Does Signal Degradation Affect Top–Down Processing of Speech?

A. Wagner, C. Pals, C. de Blecourt, A. Sarampalis and D. Başkent

Abstract Speech perception is formed based on both the acoustic signal and listeners' knowledge of the world and semantic context. Access to semantic information can facilitate interpretation of degraded speech, such as speech in background noise or the speech signal transmitted via cochlear implants (CIs). This paper focuses on the latter, and investigates the time course of understanding words, and how sentential context reduces listeners' dependency on the acoustic signal for natural and degraded speech via an acoustic CI simulation.

In an eye-tracking experiment we combined recordings of listeners' gaze fixations with pupillometry, to capture effects of semantic information on both the time course and effort of speech processing. Normal-hearing listeners were presented with sentences with or without a semantically constraining verb (e.g., crawl) preceding the target (baby), and their ocular responses were recorded to four pictures, including the target, a phonological (bay) competitor and a semantic (worm) and an unrelated distractor.

A. Wagner (✉) · C. Pals · D. Başkent
Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
e-mail: a.wagner@umcg.nl

C. Pals
e-mail: pals.carina@gmail.com

D. Başkent
e-mail: dbaskent@gmail.com

A. Wagner · C. de Blecourt · A. Sarampalis · D. Başkent
Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands

C. de Blecourt
e-mail: cdeblecourt@hotmail.com

A. Sarampalis
Department of Psychology, University of Groningen, Groningen, The Netherlands
e-mail: a.sarampalis@rug.nl

P. van Dijk et al. (eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing,* Advances in Experimental Medicine and Biology 894,
DOI 10.1007/978-3-319-25474-6_31

The results show that in natural speech, listeners' gazes reflect their uptake of acoustic information, and integration of preceding semantic context. Degradation of the signal leads to a later disambiguation of phonologically similar words, and to a delay in integration of semantic information. Complementary to this, the pupil dilation data show that early semantic integration reduces the effort in disambiguating phonologically similar words. Processing degraded speech comes with increased effort due to the impoverished nature of the signal. Delayed integration of semantic information further constrains listeners' ability to compensate for inaudible signals.

Keywords Speech perception · Degraded speech · Cochlear implants

1 Introduction

Processing of speech, especially in one's native language, is supported by world knowledge, the contextual frame of the conversation, and the semantic content. As a consequence, listeners can understand speech even under adverse conditions, where it is partially masked or degraded. Access to these signal-independent sources of information can, however, be compromised if the entire speech signal is degraded, rather than parts of it. This is the case for profoundly hearing impaired listeners who rely on the signal transmitted via a cochlear implant (CI) for verbal communication. Though CIs allow listeners to perceive speech, this remains an effortful task for them.

In optimal conditions, effortless processing of speech depends on the integration of analyses along a hierarchy of processing stages, as they are described in models of speech perception. These models differ in the way they view the spread of information across various analysis stages (e.g. TRACE: McClelland and Elman 1986; Shortlist: Norris 1994; Shorlist B: Norris and McQueen 2008), but they do agree on the presence of lexical competition. Lexical competition is the process through which listeners consider all the mental representations that overlap with the heard signal as candidates for the word intended by the speaker. Before making a lexical decision listeners thus subconsciously consider multiple words, including homonyms (e.g., 'pair' and 'pear') and lexical embeddings (e.g., *paint* in *paint*ing). In optimal conditions, lexical competition is resolved (i.e. phonologically similar words are disambiguated) very early in the course of speech perception because listeners can rely on a plethora of acoustic cues that mark the difference between phonologically overlapping words (e.g., Salverda et al. 2003), and further also benefit from semantic information in sentences (Dahan and Tanenhaus 2004).

These models are based on data on natural speech perception in optimal conditions, so the question of how analysis of speech is affected by constant degradation of the signal remains unanswered. The present study investigates the time course of lexical competition and semantic integration when processing degraded speech. Furthermore this study will also query whether semantic integration can reduce the mental effort involved in lexical competition in natural and degraded speech. This question has not been studied before since understanding speech in optimal conditions is commonly perceived as effortless. To address these questions we will adapt the approach of Dahan and Tanenhaus (2004), and perform an eye tracking experiment in which listeners are presented with natural and degraded speech. We will further combine the recordings of gaze fixations with pupillometry to obtain a measure of processing effort.

