Comprehension of degraded speech: Exploring the role of attention and speed of processing in top-down prediction

Pratik Bhandari

Department of

Saarland University

A thesis submitted for the degree of

Doctor of Philosophy

xxxx 2021

Dedicated to ...

Acknowledgements

Here I acknowledge lots of people including my GP, neurologists and counselors.

Pratik Bhandari
Universität des Saarlandes,
Saarbrücken
10 March 2021

Abstract

Listening in an adverse environment poses a challenge – it is difficult to understand what is being said when there is background noise, or when the speaker's speech signal is distorted. Nevertheless, listeners show remarkable success in understanding the distorted speech by utilizing context information to form predictions about upcoming linguistic events. The extent to which such top-down predictions are useful is still a matter of debate. Additionally, the role of other factors such as attention and rate of flow of information in degraded speech comprehension are understudied. In this thesis, I present a broader overview of the role of semantic predictions on degraded speech comprehension across lifespan, and how it interplays with attention, and rate of flow of information. In the first experiment (Chapter 4.1), I show that listeners can flexibly pay attention to a portion of speech stream; and attending to the sentence context is necessary to utilize the context and form top-down predictions. I show in the second experiment (Chapter 4.2) that as listeners utilize the context information, they form semantic predictions about upcoming linguistic events in a graded manner when the speech is moderately degraded. Semantic predictions are not restricted to only most highly likely sentence endings. I also argue for a novel metric to measure language comprehension, and show that sensory adaptation to degraded speech is disrupted by change in higher-level semantic features of speech. Perception, processing and comprehension of degraded speech is difficult and effortful. In the third experiment (Chapter 4.3), I show that when the rate of flow of information is changed by increasing the speed of speech, the facilitatory effect of predictability is observed even at a mildly degraded speech. That is, earlier (in Chapter 4.2), mild degradation was easier to process; but increase in speed rendered the speech processing difficult such that predictability had a facilitatory effect. To examine the general age differences in the facilitatory effect of predictability, I conducted a fourth experiment (Chapter 4.4) where younger adults (age range 18-30) and older adults (age range = ...) were recruited. Highest age difference in the use of sentence context for language comprehension was observed at the moderate level of speech degradation. This supported the hypothesis that sensory decline with aging tips older adults to rely on context information more than younger adults do. I show the neural markers of these age differences in the fifth experiment (Chapter 4.5) ... ... Taken together, this thesis and the

results herein support the views that sentence context and semantic prediction facilitate comprehension of moderately degraded speech in a graded manner, and reliance on context increases with age.

Contents

List of Figures

List of Tables

List of Abbreviations

HP, MP, LP . High-, Medium-, or Low-predictability

YA, OA . . . . Younger, or Older adults

ch . . . . . channels

1 General Introduction

Contents

1.1 Overview of the thesis

1.2 Theories of language comprehension

1.2.1 Predictive language processing

1.3 Speech degradation

Speech can be distorted by variability in speakers' production, like, accented speech, soft/rapid speech, or it can arise from listener-related factors like, hearing loss, auditory processing disorder. It can also be a result of noise from transmission, like ambient noise, or distortion in the transmission (e.g., telephone line). All these

sources of distortion make listening condition adverse. In laboratory setup, the effect of speech distortion, and the mechanism of listening in adverse listening condition is studied using artificial distortion of speech. For example, white or pink noise signal is superimposed on the top of speech signal so as to add a background noise.

In the early 1950s, plastic tapes with magnetic recorder were mechanically cut, spliced and pasted (named as *chop-splice method*) to increase the rate of speech (**Garvey1953**). Such a method was developed used to overcome the effect of frequency shift on intelligibility that was an undesired result of accelerated speech recorded on sound film or discs in the late 1920s (Fletcher 1929). In recent days, algorithms like Pitch Synchronous Overlap-Add Technique [PSOLA; Charpentier and Stella (1986); Moulines and Charpentier (1990)] and Overlap-Add Technique Based on Waveform Similarity [WSOLA; Verhelst and Roelands (1993)] are used to compress or elongate an auditory signal so as to change the rate of speech. These methods preserve the phonemic properties of speech to a large extent (see Section X.X.X).

In the same vein, noise-vocoding is used to remove the spectral detail of the speech signal only leaving its temporal and periodicity cues (see Section X.X.X). Noise-vocoding was initially developed as a means to reduce the information in speech signal to be transmitted through the telephone line (Dudley 1939; "The vocoder" 1940) — [Re-read this thoroughly]. Shannon and colleagues later used the same technique as an analogue to cochlear implant (Shannon, Zeng, et al. 1995; Loizou et al. 1999; Shannon, Fu, et al. 2004) – number of channels used in a cochlear implant are similar to the number of noise-vocoding channels in terms of their speech output and intelligibility [... cite probably Wagner et al. ...].

1.4 Comprehension of degraded speech

The first factor that determines the intelligibility of noise-vocoded speech is the number of channels. With an increase in noise-vocoding channels, speech intelligibility increases (e.g., Davis et al. 2005; Shannon, Zeng, et al. 1995). For example, speech processed through 8 channels noise vocoding is more intelligible than the speech processed through 4 channels noise vocoding (Loizou et al. 1999). Participant related variables (age, vocabulary), test materials (words, sentences, accented speech), and listening conditions (quiet, background noise) also influence the intelligibility of noise vocoding speech. Accuracy is higher for sentences than for words in isolation (CITE). Compared to quiet, response accuracy was reduced when listeners were presented with vocoded speech in the presence of a background noise. At the same level of degradation, accuracy is higher for younger adults than for older adults, i.e., with age, keeping all other variables constant, comprehension of degraded speech decreases. Other factors like sentence context and vocabulary also play a role, which will be discussed below. In sum, comprehension of degraded speech is a not only the amount of spectral details available, but also other listener and speaker related factors. How a listener utilizes available context information to 'make-up' for the impoverished auditory information is the critical factor determining intelligibility and comprehension of degraded speech.

1.4.1 Role of sentence context

Literature from sentence reading provides us with an insight how readers use the information available as the words are presented to them to make predictions about what word they'll see next. In visual world paradigm, Altmann and Kamide (1999) showed that a listeners predict upcoming word of a sentence using the cue provide by the sentence context. For example, they presented participants a picture of four objects: cake, XXX, XXX, and XXX while the participants were listening to the sentence '*The boy will eat the . . .*'. Even before hearing *cake*, participants fixated at the picture of cake. This finding has been replicated multiple times

[Kamide et al. (2003); Altmann and Kamide (2007); CITE other recent papers] in different languages (**MISHRA**). This is observable in behavioral measures as well as electrophysiological measures.

Kutas and Hillyard (1984) reported smaller N400 amplitude for highly probable sentence endings than for less probable sentence endings given different sentence contexts. They found that the N400 amplitude was more sensitive to sentence ending than to the constrain imposed by the preceding words. – Check new studies on this line – DeLong et al. (2005) showed that listeners form probabilistic predictions about upcoming words in a sentence. In a highly predictable sentence context like 'The day was breezy so the boy went outside to fly . . . ', N400 amplitude was much smaller for an expected continuation 'a kite' than for an unexpected continuation 'an airplane'. This study has been further replicated in Spanish-English bilinguals [CITE] showing that when presented with . . . . . Similarly, it has been observed that readers tend to skip the predictable words more than unpredictable words while reading, and predictable words are read faster than unpredictable words (Frisson et al. 2005; Rayner et al. 2011). Semantic context already provides listeners information about what the predictable word is going to be, therefore their fixation time on the predictable words is lesser than unpredictable words, and it takes lesser time to read the predictable words.

Taken together, these studies show that as the sentence unfolds, a human comprehender forms the meaning representation of the available context information, and generates prediction about what linguistic input is going to come next.

In a noisy environment, however, it is difficult to understand the context itself. While reading text that is visually degraded, readers rely on the available context information (e.g., Clark et al. 2021). Listeners generally rely on sentence context moreso in an adverse listening condition than in a clear listening environment. For example, Sheldon et al. (2008b) showed that when the speech signal is degraded, word recognition is improved by sentence context in both younger and older adults. They presented listeners with senteces with high and low context information,

noise-vocoded at different levels of spectral degradation. They found that response accuracy was higher for sentences with high context information than for the sentences with low context information. In cochlear implantees, XYZ et al., have shown that high predictability sentences result in higher accuracy than low predictability sentences in a word recognition task. Contrary to the findings of most of the studies using clean speech and reading clean text, sentence context does not *always* help in comprehension of all noise-vocoded speech. When the noise-vocoded speech is least degraded, listeners might rely mostly on the bottom-up auditory input than on top-down predictions generated from the sentence context; hence, context does not render benefit in this case. For example, in the 32-channels noise-vocoded speech in Obleser and Kotz (2009), sentence context, and consecutively top-down semantic prediction does not yield any benefit in sentence comprehension when compared to 4-channels noise-vocoded speech.

