# TESTING SPEECH PERCEPTION TODAY AND TOMORROW: SERIOUS COMPUTER GAMES AS PERCEPTION TESTS

*Natalie Lewandowski*[1], *Daniel Duran*[1,2]

[1]*Institute for Natural Language Processing, University of Stuttgart,* [2]*University of Freiburg*
*natalie.lewandowski@ims.uni-stuttgart.de*

**Abstract:** We present a methodological comparison between a computerized version of a classical perception experiment for the detection and categorization of fine phonetic detail and a newly designed serious computer game. We discuss general methodological consequences for experiments on auditory perception and attention to phonetic dimensions of speech and the important role individual differences play in the evaluation of the presented test set-ups.

## 1 Introduction

Serious games have been increasingly used in cognitive and psychological studies, and lately also in phonetic experiments [1, 2, 3, 4, 5]. Based on our studies we argue that serious computer games excel classical experimental designs in phonetic research in several aspects. While classical designs can oftentimes be perceived as repetitive and abstract, emanating from – and at the same time fostering – the artificial laboratory situation, games can substantially increase the naturalness and validity of collected data [6, 7]. Categorization tests (and any other linguistic/phonetic perception tasks) can be implemented in a way not requiring explicit instructions or revealing experimental goals, as classic designs largely do. They also allow to include another implicit variable – that of attention towards the phonetic aspects of the signal. Previous tests aiming at selective attention as the Flanker Test [8], inhibition (e.g. Stroop Test or Simon Test [9]) or attention switching as the Alternating Runs Paradigm [10, 11] or the Trail-Making-Test [12] for attention switching, contain at best verbal input in the visual domain (oftentimes single letters), much less often auditory stimuli (e.g. the *Auditory Stroop Test* [13], the Test of Everyday Attention [14]), and virtually never acoustic cues embedded within natural and meaningful speech. Although the Alternating Runs Paradigm, for instance, has been successfully adapted to include phonetic dimensions of the speech signal and for the usage in an L2 learning study (*Speeded Set-Switching Task*, with nasality as the phonetic cue, [15]), with the grid design and explicit instructions of the original version taken over, it leaves again very little room for naturalness. Classic phonetic categorization tests and attention tests share the feature of being heavily supervised and forcefully directing subjects' attention on specific attributes of the auditory/visual signal. Being faced with a fixed experimental display on screen or on paper with very limited visual or auditory attributes immediately gives away, albeit only implicitly, what the purpose of the experiment is. It also allows the subjects to focus all of their attention on the task at hand, which is, potentially, much more than they would have allocated to this specific task were it encountered in a natural, everyday-life scenario offering an environment much richer in cues and signal dimensions to choose from and attend to. Our game therefore is, rather than a means to investigate the maximum capacity of attention to phonetic properties of an incoming speech signal, a way of testing how much attention a listener *usually* would devote to phonetic properties in a signal, without being explicitly instructed to do so. Subjects in our game were not explicitly pointed towards the phonetic cues in the stimuli and had limited play time but even under these difficult conditions, particularly skilled listeners were successful in their categorization. Games can therefore provide a good balance between spontaneous, natural data and a full experimental control, and allow for a better distinction of participants' individual differences in a natural scenario resembling everyday communication.

In this paper we present a pilot study with a first comparison of the computer game with a classical setup for computer-based perception tests. We compare the performance of our subjects in both experiments, with the same auditory stimuli as the basis, albeit with a different testing order. We expect the performance of Group 1 to be worse in their first experiment (the game) than Group 2 in their first experiment (the classic categorization), since both settings contrast heavily in their explicitness as to the purpose of the test (i.e. the target cue to attend to) and also in the amount of attention that can be allocated exclusively towards this target (more attentional resources and fewer distractors available in the categorization experiment than in the game). We also predict that Group 2 should perform better on the game, since they already learnt that they should focus on the sound, rather than on the meaning, or any other property of the signal.