Eye-tracking has been used to study the time course of lexical competition (e.g., Allopenna et al. 1998), since listeners' gazes to pictures on the screen reflect their lexical considerations during lexical access as they gradually match the heard signal to an object on the screen. To study the effort involved in processing speech we will record also listeners' change in pupil size. Pupil dilation is a measure that has been used to study effort involved in solving various cognitive tasks (e.g., Hoeks and Levelt 1993). An increase in pupil dilation has also been shown for listeners presented with degraded speech relative to highly intelligible speech (e.g., Zekveld et al. 2014). Pupil dilation reflects next to adaptations to changes in luminance or lightness, occurring within the timescale of 200–500 ms, also a slower evolving response to mental effort, in the timescale of about 900 ms (Hoeks and Levelt 1993).

2 Methods

2.1 *Participants*

Twenty-eight native speakers of Dutch, aged between 20 and 30 years (mean = 26), participated in this experiment. None of the participants reported any known hearing or learning difficulties. Their hearing thresholds were normal, i.e. below 20 dB HL on audiometric frequencies from 500 to 8000 kHz. All the participants signed a written consent form for this study as approved by the Medical Ethical Committee of the University Medical Center Groningen. The volunteers received either course credit or a small honorarium for their participation.

2.2 *Stimuli*

AQ1

The set of stimuli consisted of the materials used by Dahan and Tanenhaus (2004), and an additional set constructed analogously, resulting in a total of 44 critical items. The critical items were quadruplets of nouns, which were presented together as pictures on the screen. To study the time course of lexical competition we created pairs of critical Dutch words with phonological overlap at the onset, e.g., the target 'pijp' [pipe] was combined with the phonological competitor 'pijl' [arrow]. To study whether disambiguating semantic context reduces lexical competition between acoustically similar words, the two phonologically similar items were presented within sentences, in which a verb that was coherent with only one of these two nouns (e.g. 'rookte' [smoked]) either preceded or followed the noun. The critical pair was presented as pictures together with two Dutch nouns, of which one was semantically viable to follow the verb (e.g., 'kachel' [heater]), the semantic distractor, and the other a phonologically and semantically unrelated distractor ('mossel' [mussel]).

Next to the critical items we constructed 60 sets of filler items. The verbs used in all of these filler sentences were coherent with two nouns, the target and the semantic distractor. The filler items were also presented in quadruplets, and the two remaining distractor nouns were not semantically coherent subjects for the verb. To create a balance between the critical and the filler items, in 20 of the filler items the distractor nouns were phonologically overlapping at the onset. The remaining 40 sets of distractors were phonologically unrelated.

All sentences began with a prepositional phrase, such as "Never before." or "This morning." The sentences were recorded from a male native speaker of Dutch. Black and white drawings were created as display pictures, specifically for the purpose of this study.

Two listening conditions were used in the experiment; natural speech (NS) and degraded speech (DS). The degraded stimuli were created using a noise-band-vocoder to simulate CI processing. The stimuli were first bandlimited to 80–6000 Hz, and were subsequently bandpass-filtered into 6 channels. Sixth order Butterworth filters were used, with a spacing equal to the distances in the cochlea as determined using the Greenwood function. The slow-varying amplitude envelopes were extracted from each channel via lowpass filtering, and these envelopes were then used to modulate carrier wideband noise, the resulting 6 channels were finally bandpass filtered once more using the same 6 bandpass filters. The processed stimuli were the summed signals from the output of all channels. This manipulation lead to stimuli with unnatural spectrotemporally degraded form, hence stimuli that simulate the signal conveyed via CIs.

2.3 Procedure

Before data collection, participants were familiarized with the pictures and the nouns that refer to the pictures. They were then seated in a comfortable chair facing the monitor, and an Eyelink 500 eye-tracker was mounted and calibrated. This head mounted eye-tracker contains two small cameras, which can be aligned with the participants' pupil to track the pupil's movements and size continuously during the experiment. Pupil size was recorded together with gaze fixations using a sampling rate of 250 Hz.

The stimuli were presented via a speaker in sound attenuated room. The lighting in this room was kept constant throughout the experiment to avoid effects of ambient light intensity on the pupil diameter. The participants' task was to listen to the stimuli and to click on the picture corresponding to the target noun in the sentence.

Each participant was presented with stimuli blocked into an NS and DS condition. Before the DS condition, the participants were familiarized with the degradation used in this study by listening to 30 degraded sentences and selecting the correct one from a set of sentences presented on the screen.

Each experimental item was presented only once in either the context or neutral sentence, and in either NS or DS. Between the two blocks (NS and DS) there was a break. Four practice trials preceded each block (using filler items), and a block consisted of 48 experimental items; 22 critical items and 26 filler items. The order of the presentation between blocks and items was quasi-random.