In summary, sentence context provides information necessary to generate top-down predictions. Listeners use this context information and form predictions to better comprehend degraded speech.

1.4.2 Effect of aging

1.5 Research motivation

2

General Methods

Contents

2.1 Stimulus sentences

..... - Sy How sentences were constructed. - Say how cloze probabilites were collected. - Say how 120 and then 360 auditory stimuli are created.

2.2 Speech processing

All 360 sentences used in the experiments in this thesis were first recorded and digitized at 44.1 kHz with 32 bit linear encoding. The quality of auditory stimuli – in chapters 5 through 8 – are manipulated by noise-vocoding (chapter 5 to 8), and speech compression algorithm PSOLA (chapter 7). These speech processing

... .

2.2.1 Noise-vocoding

Noise-vocoding is used to parametrically vary and control the quality of speech signal and a graded manner. It largely removes the spectral details of the speech signal but preserves the temporal and preiodicity cues (Rosen et al. 1999).

Noise-vocoding distorts speech by dividing a speech signal into specific frequency bands corresponding to the number of vocoder channels. The frequency bands are analogous to the electrodes of cochlear implant (Shannon, Zeng, et al. 1995; Loizou et al. 1999; Shannon, Fu, et al. 2004). The amplitude envelope – fluctuations of amplitude – within each band is extracted and is used to modulate noise of the same bandwidth. It renders vocoded speech harder to understand by replacing the fine structure of the speech signal with noise while preserving the temporal characteristics and periodicity of perceptual cues (Rosen et al. 1999).

If the cut-off frequencies of the bandwidth of the speech signal (i.e., the analysis band) and the bandwidth of the noise do not match then the resulting noise-vocoded speech becomes spectrally shifted (e.g., Faulkner et al. 2012). The cut-off frequencies of the speech signal and the to-be-modulated noisebands are identical for all the speech stimuli in the current study.

Sentences were noise-vocoded through 1-, 4-, 6- and 8-channels in Experiment 1, 2, and 4 using custom scripts originally written by **Darwin2005**. In Experiment 3, they were vocoded only through 4-channels. The cut-off boundary frequencies were set between 70 Hz and 9000 Hz. Upper and lower bounds for band extraction within each bandwidth of each noise-vocoding condition are shown in Table X.X which follows Greenwood's cochlear frequency position function (Greenwood 1990; Erb 2014). Scaling was performed to equate the root-mean-square values of the original undistorted signal and the final noise-vocoded sentences.

Spectrograms of clear speech and noise-vocoded speech (1-, 4-, 6- and 8-channels) for the word 'Aufgabe' are shown in Figure X.X. It shows that with a decrease in the number of noise-vocoding channels, speech signal becomes more and more similar to noise.

2.2.2 Speech compression

Uniform time-compression algorithms like pitch-synchronous overlap-add technique [PSOLA, Charpentier and Stella (1986); Moulines and Charpentier (1990)) are used to compress the speech signal and increase the rate of speech. PSOLA analyzes the pitch of an auditory signal, in the time domain of its digital waveform, to set pitch marks and then segments the signal into successive analysis windows centered around those pitch marks. To create synthesized speech, a new set of pitch marks are calculated and the analysis windows are rearranged. Depending on the time-compression factor, some analysis windows are deleted, and the remaining windows are concatenated by superimposing and averaging the neighboring analysis windows. Hence the resulting speech signal is compressed, i.e., it is perceived to be faster than the original speech [e.g., CITE] The distortion of phonemic properties of speech signals are minimal when accelerating and slowing down within the range of factor 2 or below (Moulines and Charpentier 1990). In Experiment 3, PSOLA algorithm in Praat software is used to increase the rate of speech by a factor of 0.65 before passing it through 4-channels noise-vocoding.

Spectrograms of clear speech, speeded speech, and noise-vocoded speeded speech for the word 'Aufgabe' are shown in Figure X.X. It shows that .....

2.3 Measurement of language comprehension

3

General data collection methods

Contents

3.1 Data collection in the laboratory

3.2 Online data collection

3.2.1 Designing experiments

3.2.2 Hosting platform

3.2.3 Recruiting participants

Contents

3.1 Data collection in the laboratory 9
3.2 Online data collection 9
3.2.1 Designing experiments 9
3.2.2 Hosting platform 9
3.2.3 Recruiting participants 9

4

General statistical approach

4.1 Linear regression

In this thesis, we use binomial mixed effects logistic regression models with crossed random effects (Baayen et al. 2008). These models are, simply speaking, extensions of logistic regression models. A logistic regression models a dependent variable (or an *outcome*, or a *response* variable) as a function of one or more independent predictor variables (or *factors*, or *explanatory* variables). That is, an outcome y is modeled as a function of explanatory variables $x_1, x_2, x_3..., x_n$, and an error term ε.

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon \tag{4.1}$$

The intercept α, and the regression coefficients β_1, β_2, and β_3 for each explanatory variable are estimated to achieve the model that best fits the data. Analysis of Variance (ANOVA) is a special case of logistic regression (Chatterjee in Jaeger's thesis page 40; Shravan's book and blog) that is one of the most common statistical tools in psychology and psycholinguistics [...CITE...]. These linear regressions as shown above (Equation 4.1) and ANOVAS however, are not well suited for categorical data like response to multiple choice questions or yes/no questions, confidence ratings, etc. For example, in all the experiments in the current thesis, the

response variables are response accuracy, given binary correct/incorrect responses. Output of linear regression model ranges from $+\infty$ to $-\infty$ while accuracy (or probability) ranges from, 0 to 1. Additionally, simple regression models do not take into account the variability across individual participants and items. These problems in psychological sciences and psycholinguistics research has been long pointed out as early as 19XX [. . . CITE language as a fixed effects fallacy. . .], and later [. . .]. They are addressed to some extent by binomial logistic regression, and for our purpose by incorporating mixed effects model to binomial logistic regression.

Below we briefly introduce binomial logistic regression and mixed effects model. Then we show a simple example of how binomial logistic mixed effects model is used in the experiments in this thesis.

4.2 Binomial logistic regression

The response variable in this experiments in this thesis are binary. Participants' written response to what they hear are coded as either correct or incorrect. A binomial logistic regression model is best suited for such a categorical data (Jaeger 2008). We will use the term logistic regression model and binomial logistic regression model interchangeably henceforth.

As the name suggests, the output variable in a logistic regression model is in logit scale. The model therefore predicts logits of an outcome variable. Logits are log with base e, i.e. ln.

Probability ranges from 0 to 1 only while *odds* ranges from 0 to $+\infty$. Fitting a linear regression model with probability or odds would assume the range to be between 0 and 1, and between 0 and $+\infty$ respectively. This restricts the range, and is incorrect for a linear model. Therefore, in a binomial logistic regression model, log-odds are used which range from $-\infty$ to $+\infty$.

A simple binomial logistic regression model is shown in Equation 4.2:

$$\ln(\frac{p}{1-p}) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon \tag{4.2}$$

This is equivalent to,

$$p = \frac{exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon)}{1 + exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon)} \tag{4.3}$$

$$= \frac{exp(\ln(\frac{p}{1-p}))}{1 + exp(\ln(\frac{p}{1-p}))} \tag{4.4}$$

where,

$$\ln(\frac{p}{1-p}) = logit(p) \tag{4.5}$$

Log-odds of correct response obtained from Equation 4.2 can be transformed to probability of correct response. Equations 4.4, and 4.5 provide the relationship between probability, logit (or log-odds), and odds ($\frac{p}{1-p}$)).

Some of the assumptions made for binomial logistic regression models are violated in our data. One of them being non-independence of observations, i.e., all data points are independent from one another. This assumption is violated in unbalanced design, and at times even for balanced design. Same participant responds multiple trials of same experimental condition within an experiment. Although the design itself is balanced, after removal of outliers and/or trials which are not appropriate for comprhension measures (see section XXX for details), number of trials in analyses are unequal for each participant, item, and experimental condition. This introduces a bias in the model [Jaeger2008; other papers on GLMM].

Another intrinsic property or feature of logistic regression is that it assumes a common mean for each predictors. It has been shown that this is in fact not true: the effect of a predictor can vary depending on different random variables like participants, or items. To account for these variances, mixed effects models are used. In recent days, these models are frequently used and advocated for by psycholingustis and statisticians [... cite ...].