# 2 Methodology

## 2.1 Acoustic stimuli

The same stimuli were used in the game and the classic categorization experiment. A female native speaker of Standard High German recorded short sentences and phrases in German, with varying degrees of semantic relatedness to the game scenario. Some of the utterances matched the game content well, as e.g. "Der Captain ist tot" ("The captain is dead"), others had an intermediate relation to the game content and contained information that could be linked to the landscape and architecture present in the game levels, as. e.g. "Oh, Kokosnüsse!" ("Oh, coconuts!", plausible since there were palm trees at the island level). The rest of the phrases was devoid of any meaningful relation to the game storyline or the level appearance. The choice of a continuum of contentwise related and unrelated stimuli was motivated by, on the one hand, the desire to create a maximally authentic acoustic layer fitting the storyline of the game, while on the other hand, providing enough off-topic stimuli for the participants to implicitly realize that *meaning* is not the essential cue to differentiate between the two agent categories. Altogether 64 stimuli (16 for the trainings, 48 for the test phases) were used in both experiments. In order to create the "dialect" of the alien agents the original recordings were manipulated by altering the following phonetic characteristics: extended F0 range, altered fricative spectrum and shifted second formant of the vowel segments. The human agents were always assigned the original recordings, and the aliens always received the altered versions. The first eight stimuli in both test conditions belonged to the training phase. The order of items in every game level/test block was randomized for each participant.

## 2.2 Computer game design

The computer game was developed as a testing environment for the implicit perception and attention for phonetic detail [16, 17], following similar applications of serious games in speech perception experiments [1]. The story involves an alien invasion on Earth, and the task is to save all humans and catch all aliens. The acoustic stimuli were uttered by the human agent in the game, who could, however, also be a disguised alien. Each level contained only one type of original and manipulated stimuli, i.e. either fricatives, or F0, or F2. Every level starts with a training phase, where visual labels (with colors) are present, helping to distinguish aliens from humans in addition to the sound being played. These visual aids fade after eight trials and the auditory stimuli become the only way of successfully identifying the two categories. Both human and alien agents have the exact same appearance, that of a female human, with the aliens only changing to their real form after a response has been given by the player (pressing response key or mouse). The story within the game creates tasks for the players to solve, using their natural skill to pay attention to the target cues. The subjects, however, were not told which dimension of the auditory stimulus (sound, meaning, syntax, lexical content, etc.) was crucial for the distinction of agents but had to figure it out by themselves while playing and encountering the in-game agents.

## 2.3 Classical categorization experiment

The classical experiment was a categorization test with acoustic stimuli, designed in a way to maximally resemble the game scenario (involving the category labels "human" and "alien", just as in the game). All manipulated items belonged to the "alien" category, whereas the original recordings were used as the "human" samples. The nature of the manipulation was not communicated to the participants, however, since the setting was an auditory categorization test, it was obvious to the participants that they were supposed to focus on cues in the sound of the stimuli. This is in stark contrast to the game scenario, where the target dimension of the signal was never explicitly nor circumstancially revealed to the participants. Similarly to the game, after a short training phase, subjects had to categorize the stimuli in three blocks, with one manipulation at a time (as in the three game levels).

## 2.4 Participants

Our subjects were 24 German native speakers (age 20–31, 12 female) divided into two groups with 12 subjects which differed in testing order (game first vs. perception test first). The test group – Group 1 (G1) – played the game first and then completed the classic perception test, the control group – Group 2 (G2) – took part in the classic perception test first and played the game afterwards. The two test sessions followed each other with a 3–7 days' break. Analyzed were accuracy and reaction times, as well as individual post-hoc questionnaires on the evaluation of the two methods. Two participants suffered from a mild case of cybersickness while playing the game. After a short break, however, they were able to continue with the experiment. Since the break occurred still within the training phase before any RTs were measured, the data was included in the evaluation.

## 2.5 Post-hoc questionnaires

The first post-hoc questionnaire for every participant included sociodemographic information and questions on the usage of computers and other electronic devices, and the frequency and type of games played either on the computer, console or smartphone. The data was summarized in the following variables: *isGamer* (yes/no), *GamingFrequency* in days per week, and a *GamingScore* (i.e. How many types of games and on how many devices are usually played). The second questionnaire was filled out directly after the respective experiment (game and perception test) and included a.o. questions on the difficulty and fun of the game/test (on a Likert scale from 1-5), and also questions on the used "strategy" during the experiment in order to distinguish between aliens and humans.