2.4 Analysis

Trials in which participants clicked on the wrong picture were excluded from the analysis. Trials with eye blinks longer than 300 ms were also excluded. Shorter blinks were corrected for by means of linear interpolation.

2.4.1 Gaze Fixations

To address the question of how semantic context affects lexical competition between phonologically similar words the statistical analyses focus on listeners' gaze fixations towards the phonological competitor and the semantic distractor. The probabilities of gaze fixations towards this competitor and this distractor were statically analyzed by means of growth curves (Mirman 2014). R (R Core team 2013) with lme4 package (Bates et al. 2014) was used to model the time curves of fixations as 4th order polynomials within the time window of 200–2000 ms after word onset. Two logistic-regression multi-level models were used, with fixations to either the phonological competitor or the semantic distractor, coded as a binomial response. The time course curves were described in four terms: intercept, the overall slope of the curve, the width of the rise and fall around the inflection, and the curvature in the tails. The probability of fixations along the time course was modeled as a function of Context (neutral versus context), Presentation (NS versus DS) and the possible three-way interactions between these two factors and all four terms describing the curves. As random effect, we included individual variation among participants and items on all four terms describing the time curve. Model comparison was used to estimate the contribution of individual predictors to the fit of the model. For this, individual fixed effects were sequentially added, and the change in the model fit was evaluated by means of likelihood ratio test.

2.4.2 Pupil Dilation

To investigate the effort involved in the process of lexical competition with and without semantic context, the pupil dilation data per participant were baseline-corrected to the 200 ms preceding the presentation of the experimental item. The baseline-corrected data were normalized to correct for individual differences in pupil size, according to the equation:

$$\%\,Event\ Related\ Pupil\ Dilation = (observation - baseline)\,/\,baseline * 100.$$

For the statistical analysis, the pupil size data, as captured by the event-related pupil dilation (ERPD), were analyzed analogously to the fixation data, as time curves of pupil dilation. The time-course functions were analyzed as 3rd -order polynomials, since, during fitting, the fourth order turned out to be redundant to the description of these curve functions. The terms describing the curves are: intercept, the slope of the curve, and a coefficient for the curvature around the inflection point. These time curves were analyzed by means of multi-level nonlinear regression model. The statistical models contained in addition to the terms describing the curves per participant also random effects on these three terms per participant, and for the phonological competitor model also random effects per item.

3 Results

3.1 Gaze Fixations

Figure 1 displays the time curves of fixations to all four pictures displayed within the NS blocks for C (a), and N (b), and for the DS blocks for C (c) and N (d). These figures show proportions of fixations to the four pictures displayed, averaged across participants, and the 95 % confidence intervals for the fixations to the target and competitor.

Of particular interest for this study are the three-way interactions between Context (C versus N) and Presentation (NS versus DS) and the terms describing the course of the curves. For the fixations to the phonological competitor, as significant emerged the three way interactions with the first term (the intercept) of the curve ($\chi2(18)=28476$, $p<0.001$, the interaction with the quadratic term (the slope), ($\chi2(18)=28184$, $p<0.001$), the interaction between the cubic term (rise and fall around the central inflection), ($\chi2(18)=27632$, $p<0.001$), and the quartic term (curvature in the tails), ($\chi2(18)=27651$, $p<0.05$). The interaction with the intercept shows that the context sentences reduced the area under the fixation curves to the competitor for NS (red lines in Fig. 1a versus b), and that this reduction was smaller for DS (red lines in Fig. 1c versus d). The interaction with the slope shows that the growth of fixations to the competitor is shallower for DS in the neutral context than it is for NS in neutral context. The interaction with the cubic term reflects that the location of the peak of fixations towards the competitor in DS is delayed for about 300 ms relative to the location of the peak for NS, and that the course of this curve is more symmetric than for NS, and this mainly for the items presented in neutral