4.3 Mixed effects modeling

To overcome the limitations of logistic models, like violation of assumption of non-dependence of observations, and to account for the variability in the subject and/or item related parameter, mixed effects models are used. Mixed effects models contain 1) both linear and logistic regressions, and 2) *fixed effects* and *random effects*, hence the name *mixed effects*. Fixed effects term, e.g., levels of degradation assumes that all levels of degradation used in the experiment are independent from one another and they share a common residual variance. The random effects term with only varying intercept, e.g., subject as intercept, assumes that if there are 100 subjects then the mean accuracy of those 100 subjects are only a subset of possible global accuracies drawn from a set of population mean. When a slope, e.g., levels of predictability, is included to the random effects structure in addition to the varying intercept (e.g., subjects), then the model assumes that the effect of predictability on response accuracy varies across subjects.

4.4 Binomial logistic mixed effects modeling

A binomial logistic mixed effects model with varying intercepts and slopes for items and subjects is shown in Equation 4.6 below.

$$\begin{aligned}\ln(\frac{p}{1-p}) = \alpha + u_\alpha + w_\alpha + (\beta_1 + u_{\beta_1} + w_{\beta_1}) \cdot x_1 + \\ (\beta_2 + u_{\beta_2} + w_{\beta_2}) \cdot x_2 + ... + (\beta_n + u_{\beta_n} + w_{\beta_n}) \cdot x_n\end{aligned} \tag{4.6}$$

where,

- α is the Intercept.
- Fixed effects: $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (or effects) of $x_1, x_2, ..., x_n$.
- $\boldsymbol{u} = \langle u_\alpha, u_{\beta_1}, u_{\beta_2}, ..., u_{\beta_n} \rangle$: Varying intercept and slopes for random effect term like, *subject*.
- $\boldsymbol{w} = \langle w_\alpha, w_{\beta_1}, w_{\beta_2}, ..., w_{\beta_n} \rangle$: Varying intercept and slopes for random effect term like, *item*.

In this thesis, statistically, we examine the effect of predictability, speech degradation and speech rate (see section X.X, X.X, X.X) on response accuracy. And hence we use these variables in the fixed effects term. Subjects and items are used as random intercepts with by-subject and by-item slopes. The details of the models fitted to data from each experiment are given in Chapter X, X, X and X.

We therefore use binomial logistic mixed effects model as our primary statistical analysis tool in all the experiments reported in this thesis. We primarily follow the recommendations of Baayen et al. (2008), Barr et al. (2013), and Bates, Kliegl, et al. (2015).

4.5 Running the model in R

Data preprocessing and analyses were performed in R-Studio (Version 3.6.1; R Core Team, 2019; . . .). Accuracy was analyzed with Generalized Linear Mixed Models (GLMMs) with lmerTest (Kuznetsova et al. 2017) and lme4 (Bates, Mächler, et al. 2015) packages. Binary responses (correct responses coded as 1 and incorrect responses coded as 0) for all participants were fit with a binomial logistic mixed-effects model.

On the data in from each experiment, we fitted models with maximal random effects structure that included random intercepts for each participant and item (Barr et al. 2013). By-participant and by-item slopes included in the model are discussed in the Analysis sections of Chapter X, X, X. Model selection was based on Akaike Information Criterion (AIC) (Grueber et al. 2011; Richards et al. 2011) unless otherwise stated. Random effects not supported by the data that explained zero variance according to singular value decomposition were excluded to prevent overparameterization. This gave a more parsimonious model (Bates, Kliegl, et al. 2015) which was then extended separately with: i) item-related correlation parameters, ii) participant-related correlation parameter, and iii) both item- and participant-related correlation parameters. The best fitting model

among the parsimonious and extended models was then selected as the optimal model for our data.

Everyone knows what attention is.

— William James (James 1890)

5

Experiment 1: Predictability effect of degraded speech are reduced as a function of attention

Contents

This chapter comes from the manuscript that is under prep for JML.

5.1 Introduction

Understanding speech is highly automatized and seemingly easy when conditions are optimal. However, in our day-to-day communication, conditions are often far from being optimal. Intelligibility and comprehension of speech can be compromised at the source (speaker), at the receiver (listener), and at the transmission of the speech signal (environmental factor; Shannon, 1948). For example, conversation with a friend over the phone can be corrupted by a poor transmission of the speech signal which in turn hampers language comprehension. Interestingly, although the speech signal is sometimes bad (distorted, or noisy), listeners do not always fail to understand what a friend is saying over the phone. Instead, listeners are successful in understanding distorted speech by utilizing context information which contains information in a given situation about a topic of conversation, semantic and syntactic information of a sentence structure, world knowledge, visual information, etc.(Kaiser and Trueswell 2004; Knoeferle et al. 2005; Altmann and Kamide 2007; Xiang and Kuperberg 2015; **Stilp2020**). To utilize the context information, listeners must attend to it and build up a meaning representation of what has been said. Listeners attend to the context information in clear speech with minimal effort, but processing and comprehending degraded speech is more effortful and requires more attentional resources (**Peelle2018**; **Wild2012**; Eckert et al. 2016).

In this chapter we examine how attention modulates the predictability effects brought about by contextual information or cues, in an adverse listening condition. We address the existing unclarity in the literatureß regarding how listeners distribute their attentional resources in adverse listening conditions: On the one hand, listeners can attend throughout the whole stream of speech and may thereby profit from the context information to predict sentence endings. On the other hand, listeners can focus their attention on linguistic material at a particular time point in the speech stream and, as a result, miss critical parts of the sentence context. If the goal is to understand a specific word in an utterance, there is a trade-off between allocating

attentional resources to the perception of that word vs. allocating resources also to the understanding of the linguistic context and generating predictions.

We report a study that was run with an aim to investigate how the allocation of attentional resources induced by different task instructions influence language comprehension and, in particular, the use of context information under adverse listening conditions. To examine the role of attention on predictive processing under degraded speech, we conducted two experiments in which we manipulated task instructions. In Experiment 1, participants were instructed to only repeat the final word of the sentence they have heard, while in Experiment 2, they were instructed to repeat the whole sentence, and by this drawing attention to the entire sentence including the context. In both experiments we varied the degree of predictability of sentence endings as well as the degree of speech degradation.

5.1.1 Predictive processing and language comprehension under degraded speech

As we have discussed earlier in Chapter 1, it is well documented in literature that human comprehenders generate expectations about upcoming linguistic material based on context information (**Kuperberg2016**; **Nieuwland2019**; for reviews, see Pickering and Gambi 2018; Staub 2015). These expectations are formed while sentence unfolds. The claims about the predictive nature of language comprehension are based on a variety of behavioral and electrophysiological experimental measures including eye-tracking and electroencephalography (EEG). For instance, in the well-known visual world paradigm, listeners fixate at a picture of an object (e.g., the cake) that is predictable based on the prior sentence context (e.g., 'The boy will eat the . . . ') even before hearing the final target word (**Ankener2018**; e.g., Altmann and Kamide 1999; Altmann and Kamide 2007). Moreover, highly predictable words are read faster and are skipped more often compared to less predictable words (Frisson et al. 2005; Rayner et al. 2011).

In EEG studies, the N400, a negative going EEG component, that usually peaks around 400 ms post-stimulus is considered as a neural marker of semantic

unexpectedness (Kutas and Federmeier 2011). For instance, in the highly predictable sentence context 'The day was breezy so the boy went outside to fly . . . ,' DeLong et al. (2005) found that the amplitude of the N400 component for the expected continuation 'a kite' was much smaller than for the unexpected continuation 'an airplane'. Although these studies demonstrated that as the sentence context builds up, listeners form predictions about upcoming words in the sentence, the universality and ubiquity of predictive language processing has been questioned (**Huettig2016**). Also, the use of context for top-down prediction can be limited by factors like literacy (**Mishra2012**), age, and working memory (**Federmeier2010**; **Federmeier2002**), as well as by the experimental setup (**Huettig2019**). While these language comprehension studies investigating predictive processing have used clean speech and sentence reading, the present study focuses on examining how attention influences the use of context to form top-down prediction under adverse listening conditions

There is already some evidence that when the bottom-up speech signal is less reliable due to degradation, listeners tend to rely more on the context information to support language comprehension (Amichetti et al. 2018; Obleser and Kotz 2009; Sheldon et al. 2008b). For example, Sheldon et al. (2008b) (Figure 2) estimated that for both younger and older adults, the number of noise-vocoding channels required to achieve 50% accuracy varied as a function of sentence context. Compared to high predictability sentences, a greater number of channels (i.e., more bottom-up information) was required in less predictability sentences to achieve the same level of accuracy. Therefore, they concluded that when speech is degraded, predictable sentence context facilitates word recognition. Obleser, Wise, et al. (2007) found that at a moderate level of spectral degradation, listeners' word recognition accuracy was higher for constraining sentence contexts than for non-constraining ones. However, while listening to the least degraded speech, there was no such beneficial effect of sentence context (see also Obleser and Kotz 2009). Hence, especially when the bottom-up speech signal is less reliable due to moderate degradation, information available from the sentence context is used to enhance language comprehension,

suggesting that there is a dynamic interaction between top-down predictive and bottom-up sensory processes in language comprehension (Bhandari et al. 2021).