# 3 Results

The data sets were transformed and prepared for analysis using R version 3.4.3 [18] and the packages *tidyverse* [19], *dplyr* and *stringr*. The statistical analyses were performed using *afex* [20] and *lmerTest* [21], as well as *ggplot2* [22]. Raw reaction times were first log-transformed before supplying it to the model. Visual inspection of normality plots did not show any obvious deviations. Descriptive statistics is presented in Table 1. The best fitting linear mixed model (lmer) for predicting the variable*RT(log)* was obtained by maximum likelihood t-tests using Satterthwaite approximations to degrees of freedom (lmerMod) after fitting a large model first and applying a combination of automatic and manually supervised stepwise reduction with the *step* procedure in the *lmerTest* package. The resulting best fitting model contains random intercepts for *stimulus* and *subject*, and the fixed factors shown in Table 2 (model parameters: AIC 816.9, BIC 887.1, logLik -394.5, deviance 788.9, df.resid 1102). The number of *correct responses* in the two test scenarios was predicted by fitting a maximal generalized linear model (GLM) of type *binomial* and a subsequent reduction of factors to achieve the best fit (see Table 3, model formula: *correct ~test * (group + difficulty)*; null deviance: 1416.5 on 1115 df, residual deviance: 1295.1 on 1110 df, AIC: 1307.1). The linear mixed model reveals that correct responses came hand in hand with shorter reaction times, and perceived *fun* in the experiments also reduced RTs. Furthermore, there was an effect for the type of the acoustic manipulation of the stimuli and strong interactions between *test*group* and *test* and participants' *gaming score*, with more gaming experience actually prolonging reaction times in the game (see Table 2). Post-hoc pairwise comparisons with Tukey HSD Tests were performed on the factors in the fixed effects of the linear mixed model. For the interaction of *group* and *test* all between- and within-group comparisons reached significance, indicating that subjects in both groups and tests responded to the stimuli with differing RTs. The GLM for *accuracy* shows an effect for test type (i.e. a considerable negative effect for the *game*), and a main negative effect of perceived difficulty of the experiment. The subjective evaluation of the game's difficulty level seems to correlate with an actual decrease in accuracy for the classic test, which is, however, reversed for the game. The significant interaction of *group* and *test* confirms that G2 performed better in the game than G1. There also is a small bias for *fun* in favor of the game, mediated by group (post-hoc Tukey: game(g2)-classic(g1), diff 0.24673487, p adj. = 0.0055663).

**Table 1** – Proportion correct responses (accuracy) and reaction times (sec) in both tests and groups, without the training phase.

| Group | Test | Accuracy Mean | Accuracy SD | RT Mean (sec) | RT SD |
|---|---|---|---|---|---|
| 1 | game | 0.42 | 0.49 | 3.09 | 1.60 |
| 1 | classic | 0.69 | 0.46 | 3.96 | 1.22 |
| 2 | classic | 0.80 | 0.40 | 3.30 | 1.05 |
| 2 | game | 0.76 | 0.43 | 1.54 | 0.68 |

A further analysis focused on the performance of both groups on their respective first test – *Time 1* – treating the game for G1 and the categorization test for G2 as two conditions of one variable, since the subjects had no knowledge as to the nature of the target cues prior to Time 1. The difference in accuracy on the first performed test per group was significant (compare Table 1, *Wilcoxon Rank Sum*: W = 23705, p-value <2.2e-16) – G2 was better able to correctly categorize the stimuli in the perception test than G1 was in the game. The same was true for Time 2 – G2 playing the game (76% correct) outperformed G1 completing the perception test with 69% correct (W = 23705, p-value <2.2e-16). For the logged RTs, the differences between both tests at Time 1 (*Tukey multiple comparisons of means*: diff. 0.156863, $p$ <1.04e-05) and

**Table 2** – Fixed factors in lmer: *RTlog ~test * (group + GamingScore) + fun + manipulation + correct + (1|stimulus) + (1|subject)*. Random effects: *stimulus (Intercept)*, var. 0.0032, SD 0.0564; *subject (Intercept)*: var 0.0324, SD 0.1801; *resid.*: var. 0.1100, SD 0.3317.