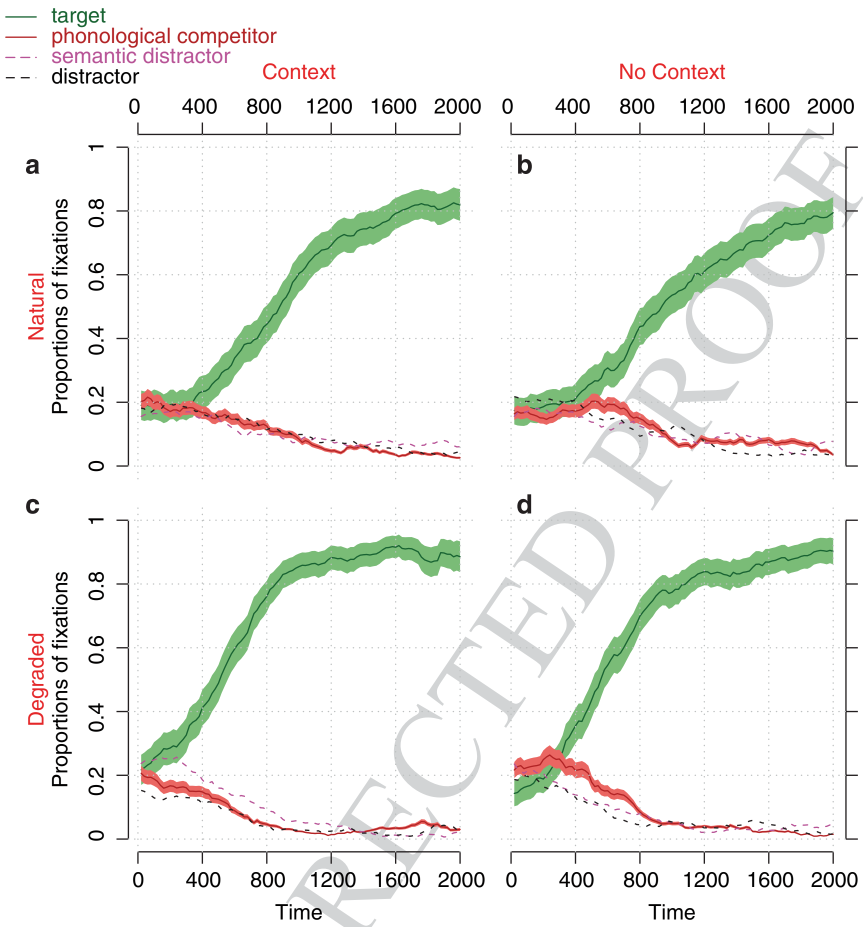

Fig. 1 Time curves of fixations to the pictures displayed for NS (**a** & **b**) and DS (**c** & **d**), and for items presented in context sentences (**a** & **c**), and neutral sentences (**b** & **d**)

context. The interaction with the quartic term reflects a slower decline of fixations towards the competitor in DS versus NS, and shallower for items in context than in neutral sentences.

For the fixations to the semantic distractor, as significant emerged the interactions between Context and Presentation and the intercept of the curve ($\chi 2(3) = 2268.6$ $p < 0.001$), the interaction with the quadratic term ($\chi 2(3) = 337.25$, $p < 0.001$), the interaction between the cubic term, ($\chi 2(3) = 69.41$, $p < 0.001$), and the quartic term ($\chi 2(3) = 19.09$, $p < 0.05$). These interactions reflect what can also be seen in a comparison of between NS and DS in Fig. 1. Namely that in NS, listeners fixate the semantic competitor more often in the context sentences than in neutral context. This effect is absent for DS.

3.2 *Pupil Dilation*

Figure 2 displays the time course of pupil dilation for NS and DS and for the two contexts C and N.

The curves for NS and DS in the neutral condition show a constant increase in pupil size over time as a function of lexical competition. The curves for the context condition show a decline in pupil size growth starting at around 800 ms after the onset of the target word. The statistical analysis revealed significant three way interactions with Context (N versus C) and Presentation (NS versus DS) on all terms describing the curves: Intercept ($\chi 2$ (3)$=301.90$, $p<0.001$), slope ($\chi 2$ (3)$=145.3$, $p<0.001$), and the cubic term, the curvature around the peak ($\chi 2$ (3)$=272.52$, $p<0.001$). This implies that pupil dilation was sensitive in capturing the reduced effect of lexical competition in the context sentences versus neutral context, but this effect was delayed and smaller in DS than in NS.

A look at this figure suggests that the effort involved in lexical competition for DS was overall smaller for DS than for NS. This overall smaller increase in pupil dilation can be explained by the fact that these curves are normalized to a baseline