5.1.2 Attention and predictive language processing

Not only the quality of speech signal influences the reliance and use of predictive processing but also attention to auditory input is important. Auditory attention allows a listener to focus on the speech signal of interest (**Fritz2007**; **Lange2013**). For instance, it has been shown that a listener can attend to and derive information from one stream of sound among many competing streams as demonstrated in the well-known *cocktail party effect* (**Cherry1953**; **Hafter2007**). When a participant is instructed to attend to only one of the two or more competing speech streams in a diotic or dichotic presentation, response accuracy to the attended speech stream is higher than to the unattended speech (**Toth2020**). Similarly, when a listener is presented with a stream of tones (e.g., musical notes varying in pitch, pure tones of different harmonics) but attends to any one of the tones appearing at a specified time point, this is reflected in a larger amplitude of N1 (**Lange2010**; **Sanders2008**) which is the first negative going ERP component peaking around 100 ms post-stimulus considered as a marker of auditory selective attention (**Naatanen1987**; **Thorton2007**). Hence, listeners can draw attention to and process one among multiple competing speech streams.

So far, most previous studies investigated listeners' attention within a single speech stream by using acoustic cues like accentuation and prosodic emphasis. For example, **Li2014** examined whether the comprehension of critical words in a sentence context was influenced by a linguistic attention probe such as "ba" presented together with accented or de-accented critical word. The N1 amplitude was larger for words with such attention probe than for words without a probe. These findings support the view that attention can be flexibly directed either by instructions towards a specific signal or by linguistic probes (**Li2017**). Thus, listeners are able to select a part or segment of stream of auditory stimuli to pay attention to.

The findings on the interplay of attention and prediction mentioned above come from studies most of which used a stream of clean speech or multiple streams of clean speech in their experiments. They cannot tell us about the attention-prediction interplay in degraded speech comprehension. Specifically, we do not know what role attention to a segment of speech stream plays in the contextual facilitation of degraded speech comprehension, although separate lines of research show that listeners attend to most informative portion of speech stream (**Astheimer2011**), and semantic predictability facilitates comprehension of degraded speech (e.g., Obleser and Kotz 2009). In two experiments, we therefore examined whether context-based semantic predictions are automatic during effortful listening to degraded speech, when participants are instructed to report only the final word of the sentence, or the entire sentence. We varied the task instructions to the listeners from Experiment 1 to Experiment 2 which required them to differentially attend to the target word, or to the target word including the context. We hypothesized that when listeners pay attention only to the contextually predicted target word, they do not form top-down predictions, i.e., there should not be a facilitatory effect of target word predictability. In contrast, when listeners attend to the whole sentence, they do form expectations such that the facilitatory effect of target word predictability will be observed.

5.2 Experiment 1A

This experiment was designed such that processing the context was not strictly necessary for the task. Listeners were asked to report the noun of the sentence that they heard which was in the final position of the sentence. This instruction did not require listeners to pay attention to the context which preceded the target word.

5.3 Methods

5.3.1 Participants

We recruited 50 participants online via Prolific Academic. One participant whose response accuracy was less than 50% across all experimental conditions was removed. Among the remaining 49 participants ($\bar{x} \pm$ SD = 23.31 $\pm$ 3.53 years; age range = 18 - 30 years), 27 were male and 22 were female. All participants were native speakers of German residing in Germany, and did not have any speech-language disorder, hearing loss, or neurological disorder (all self-reported). All participants received 6.20 Euro as monetary compensation for their participation. The experiment was approximately 40 minutes long. The German Society for Language Science ethics committee approved the study and participants provided an informed consent in accordance with the declaration of Helsinki.

5.3.2 Stimuli

We used the stimuli created by the method described in Section X.X.X in Chapter X.X which consisted of 360 German sentences spoken by a female native German speaker in an unaccented normal rate of speech. 120 nouns were used to create three categories of sentences differing in the cloze probability of the target words (nouns) which appeared as the final word of the sentence. Thus, we compared high, medium and low predictability sentences which were sentences with low, medium, and high cloze target words respectively. This gave 360 sentences that consisted of pronoun, verb, determiner, and object (noun). The mean cloze probabilities of target words for low, medium and high predictability sentences were 0.022 $\pm$ 0.027 ($\bar{x} \pm$ SD; range = 0.00 - 0.09), 0.274 $\pm$ 0.134 ($\bar{x} \pm$ SD; range = 0.1 - 0.55), and 0.752 $\pm$ 0.123 ($\bar{x} \pm$ SD; range = 0.56 - 1.00) respectively. Speech degradation was achieved by noise vocoding through 1, 4, 6, and 8 channels.

Each participant was presented with 40 high predictability, 40 medium predictability, and 40 low predictability sentences. Levels of speech degradation were also balanced across each predictability level, so that for each of the three predictability conditions (high, medium and low predictability), ten 1 channel,

ten 4 channels, ten 6 channels, and ten 8 channels noise vocoded sentences were presented, resulting in 12 experimental lists. The sentences in each list were pseudo-randomized so that no more than three sentences of same degradation and predictability condition appeared consecutively.

5.3.3 Procedure

Participants were asked to use headphones or earphones. A sample of noise vocoded speech not used in the practice trial and the main experiment was provided so that the participants could adjust the loudness to a preferred level of comfort at the beginning of the experiment. The participants were instructed to listen to the sentences and to type in the target word (noun) by using the keyboard. The time for typing in the response was not limited. They were also informed at the beginning of the experiment that some of the sentences would be 'noisy' and not easy to understand, and in these cases, they were encouraged to guess what they might have heard. Eight practice trials with different levels of speech degradation were given to familiarize the participants with the task before presenting all 120 experimental trials with an inter-trial interval of 1000 ms.

5.4 Analyses

We preprocessed and analysed data in R-Studio (Version 3.6.3; R Core Team, 2020) following the procedure described in Chapter 4.4.4. At 1 channel, there were only 5 correct responses, one each from 5 participants among 49. Therefore, the 1 channel speech degradation condition was excluded from the analyses.

Response accuracy was analyzed with Generalized Linear Mixed Models (GLMMs) with lme4 (Bates, Mächler, et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages. Binary responses (correct/incorrect) for all participants were fit with a binomial logistic mixed-effects model(**Jaeger2006**; Jaeger 2008). Noise condition (categorical; 4, 6, and 8 channels noise vocoding), target word predictability (categorical; high, medium, and low), global channel context (categorical; predictable

channel context and unpredictable channel context), and the interaction of number of channels and target word predictability were included in the fixed effects.

We first fitted a model with maximal random effects structure that included random intercepts for each participant and item (Barr et al. 2013). Both, by-participant, and by-item random slopes were included for number of channels, target word predictability, and their interaction. Non-significant higher-order interactions were excluded from the fixed-effects structure and from the random-effects structure in a stepwise manner. Random effects not supported by the data that explained zero variance were excluded and a more parsimonious model was obtained (Bates, Kliegl, et al. 2015). Such a model was then extended separately with i) item-related correlation parameters, ii) participant-related correlation parameters, and iii) both item- and participant-related correlation parameters when applicable. Aiming for model parsimony, the best fitting model among the parsimonious and extended models was then selected as the optimal model for our data. Model selection was based on Akaike Information Criterion (AIC) unless otherwise stated (**Burnham2002**; Grueber et al. 2011; Richards et al. 2011).

We applied treatment contrast for number of channels (8 channels as a baseline) and sliding difference contrast for target word predictability (low predictability vs. medium predictability, and low predictability vs. high predictability sentences). We report the results from the optimal model, and are shown in Table X.X.X.

5.5 Results and Discussion

Mean response accuracies for all experimental conditions are shown in Table X.X.X and Figure X.X.X. It shows that accuracy increases with an increase in the number of noise vocoding channels, i.e., with the decrease in speech degradation. However, accuracy does not increase with an increase in target word predictability. The results of statistical analyses confirmed these observations.