| | **Estimate** | **Std. Error** | **df** | **t value** | **Pr(>\|t\|)** |
|---:|---:|---:|---:|---:|---:|
| (Intercept) | 1.82 | 0.17 | 44.12 | 10.90 | 0.00 |
| testgame | -0.05 | 0.07 | 1081.66 | -0.68 | 0.49 |
| group | -0.14 | 0.08 | 24.47 | -1.80 | 0.08 |
| GamingScore | -0.05 | 0.03 | 24.63 | -2.02 | 0.05 |
| fun | -0.07 | 0.02 | 427.61 | -3.29 | 0.00 |
| manipulationF2 | 0.21 | 0.05 | 26.21 | 4.07 | 0.00 |
| manipulationFRIC | 0.11 | 0.05 | 31.66 | 2.26 | 0.03 |
| manipulationOriginal | 0.14 | 0.04 | 28.18 | 3.35 | 0.00 |
| correct | -0.07 | 0.02 | 1091.13 | -2.90 | 0.00 |
| testgame:group | -0.47 | 0.04 | 1067.83 | -11.69 | 0.00 |
| testgame:GamingScore | 0.07 | 0.01 | 1063.90 | 5.27 | 0.00 |

**Table 3** – Output of GLM *binomial* for the proportion of correct responses (accuracy) in both tests and groups.

| | **Estimate** | **Std. Error** | **z value** | **Pr(>\|z\|)** |
|---:|---:|---:|---:|---:|
| (Intercept) | 1.3275 | 0.4327 | 3.07 | 0.0022 |
| test(game) | -3.2427 | 0.5900 | -5.50 | 0.0000 |
| group | 0.3055 | 0.2127 | 1.44 | 0.1509 |
| difficulty | -0.3399 | 0.0902 | -3.77 | 0.0002 |
| test(game):group | 1.2073 | 0.2852 | 4.23 | 0.0000 |
| test(game):difficulty | 0.3622 | 0.1229 | 2.95 | 0.0032 |

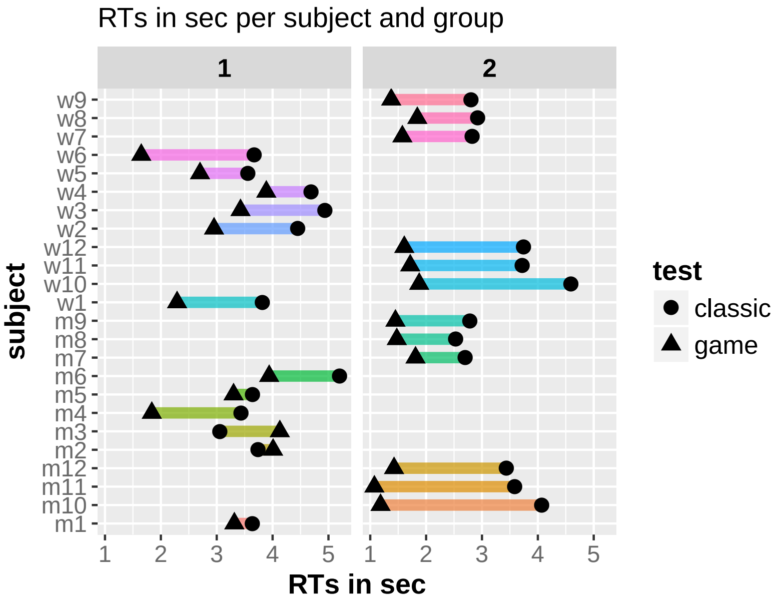

**Figure 1** – RTs in seconds in both tests per subject and group.

at Time 2 were significant (diff. -0.977677, $p$ = 0.000). Figures 1 and 2 display the individual differences in performance of our subjects in both tests and groups.

# 4 Discussion and Conclusion

Group 1 performed on average worse than Group 2 in both the game, and also, counterintuitively, in the categorization test. However, they also showed more variance throughout, as

