computer program, 2016. Version 5.

[12] Lisa Lange, Bartholomäus Pfeiffer, and Daniel Duran. ABIMS – auditory bewildered interaction measurement system. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1074–1075, Dresden, 2015. ISCA Archive.

[13] Andreas Frey, Johannes Hartig, André Ketzel, Axel Zinkernagel, and Helfried Moosbrugger. The use of virtual environments based on a modification of the computer game Quake III Arena® in psychological experimenting. *Computers in Human Behavior*, 23(4):2026–2039, July 2007.

[14] Paul Boersma and David Weenink. Praat: doing phonetics by computer, 2016. Version 6.