We found that there was a significant main effect of number of channels, indicating that response accuracy in the 8 channels noise vocoded speech was

higher than in both 4 channels (β = -3.49, SE = 0.23, z (4246) = -15.30, $p < .001$) and 6 channels noise vocoded speech (β = -.69, SE = .22, z (4320) = -3.12, p = .002). That is, when the number of channels increased to 8, listeners made more correct responses (see Figure X.X.X). However, there was no significant main effect of target word predictability (β = -.07, SE = .17, z (4246) = -.42, p = .68, and β = -.003, SE = .16, z (4246) = -.02, p = .98), and no significant interaction between number of noise vocoding channels and target word predictability (all ps > .05).

The results of Experiment 1 indicated a decrease in response accuracy with an increase in speech degradation from 8 channels to 6 channels noise vocoding condition, and from 8 channels to 4 channels noise vocoding condition. However, response accuracy did not increase with an increase in target word predictability, and the interaction between number of noise vocoding channels and target word predictability was also absent, in contrast to previous findings (Obleser, Wise, et al. 2007; Obleser and Kotz 2011; **Hunter2018**). These results suggest that the task instruction, which asked participants to only report the final word, indeed lead to neglecting the context, and therefore the facilitatory effect of prediction was not observed. However, to further test the hypothesis – as mentioned in the beginning of this chapter – that predictability effect is dependent on attentional effect, we conducted second experiment. In the second experiment, we changed the task instruction to draw participants attention on the entire sentence such that they would decode the whole sentence including the context.

5.6 Experiment 1B

Following up on the first experiment (Experiment 1A), we conducted second experiment (Experiment 1B) on a separate group of participants with a different task instruction. This experiment was intended to test the hypothesis that facilitatory effect of top-down predictions is observed only when listeners attention is unrestricted, and context is also included within the attentional focus.

5.7 Methods

5.7.1 Participants and Materials

We recruited 48 participants ($\bar{x}$ $\pm$ SD = 24.44 $\pm$ 3.5 years; age range = 18 - 31 years; 32 males) online via Prolific Academic. Same procedure as Experiment 1A was followed. We used the same materials that were used in Experiment 1A.

5.7.2 Procedure

We followed the same procedure as in Experiment 1A with one difference. Instead of only the final word of the sentence, participants were asked to report the entire sentence by typing in what they heard.

5.8 Analyses

We followed the same data analyses procedure as in Experiment 1A. The 1 channel noise vocoding condition was excluded from the analysis. We only considered the final words of the sentences (i.e., the target words) to be either correct or incorrect; other words were not considered in the analyses. Like Experiment 1A, the results from the optimal model are reported.

5.9 Results and Discussion

Mean response accuracy for different conditions are shown in Table X.X.X. and are presented in Figure X.X.X. It shows that the accuracy increased with an increase in both the number of noise vocoding channels, and the target word predictability. These observations are confirmed by the results of statistical analyses (Table X.X.X): We again found a main effect of number of noise vocoding channels such that response accuracy at 8 channels was higher than both 4 channels (β = -3.49, SE = .23, z (4320) = -15.29, $p < .001$), and 6 channels noise vocoding (β = -0.61, SE = .20, z (4320) = -3.07, p = .002).

5. Experiment 1: Predictability effect of degraded speech are reduced as a function of attention

In contrast to Experiment 1A, there was also a main effect of target word predictability: Response accuracy in high predictability sentences was significantly higher than in low predictability sentences (β = 1.25, SE = .28, z (4320) = 4.50, $p < .001$). We also found a statistically significant interaction between speech degradation and target word predictability (β = -.95, SE = .30, z (4320) = -3.14, p = .002). Subsequent subgroup analyses of each channel condition showed that the interaction was driven by the difference in response accuracy between high predictability sentences and low predictability sentences at 8 channels (β = 1.42, SE = .62, z (1440) = 2.30, p = .02), and 6 channels noise vocoding conditions (β = 1.14, SE = .34, z (1440) = 3.31, $p < .001$); at 4 channels noise vocoding condition, the difference between high and low predictability sentences was not significant (β = .28, SE = .18, z (1440) = 1.59, p = .11).

In contrast to Experiment 1A, these results indicate an effect of target word predictability, that is, response accuracy was higher when the target word predictability was high as compared to low. Also, the interaction between predictability and speech degradation, which was not observed in Experiment 1, showed that semantic predictability facilitated the comprehension of degraded speech already at moderate degradation levels (like, 6 and 8 noise vocoding channels). In line with the findings from Experiment !a, response accuracy was better with a higher number of channels.

To test whether the difference between experimental manipulations is statistically significant, we combined the data from both the experiments in a single analysis. We ran another binomial linear mixed-effects model on response accuracy and followed the same procedure as Experiment 1 and Experiment 2 to obtain the optimal model. The model summary is shown in Table X.X.X. The model revealed that the critical interaction between experimental manipulation and target word predictability was indeed statistically significant (β = -.45, SE = .18, z (8566) = -2.55, p = .011);, i.e., the effect of predictability was larger in the group that was asked to type in the whole sentence. Together, these findings suggest that the change in task instruction, which draws attention either to the entire sentence or only to the final word, is critical for making use of the context information under degraded speech.

5.10 General Discussion

The main goals of the present study were to investigate whether online semantic predictions are formed in comprehension of degraded speech when task instructions encourage attention to the processing of the context information, or only to the critical target word. The results of two experiments revealed that attentional processes clearly modulate the use of context information for predicting sentence endings when the speech signal is moderately degraded.

In contrast to the first experiment, the results of our second experiment show and interaction between target word predictability and degraded speech. This is generally in line with existing studies that found a facilitatory effect of predictability at different levels of speech degradation when the participants were instructed to pay attention to the entire sentence [e.g., at 4 channels or 8 channels noise vocoded speech; Obleser, Wise, et al. (2007); Obleser and Kotz (2009)]. The important new finding that our study adds to the present literature is that this predictability effect may be weakened or even lost, when listeners are instructed to report only the final word of the sentence that they heard, like in Experiment 1A. The lack of predictability effect and contextual facilitation can most likely be attributed to listeners not successfully decoding the meaning of the verb of the sentence, as the verb is the primary predictive cue for the target word (noun) in our stimuli. Hence, this small change in task instructions from Experiment 1A to Experiment 1B sheds light on the role of top-down regulation of attention on using context for language comprehension in adverse listening conditions. In adverse listening condition, language comprehension is generally effortful so that focusing attention only a part of the speech signal seems much beneficial in order to enhance stimulus decoding. However, the results of this study also show that this comes at the cost of neglecting the context information that could be beneficial for language comprehension. Our findings hence demonstrate that there is a trade-off between the use of context for generating top-down predictions vs. focusing all attention on a target word. Specifically, the engagement in the use of context and generation of

top-down predictions may change as a function of attention (**Li2014**). This claim is also corroborated by the significant change in predictability effects (or contextual facilitation) from Experiment 1A to Experiment 1B, in the combined dataset.

At this point we note the differences in response accuracies across different levels of speech degradation, and contextual facilitation therein. At 8 channels, the speech was least degraded, and listeners recognized more words than in the 4 and 6 channels noise vocoded conditions, which is in line with prior studies that have found an increase in intelligibility and word recognition with an increase in number of channels (Davis et al. 2005; Obleser and Kotz 2011). Speech signal passed through 4 channels noise vocoding was the most degraded after excluding the 1 channel noise vocoded speech for analyses. Therefore, in the second experiment, at 4 channels, attending to the entire sentence did not confer contextual facilitation because decoding the context itself was difficult. Listeners could not utilize the context differentially across high and low predictability sentences to generate semantic predictions. At 6 channels – a moderate level of degradation – listeners could attend to, identify, and decode the context; hence we observed the significant difference in response accuracy between high and low predictability sentences. We observed a similar contextual facilitation at 8 channels as well. This is in line with previous findings which show that predictability effects can be observed at moderate degradation level of 8 channels noise vocoding or less (e.g., Obleser, Wise, et al. 2007; cf. Obleser and Kotz 2009). To summarize, our results indicate that there was a very strong difference in intelligibility between 4 and 6 channels, but that the difference in intelligibility between 6 and 8 channels was minimal. Note though that even for 8 channels, low predictability sentences were not always understood correctly.

From most theoretical accounts of language processing that align with predictive language processing, one would expect that listeners automatically form top-down predictions about upcoming linguistic stimuli based on prior context (**Friston2020**; **Kuperberg2016**; **Mcclelland1986**; **Norris2016**; Pickering and Gambi 2018). Also, when speech is degraded, top-down predictions render a benefit in word recognition and language comprehension (e.g., Corps and Rabagliati 2020; Sheldon

et al. 2008b; Sheldon et al. 2008a). Results of our study revealed new theoretical insights by showing that this is not always the case. Top-down predictions are dependent on attentional processes (**Kok2011**), directed by task instructions, thus they are anot *always* automatic, and predictability does not *always* facilitate language comprehension when speech is degraded. To this point, our findings shed light on the growing body of literature that indicate limitations of predictive language processing accounts (**Heuttig2019**; **Huettig2016**; **Mishra2012**; Nieuwland et al. 2018).