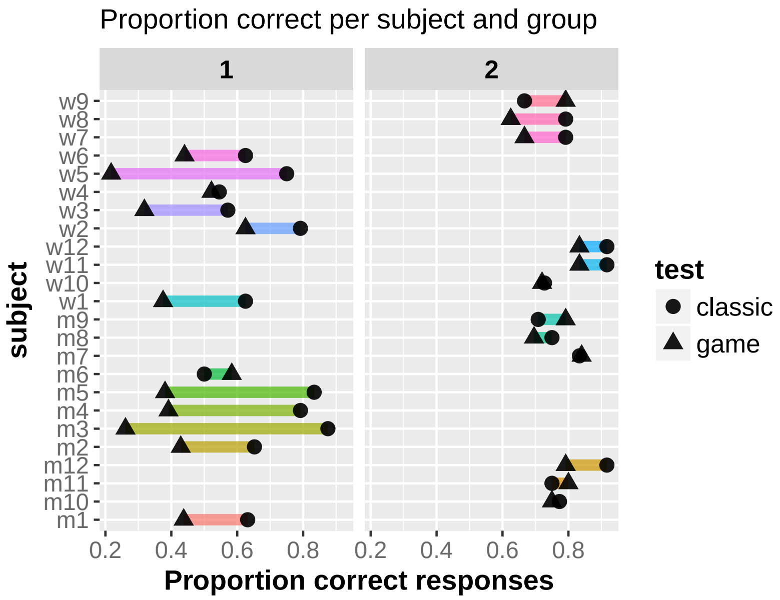

**Figure 2** – Proportion correct responses in both tests per subject and group.

expected. If we consider the three highest individual scores in the game in G1 (0.62, 0.58 and 0.52), achieved without any prior knowledge about the nature of the target cues and very limited time of trials in one level for correct categorization to occur, it seems that learning was indeed successful for a small number of subjects, supporting our claim about the importance to consider individual differences (IDs) in performance. Interestingly, the qualitative analysis of the post-hoc questionnaires revealed that more subjects from G1 actually correctly recalled words and sentences from the game and perception test than G2 from their experiments, which supports the claim that they were considerably more engaged in in-depth semantic processing of the stimuli in contrast to a shallow processing at a pure phonetic level. This is in line with Hawkins [23] stating that the natural tendency in communication is listening for meaning and this does not require a full acoustic analysis of the signal. The degree to which an additional in-depth acoustic analysis of a perceived stimulus (= above the threshold necessary for understanding a message) occurs, is highly individual. It is also possible that IDs in attention-switching skills would allow a more apt listener to more effectively switch between dimensions of an incoming speech signal (e.g. meaning vs. sound) [24], or to weigh the cues more appropriately depending on the current situation [25]. This would, combined with a naturally higher sensitivity for phonetic information in speech (i.e. phonetic *talent* [26]), also cause better attention-switchers to disengage from unsuccessfull strategies faster and turn toward the currently more adequate one. This may also shed light on the initially surprising worse performance of G1 in the categorization test (Table 1), given that they completed it after the computer game and could have been expected to at least perform similarly to G2 in the absence of any learning effects during the game, and not be negatively influenced by it. Overcoming the initial drive to focus on meaning rather than on sound and figuring out the cues to more robustly categorize the stimuli though, potentially takes longer than was allocated to the respective game levels in the current study, as previous test runs with the same game engine provide higher benchmarks for accuracy [17]. Thus, for some subjects in G1, their strategies from the game might have still prevailed during the perception test, where they listened rather for content-related than for acoustic cues, or tested further alternative hypotheses. A longer experimental time is also expected to further pronounce the fun-bias in favor of the game that could already be observed here. A more detailed look at the data does also reveal that at an individual level, G1's accuracy on the perception test actually comes close to the highscores of Group 2 (three best in G1 on classic test: 0.79, 0.83, 0.88, highscore G2: 0.92), again pointing to the fact that IDs should always be considered.

Results show, as expected, a considerable individual variation between subjects in the game. Additionally, again rather non-surprisingly, players seem to have devoted more attention to semantic information than to phonetic detail in the game environment than in the perception test. This, however, calls to question the validity of perception test data (and foremost the magnitude of the observed effects) obtained by explicitly making the subjects aware of the need to pay attention to the phonetic detail present, since it might not at all be reflective of their natural attention paid to fine phonetic detail in everyday communication. To allow for a fully-fledged comparison between the two types, the experiments would need to be further extended, to include a longer game play time and the categorization of more stimuli. Adding a third condition (with participants knowing prior to the game that *sound* is the target cue but are not familiar with the exact nature of the manipulation), which would correspond to an information level closer to the classic categorization experiment setting, might bring forward a more comprehensive comparison of the effects on attention towards phonetic-acoustic cues.

### Acknowledgments

Parts of this work are based on the thesis of Jenny Elisabeth Krüwald, supervised by Daniel Duran at the Institute for Natural Language Processing at the University Stuttgart. This work has been supported by the *DFG*, through the project A4 of the SFB732 at the University of Stuttgart.