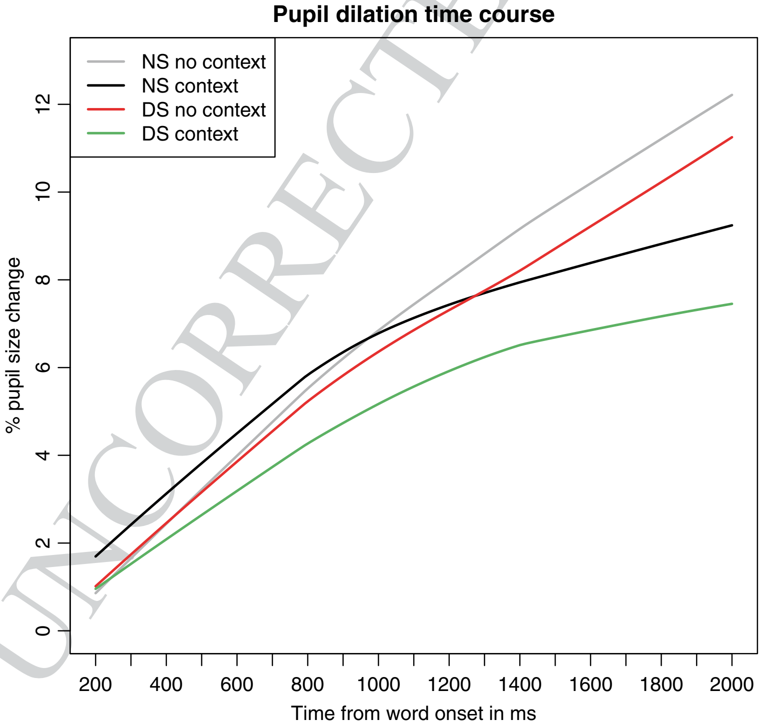

Fig. 2 Time curves of pupil dilation, averaged across participants for NS (*black* and *grey*) and DS (*red* and *green*), and for items presented in context sentences (*black* and *green*), and neutral sentences (*red* and *grey*)

of 200 ms preceding the presentation of each critical item per trial, participant and condition. Listening to degraded speech is by itself more effortful than listening to natural speech (e.g., Winn et al. 2015), and therefore there is a difference in the baseline between DS and NS. These differences in the baseline can be explained by the difference in processing degraded versus natural speech, and are independent of the effects of semantic integration on lexical competition.

4 Discussion

This present study examined the effect of semantic integration on the time course of lexical competition, and on the effort involved in solving lexical competition in natural and degraded speech. Our results show that processing natural speech comes with a timely integration of semantic information, which in turn reduces lexical competition. Listeners are then able to pre-select a displayed target based on its semantic coherence with the context, and this allows listeners to reduce the effort involved in lexical competition. When processing degraded speech the integration of semantic information is delayed, as is also lexical competition. This implies that semantic integration is not able to reduce lexical competition, which by itself is longer and occurs later. These results were also mirrored by the pupil dilation data, in which a release from lexical competition was visible but delayed. Mapping of degraded speech to mental representations is more effortful due the mismatch between the actual signal and its mental representation, and lexical context is not able to release listeners from this effort on time. In natural situations, in which words are being heard in succession, and the speech signal evolves quickly over time, such a difference in processing speed of degraded speech will accumulate effort, and draw more strongly on resources in working memory.

Acknowledgments We would like to thank Dr. Paolo Toffanin for technical support, and Prof. Frans Cornelissen (University Medical Center Groningen) for providing the eye-tracker for this study. This work was supported by a Marie Curie Intra-European Fellowship (FP7-PEOPLE-2012-IEF 332402). Support for the second author came from a VIDI Grant from the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw) Grant No. 016.093.397. The study is part of the research program of our department: Healthy Aging and Communication.

References

Allopenna PD, Magnuson JS, Tanenhaus MK (1998) Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J Memory Lang 38:419–439

Bates D, Maechler M, Bolker B, Walker S (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. http://CRAN.R-project.org/package=lme4

AQ2

Dahan D, Tanenhaus MK (2004) Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. J Exp Psychol Learn Mem Cogn 30:498–513

Hoeks B, Levelt W (1993) Pupillary dilation as a measure of attention: a quantitative system analysis. Behav Res Methods 25(1):16–26

McClelland JL, Elman JL (1986). The TRACE model of speech perception. Cognitive Psychology 18:1–86

Mirman D (2014) Growth curve analysis and visualization using R. Chapman and Hall/CRC, Florida

Norris D (1994) Shortlist: a connectionist model of continuous speech recognition. Cognition 52:189–234

Norris D, McQueen JM (2008) Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev 115(2):357–395

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Salverda AP, Dahan D, McQueen JM (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90:51–89

Winn MB, Edwards JR, Litovsky RY (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear and Hear (ahead of print)

Zekveld AA, Heslenfeld DJ, Johnsrude IS, Versfeld N, Kramer SE (2014) The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. Neuroimage 101:76–86

Chapter 31: Author Query

AQ1. “Dahan and Tananhaus 2004” was changed to “Dahan and Tanenhaus 2004” to match the reference list. Please check.

AQ2. We have updated the publisher’s location of the reference “Mirman 2014”. Please check.