A limitation of the current study should also be noted. In our experiments, we have used short Subject-Verb-Object sentences in which the verb is predictive of the noun; and we have given participants somewhat unnatural task of reporting the last word of a sentence. In a more naturalistic sentence comprehension task, participants would normally aim to understand a full utterance, and would most likely not have restricted goals such as first and foremost decoding a word in a specific position of the sentence. Instead, the speaker would usually indicate important words or concepts via pitch contours, stress, or intonation patterns, which would then direct the attention of a listener. Furthermore, the sentences uttered in most of the day-to-day conversations are longer, and context information builds up more gradually – information from several words is usually jointly predictive of upcoming linguistic units.

5.11 Conclusions

In conclusion, this study provides a novel insight into the modulatory role of attention regulation in the interaction between top-down predictive and bottom-up auditory processes. We show that task instructions affect distribution of attention to the noisy speech signal. This, in turn, means that when insufficient attention is given to the context, top-down predictions cannot be generated, and the facilitatory effect of predictability is substantially reduced. The findings of this study indicate limitations to predictive processing accounts of language comprehension.

6

Experiment 2: Semantic predictability facilitates comprehension of degraded speech in a graded manner

Contents

6.1 Background

In the previous chapter, we showed that semantic predictability facilitates comprehension of degraded speech. There was a significant difference in word recognition accuracies between low and high predictability sentences (i.e., contextual facilitation),

when listeners were instructed to attend to the entire sentence including the context. In this chapter, we examine if semantic predictability facilitates comprehension of degraded speech in a graded manner; whether prediction is probabilistic, or if it is an all-or-none phenomenon. To do so, we determine if listeners' are able to decode the context, at all instances, when they are attending to the entire sentence. We also examine if such facilitatory effect is influenced or modulated by adaptation to degraded speech.

6.1.1 Nature of predictability: Probabilistic or all-or-none?

There are two thoughts, or debate about the nature of predictability effects. One line of studies argue that prediction is all-or-none and deterministic [...CITE...]. For example, in garden path phenomenon, a parser first tries to predict the simplistic structure of a sentence. If this parsing is disconfirmed by the bottom-up input then the parser backs off and restarts – it reinterprets the context and predicts a new sentence structure. Therefore, according to XYZ, how a parser resolves garden path sentences is an evidence for all-or-none prediction.

Humans use only one of many possible continuations in a sentence. XYZ argue that it is metabolically expensive to predict many parallel linguistic units when only most of them are going to be discarded. Therefore, evolution would not pick a language processing system that predicts multiple parallel linguistic units. Another line of studies show that predictions are probabilistic, or graded. Nieuwland et al. (2018) showed in a large-scale replication study of DeLong et al. (2005) that the N400 amplitude at the sentence-final noun is directly proportional to its cloze probability across a range of high- and low-cloze words. Heilbron et al. (2020) also showed that a probabilistic prediction model outperforms a constrained guessing model, suggesting that linguistic prediction is not limited to highly predictable sentence endings, but it operates broadly in a wide range of probable sentence endings. The studies mentioned above were either reading studies, or were conducted with clean speech.

To our knowledge, only one study by Strauß et al. (2013) has directly studied the nature of prediction in degraded speech comprehension. Severely degraded (4

channels noise vocoded), moderately degraded (8 channels noise vocoded), and clear (non-degraded) speech were presented to the participants. Target word predictability was varied by manipulating its expectancy (i.e., how expected the target word is given the verb) and typicality (i.e., co-occurrence of target word and the preceding verb). Strauß et al. (2013) reported that at a moderate level of spectral degradation, N400 responses at strong-context, low-typical words and weak-context, low-typical words were largest. N400 responses at the latter two were not statistically different from each other. However, the N400 response was smallest at highly predictable (strong-context, high-typical) words. The authors interpreted these findings as a facilitatory effect of sentence predictability which might be limited to only highly predictable sentence endings at a moderate level of spectral degradation. Based on these findings, Strauß et al. (2013) proposed an 'expectancy searchlight model'. According to the expectancy searchlight model, listeners form 'narrowed expectations' from a restricted semantic space when the sentence endings are highly predictable. For less predictable sentence endings, listeners cannot preactivate those less predictable sentence endings in an adverse listening condition. Therefore, the expectancy searchlight model of Strauß et al. (2013) is in line with all-or-none prediction account. These theoretical accounts have not been further explored in degraded speech comprehension. However, it has been shown that listeners rely on context when speech is moderately degraded [CITE].

6.1.2 Predictability effects in degraded speech comprehension

When the bottom-up perceptual input is difficult to understand, listeners rely more on top-down predictions in adverse listening conditions like noisy environment with background noise [...CITE...], reverberation [...CITE...], and degraded speech (e.g., Sheldon et al. 2008b; Corps and Rabagliati 2020).

In their studies, Obleser and colleagues (Obleser, Wise, et al. 2007; Obleser and Kotz 2009; Obleser and Kotz 2011), used sentences of two levels of semantic predictability (high and low) and systematically degraded the speech signal by

passing it through various numbers of noise vocoding channels ranging from 1 to 32 in a series of behavioral and neuroimaging studies. They found that semantic predictability facilitated language comprehension at a moderate level of speech degradation (at 4 channels, or 8 channels noise vocoding). That is, participants relied on the sentence context when the speech signal was degraded but intelligible enough. Accuracy of word recognition was found to be higher for highly predictable target words than for lowly predictable target words at such moderate levels of speech degradation (Obleser and Kotz 2009). For the extremes, i.e., when the speech signal was highly degraded or when it was clearly intelligible, the word recognition accuracy was similar across both levels of sentence predictability, meaning that predictability did not facilitate language comprehension. It can be safely concluded from these findings that at moderate levels of degradation, participants rely more on the top-down prediction generated by the sentence context and less on the bottom-up processing of unclear, less intelligible (but intelligible enough), and degraded speech signal (Obleser 2014). In other words, reliance on prediction results in higher word recognition accuracy for high-cloze probability target words than for low-cloze probability target words. In the case of a heavily degraded speech signal, participants may not be able to understand the sentence context and, therefore, they are unable to predict the target word; or their cognitive resources may already be occupied by decoding the signal, leaving little room for making predictions [. . . CITE. . . look Ryskin2021]. Thus, there is no differential effect of levels of sentence predictability. On the other extreme, when the speech is clear and intelligible (at the behavioral level, i.e., when the participants respond what the target word of the sentence is), participants recognize the intelligible target word across all levels of sentence predictability. Hence, no differential effect of levels of predictability of target word can be expected.

In contrast to clear speech perception, listeners adapt to degrade speech and their performance has been shown to improve over the course of experiment [e.g., CITE]. Therefore, *adaptation* has to be considered when we review and examine predictability effects on degraded speech.

6.1.3 Adaptation to degraded speech

Listeners quickly adapt to novel speech with artificial acoustic distortions (e.g., Dupoux and Green 1997). Repeated exposure to degraded speech leads to improved comprehension over time (for a review, Samuel and Kraljic 2009; Guediche et al. 2014). When the noise condition is constant throughout the experiment, listeners adapt to it and the performance (e.g., word recognition) improves with as little as 20 minutes of exposure (e.g., Rosen et al. 1999). For example, Davis et al. (2005, Experiment 1) presented listeners with only 6 channels noise vocoded sentences and found an increase in the proportion of correctly reported words over the course of experiment. Similarly, Erb et al. (2013) presented participants with only 4 channels noise vocoded sentences and reached a similar conclusion. In these experiments, a single noise condition (6 channels or 4 channels) was presented in one block. Therefore, from the participants' perspective, the next-trial speech degradation level was predictable. Additionally, target word predictability of the sentences were not varied by any measures.

When multiple noise conditions are presented in a (pseudo-)randomized order within a block then a listener is uncertain about any upcoming trial's noise condition, i.e., if such multiple levels of degradation are due to presentation of multiple channels of noise vocoded speech, then the global channel context is unpredictable or uncertain. This can potentially influence perceptual adaptation. For instance, Mattys et al. (2012) note the possibility of total absence of perceptual adaptation, when the characteristics of auditory signal change throughout an experiment. We also know from Sommers et al. (1994) that trial-by-trial variability in the characteristics of distorted speech impairs word recognition (see also, Dahan and Magnuson 2006). We thus speculated that if the noise vocoded speech varies from one trial to the next then the adaptation to noise in this scenario might be different from the earlier case in which the noise condition is constant throughout the experiment. Perceptual adaptation, however, is not limited to trial-by-trial variability of stimulus property. Listeners can adapt to auditory signal at different time courses and time

scales (Atienza et al. 2002; see also, Whitmire and Stanley 2016). In addition to differences in intrinsic trial-by-trial variability and resulting short timescale trial-by-trial adaptation in two channel contexts, the global differences in the presentation of vocoded speech can result in a difference in the general adaptation at a longer timescale between predictable and unpredictable channel contexts.