## References

[1] WADE, T. and L. 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), pp. 2618–2633, 2005.

[2] MCPHERSON, J. and N. R. BURNS: *Assessing the validity of computer-game-like tests of processing speed and working memory. Behavior Research Methods*, 40(4), pp. 969–981, 2008.

[3] LIM, S.-J. and L. L. HOLT: *Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. Cogn Sci*, 35(7), pp. 1390–1405, 2011.

[4] LINDSTEDT, J. K. and W. D. GRAY: *Meta-T: Tetris® as an experimental paradigm for cognitive skills research. Behavior Research Methods*, 47(4), pp. 945–965, 2015.

[5] GRAY, W. D.: *Game-XP: Action Games as Experimental Paradigms for Cognitive Science. Topics in Cognitive Science*, 9(2), pp. 289–307, 2017.

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

[7] WASHBURN, D. A.: *The games psychologists play (and the data they provide). Behavior Research Methods, Instruments, & Computers*, 35(2), pp. 185–193, 2003.

[8] ERIKSEN, B. A. and C. W. ERIKSEN: *Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys*, 16(1), pp. 143–149, 1974.

[9] LU, C.-H. and R. W. PROCTOR: *The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. Psychon B Rev*, 2(2), pp. 174–207, 1995.

[10] ROGERS, R. D. and S. MONSELL: *Costs of a predictible switch between simple cognitive tasks*. *J Exp Psychol Gen*, 124(2), p. 207, 1995.

[11] WYLIE, G. and A. ALLPORT: *Task switching and the measurement of "switch costs"*. *Psychol Res*, 63(3-4), pp. 212–233, 2000.

[12] REITAN, R. M.: *Validity Of The Trail Making Test As An Indicator Of Organic Brain Damage*. *Percept Mot Skills*, 8(7), p. 271, 1958.

[13] MORGAN, A. L. and J. F. BRANDT: *An auditory Stroop effect for pitch, loudness, and time*. *Brain and Language*, 36(4), pp. 592–603, 1989.

[14] IAN H. ROBERTSON, V. R. I. N.-S., TONY WARD: *Test of Everyday Attention (TEA): Manual*. Thames Valley Test Company, Bury St. Edmunds, UK, 1994.

[15] DARCY, I., J. C. MORA, and D. DAIDONE: *Attention control and inhibition influence phonological development in a second language*. *Concordia Working Papers in Appl Ling*, pp. 115–129, 2014.

[16] SCHWEITZER, A., N. LEWANDOWSKI, and D. DURAN: *Attention, please! Expanding the GECO Database*. In *Proceedings of the 18th ICPhS*. Scottish Consortium of ICPhS, Glasgow, 2015.

[17] DURAN, D., N. LEWANDOWSKI, and A. SCHWEITZER: *A 3d computer game for testing perception of acoustic detail in speech*. In *Proceedings of Meetings on Acoustics (POMA)*, vol. 28:60004. Acoustical Society of America, 2016.

[18] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[19] WICKHAM, H.: *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. R package version 1.2.1.

[20] SINGMANN, H., B. BOLKER, J. WESTFALL, and F. AUST: *afex: Analysis of Factorial Experiments*, 2018. R package version 0.19-1.

[21] KUZNETSOVA, A., P. B. BROCKHOFF, and R. H. B. CHRISTENSEN: *lmerTest package: Tests in linear mixed effects models*. *J of Stat Software*, 82(13), pp. 1–26, 2017.

[22] WICKHAM, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. URL `http://ggplot2.org`.

[23] HAWKINS, S.: *Roles and representations of systematic fine phonetic detail in speech understanding*. *J Phon*, 31(3-4), pp. 373–405, 2003.

[24] SCHARENBORG, O., A. WEBER, and E. JANSE: *The role of attentional abilities in lexically guided perceptual learning by older listeners*. *Atten Percept Psychophys*, 77(2), pp. 493–507, 2015.

[25] MATTYS, S. L. and L. WIGET: *Effects of cognitive load on speech recognition*. *J Mem Lang*, 65(2), pp. 145–160, 2011.

[26] LEWANDOWSKI, N.: *Talent in nonnative phonetic convergence*. Dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2012. doi:10.18419/opus-2858.