There is a limited number of studies that has looked at how next-trial noise-uncertainty and global context of speech property influence adaptation. For example, words were presented at +3 dB SNR and +10 dB SNR in a word recognition task in a pseudorandom order (Vaden, Kuchinsky, Cute, et al. 2013). The authors wanted to minimize the certainty about the noise conditions in the block. The same group of authors (Vaden, Kuchinsky, Ahlstrom, Dubno, et al. 2015; Vaden, Kuchinsky, Ahlstrom, Teubner-Rhodes, et al. 2015; Eckert et al. 2016) proposed that an adaptive control system (cingulo opercular circuit) might be involved to optimize task performance when listeners are uncertain about upcoming trial. However, we cannot make a firm conclusion about perceptual adaptation *per se* from their studies as they do not report the change in performance over the course of experiment. Similarly, Obleser and colleagues (Obleser, Wise, et al. 2007; Obleser and Kotz 2011; Hartwigsen et al. 2015) also presented listeners with noise vocoded sentences (ranging from 2 to 32 channels of noise vocoding) in a pseudo-randomized order but did not report presence or absence of perceptual adaptation to noise vocoded speech. In the abovementioned studies, the authors did not compare participants' task performance in the blocked design against the presentation in a pseudorandomized block of different noise conditions to make an inference about general adaptation to degraded speech at a longer timescale. To examine the influence of uncertainty about next trial speech features and the global context of speech features on perceptual adaptation, we therefore compared language comprehension with a trial-by-trial variation of sentence predictability and speech degradation either in blocks, in which the noise vocoded channels were blocked, or in a randomized order.

6.2 Methods

6.2.1 Participants

We recruited two groups of participants via Prolific Academic and assigned them to one of the two groups: *unpredictable channel context* (n=48; $\bar{x}$ ± SD = 24.44 ± 3.5 years; age range = 18-31 years; 16 females) and *predictable channel context* (n=50; ± SD = 23.6 ± 3.2 years; age range = 18-30 years; 14 females). All participants were native speakers of German residing in Germany. Exclusion criteria for participating in this study were self-reported hearing disorder, speech-language disorder, or any neurological disorder. All participants received monetary compensation for their participation. The study was approved by Deutsche Gesellschaft für Sprachwissenschaft (DGfS) Ethics Committee, and the participants provided consent in accordance with the Declaration of Helsinki.

6.2.2 Materials

We used the same stimuli described in Section X.X.X in Chapter X.X. The stimuli were digital recordings of 360 German sentences spoken by a female native speaker of German in a normal rate of speech. All sentences were in present tense consisting of pronoun, verb, determiner, and object (noun) in the Subject-Verb-Object form. We used 120 nouns to create three categorizes of sentences – high predictability sentences (HP sentencs), medium predictability sentences (MP sentences) and low predictability sentences (LP sentences) – that differed in cloze probability of sentence final target words. (See Appendix A for examples.) Their mean cloze probabilities were 0.022 ± 0.027 ($\bar{x}$ ± SD; range = 0.00 - 0.09) for LP sentences, 0.274 ± 0.134 ($\bar{x}$ ± SD; range = 0.1 - 0.55) for MP sentences, and 0.752 ± 0.123 ($\bar{x}$ ± SD; range = 0.56 - 1.00) for HP sentences. The distribution of cloze probability across LP, MP and HP sentences are shown in Figure X.X.

In the unpredictable channel context, each participant was presented with 120 unique sentences: 40 HP, 40 MP and 40 LP sentences. Channel condition was also balanced across each sentence type, i.e., in each of HP, MP, and LP sentences,

ten sentences passed through each noise vocoding channels – 1, 4, 6, and 8 – were presented. This resulted into 12 experimental lists. The sentences in each list were pseudo-randomized, that is, not more than 3 sentences of same noise condition (i.e., same noise vocoding channel), or same predictability condition appeared consecutively. This randomization confirmed the uncertainty of next-trial speech quality (or degradation) in the global context of the experiment.

The same set of stimuli and experimental lists were used in the predictable channel context. Each participant was presented with 120 unique sentences blocked by channel conditions, i.e., blocked by noise vocoding channels. There were four blocks of stimuli. Thirty sentences were presented in each of the four blocks. In the first block, all sentences were 8 channels noise vocoded, followed by blocks of 6 channels, 4 channels, and 1 channel noise vocoded speech consecutively (Sheldon et al. 2008b). Within each block, 10 HP, 10 MP and 10 LP sentences were presented. All the sentences were pseudo-randomized so that not more than three sentences of the same predictability condition appeared consecutively in each block. This ascertained there was a certainty of next-trial speech quality (within each block) and an uncertianty of next-trial sentence predictability across all four blocks.

6.2.3 Procedure

Participants were asked to use headphones or earphones. A prompt to adjust loudness was displayed at the beginning of the experiment: A noise vocoded sound not used in the main experiment was presented, and participants were asked to adjust the loudness at their level of comfort. One spoken sentence was presented in each trial. Eight practice trials were presented before presenting 120 experimental trials. They were asked to enter what they had heard (i.e., to type in the entire sentence) *via* keyboard. Guessing was encouraged. The response was not timed. The experiment was about 40 minutes long.

6.3 Analyses

In the sentences used in our experiment, verbs evoke predictability of the sentence-final noun. Therefore, the effect of predictability (evoked by the verb) on language comprehension can be rightfully measured if we consider only those trials in which participants identify the verbs correctly. Verb-correct trials were considered as the sentence in which participants realized the context independent of whether they correctly understood the sentence final target noun. Morphological inflections and typos were considered as correct. We first filtered out those trials in which verbs were not identified correctly, i.e., trials with incorrect verbs. Therefore, we excluded 2469 out of 5760 trials in unpredictable channel context and 2374 out of 6000 trials in predictable channel context from the analyses. The 1 channel noise vocoding condition was dropped from the analyses as there were no correct responses in any of the trials in this condition.

We preprocessed and analysed data in R-Studio (Version 3.6.1; R Core Team, 2019) following the procedure described in Chapter 4.4.4. Response accuracy was analyzed with Generalized Linear Mixed Models (GLMMs) with lme4 (Bates, Mächler, et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages. Binary responses (correct/incorrect) for all participants in both groups (predictable channel context and unpredictable channel context) were fit with a binomial logistic mixed-effects model(**Jaeger2006**; Jaeger 2008). Noise condition (categorical; 4, 6, and 8 channels noise vocoding), target word predictability (categorical; HP, MP, LP), global channel context (categorical; predictable channel context and unpredictable channel context), and the interaction of noise condition and target word predictability were included in the fixed effects.

We first fitted a model with maximal random effects structure that included random intercepts for each participant and item (Barr et al. 2013). Both, by-participant and by-item random slopes were included for noise condition, target word predictability and their interaction. To find the optimal model for the data, non-significant higher-order interactions were excluded from the fixed-effects

structure (and from the random-effects structure) in a stepwise manner. Model selection was based on Akaike Information Criterion (AIC) (Grueber et al. 2011; Richards et al. 2011) unless otherwise stated. Random effects not supported by the data that explained zero variance according to singular value decomposition were excluded to prevent overparameterization. This gave a more parsimonious model (Bates, Kliegl, et al. 2015). Such a model was then extended separately with: i) item-related correlation parameters, ii) participant-related correlation parameter, and iii) both item- and participant-related correlation parameters. The best fitting model among the parsimonious and extended models, based on AIC, was then selected as the optimal model for our data.

We applied treatment contrast for noise condition (8 channels as a baseline; factor levels: 8 channels, 4 channels, 6 channels) and sliding difference contrast for target word predictability (factor levels: MP, LP, HP) and channel context (factor levels: unpredictable, predictable). The results from the optimal model are shown in Table X.X.X, and are reported below in the Results section.

6.4 Results

In this experiment, we tested i) whether predictability facilitates language comprehension only at a moderate level of spectral degradation, and ii) whether adaptation to degraded speech influences language comprehension. We observed that the mean response accuracy increased with an increase in number of noise vocoding channels from 4 to 6 to 8, and with an increase in target word predictability from low to medium to high (see Figure X.X). This trend is consistent across both the channel contexts; Figure X.X and Figure Y.Y show this trend for predictable channel context (i.e., blocked design) and unpredictable channel context (i.e., randomized design) respectively. Mean accuracies across all conditions are given in Table X and Y, and Figure X.X.

These observations are confirmed by the results of statistical analyses. We found a significant main effect of channel condition indicating that the response accuracy

was higher in the 8 channels than in the 4 channels (β = -2.87, SE = 0.22, z (6917) = -13.1, p <.001) and 6 channels (β = -0.66, SE = 0.19, z (6917) = -3.42, p < .001). There was a significant main effect of target word predictability suggesting that response accuracy was lower at low predictability sentences than both high predictability sentences (β = 2.18, SE = 0.3, z (6917) = 7.2, p < .001) and medium predictability sentences (β = -0.52, SE = 0.27, z (6917) = -1.97, p = .049). We also found a significant interaction between channel condition and target word predictability (β = -0.71, SE = 0.29, z (6917) = -2.44, p = .015).

We performed a subsequent subgroup analyses on each noise channel condition. They revealed that the interaction was driven by the effect of predictability at 4 channels: The accuracy at high predictability sentences was higher than medium predictability sentences (β = 1.14, SE = 0.37, z (1608) = 3.1, p < .001), which in turn was also higher than low predictability sentences (β = 1, SE = 0.24, z (1608) = 4.2, p < .001). There was no significant difference in response accuracy between low predictability and high predictability sentences at both 6 channels (β = 0.33, SE = 0.32, z (2590) = 1.04, p = .3) and 8 channels (β = -0.014, SE = 0. 32, z (2719) = -0.04, p = .97). However, response accuracy was higher in high predictability than in medium predictability sentences at both 6 channels (β = 1.83, SE = 0.65, z (2590) = 2.83, p < .005) and 8 channels (β = 1.54, SE = 0.61, z (2719) = 2.54, p = .011).

We also found a significant main effect of global channel context which showed that the response accuracy was higher in predictable channel context than in unpredictable channel context (β = -0.27, SE = 0.14, z (6917) = -2.02, p = .04).

Further, to test the effect of practice on adaptation to degraded speech, we added trial number as a fixed effect in the maximal model. Note that there were 30 trials in each block in the predictable channel context (i.e., blocked design). For comparability, we divided unpredictable channel context (i.e., randomized design) into four blocks. Then following the same procedure as above, we obtained an optimal model. We did not find a significant main effect of trial number indicating that the response accuracy did not change throughout the experiment (β = -0.0004, SE = 0.01, z (6917) = -0.05, p = 0.97). It remained constant within each block

in the predictable channel context (β = -0.02, SE = 0.01, z (3291) = -1.43, p = 0.15) as well as in the unpredictable channel context (β = 0.01 SE = 0.01, z (3291) = 1.05, p = 0.29).

6.5 Discussion

The present study had three goals: i) to examine if previously reported facilitatory effect of semantic predictability is restricted to only highly predictable sentence endings; ii) to assess the role of perceptual adaptation on the facilitation of language comprehension by sentence predictability; and iii) to use and establish a sensitive metric to measure language comprehension that takes into account whether listeners benefited from the semantic context of the sentence they have listened to.

Results of our study showed the expected interaction between predictability and degraded speech, that is, language comprehension was better for high-cloze than for low-cloze target words when the speech signal was moderately degraded by noise-vocoding through 4 channels, while the effect of predictability was absent when speech was not intelligible (noise-vocoding through 1 channel). These results are fully in line with Obleser and Kotz (2009); we partly included identical sentences from their study in the present study (see Appendix A). Importantly, in contrast to their study, we had also created sentences with medium-cloze target words (which were intermediate between high-cloze and low-cloze target words) and found that the effect of predictability was also significant when comparing sentences with medium-cloze target words against sentences with low-cloze target words at 4 channels noise-vocoding condition. Recognition of a target word was dependent on its level of predictability (measured by cloze probability), and correct recognition was not just limited to high-cloze target words. These significant differences in response accuracy between medium-cloze and low-cloze target words, and between medium-cloze and high-cloze target words at noise-vocoding through 4 channels show that the sentence-final word recognition is facilitated by semantic predictability in a graded manner. This is in line with the findings from the ERP literature where it has been

observed that semantic predictability, in terms of cloze probability of target word of a sentence, modulates semantic processing, indexed by N400, in a graded manner (DeLong et al. 2005; Wlotko and Federmeier 2012; Nieuwland et al. 2018).

The interpretation of the observed graded effect of semantic predictability at the moderate level of spectral degradation (i.e., at noise-vocoding through 4 channels) provides a novel insight into how listeners form prediction when the bottom-up input is compromised. That is, in an adverse listening condition, listeners rely more on top-down semantic prediction than on bottom-up acoustic-phonetic cues. However, such a reliance on top-down prediction is not an all-or-none phenomenon; instead, listeners form a probabilistic prediction of the target word. The effect of target word predictability on comprehension is not sharply focused solely on high-cloze target words like a 'searchlight'. But rather it is spread across a wide range including low-cloze and medium-cloze target words. As the cloze probability of the target words decreases from high to low, the focus of the searchlight becomes less precise.

6.6 Conclusion

In conclusion, this study provides novel insights into predictive language processing when bottom-up signal quality is compromised and uncertain: We show that while processing moderately degraded speech, listeners form top-down predictions across a wide range of semantic space that is not restricted within highly predictable sentence endings. In contrast to the narrowed expectation view, comprehension of words ranging from low- to high-cloze probability, including medium-cloze probability, is facilitated in a graded manner while listening to a moderately degraded speech. We also found better speech comprehension when individuals were likely to have adapted to the noise condition in the blocked design compared to the randomized design. We did not find learning effects at the trial-to-trial level of perceptual adaption – it may be that the adaptation was hampered by variation in higher-level semantic features (i.e., target word predictability). We also argue that for the examination of semantic predictability effects during language comprehension, the

analyses of response accuracy should be based on the trials in which context evoking words are correctly identified in the first place to make sure that listeners make use of the contextual cues instead of analyzing general word recognition scores.

–>

7

Experiment 3: Comprehension of degraded speech is modulated by the rate of speech

Contents

7.1 Introduction

Language comprehension depends on the features and quality of speech signal which is hampered in adverse listening conditions when speech is distorted, for example, due to change in rate of speech, and spectral degradation of speech. Studies have shown that individuals can benefit from sentence context to compensate for distorted speech, at least when the level of speech degradation is at an intermediate level (e.g., Bhandari et al. 2021; Obleser and Kotz 2009), and at varying speech rate (**Aydelott2004**; **Goy2013**). The goal of the present study is to examine if sentence context can provide benefit when speech that is degraded at an intermediate level is presented at different rates (from slow, normal, and fast); speech in day-to-day conversation is distorted and presents itself at varying rate (**Krause2004**). In the following, we first summarize the impact of speech degradation in language comprehension, and its interaction with sentence context, and then the influence of speech rate on language comprehension, and its interaction with sentence context.

7.1.1 Comprehension of degraded speech

There are a number of studies showing that speech intelligibility and language comprehension is hampered when the bottom-up input is less intelligible due to spectral degradation of the speech signal (Shannon, Zeng, et al. 1995; Davis et al. 2005). These studies have used noise vocoding as a methods of speech degradation. Here the speech signal is first divided into a specific number of frequency bands that corresponds to the number of vocoder channels. The amplitude envelope within each frequency band is extracted, and the spectral information within it is replaced by noise. The resulting vocoded speech contains temporal cues of the original speech, but it is difficult to understand – the lesser the number of vocoder channels, the lesser is the intelligibility. More attentional resources are required to process and comprehend such degraded speech as compared to clean speech (**Wild2012**; e.g., Eckert et al. 2016).

Listeners rely more on top-down predictions when the speech signal is less intelligible due to spectral degradation. Hence, they use the context information of the sentence to narrow down their predictions to a smaller set of semantic categories or words (Strauß et al. 2013; see also, Corps and Rabagliati 2020). However, it is important that the context itself is 'intelligible enough' which is the case when the speech is only moderately degraded. For example, Obleser and colleagues (Obleser, Wise, et al. 2007; Obleser and Kotz 2009; Obleser and Kotz 2011) found that at moderate levels of speech degradation, target words (the sentence final words) were better recognized when it was predictable from the sentence context than when it was unpredictable. When the speech signal is clear or only very mildly degraded, there is typically no effect of predictability on comprehension, as even unpredictable words can be understood well in this condition (intelligibility is at ceiling). In contrast, when the speech signal is extremely degraded (for instance at 1 channel noise vocoding), no facilitation from the context can be observed as the context itself cannot be understood and hence it cannot help with comprehension (Bhandari et al. 2021; Obleser, Wise, et al. 2007; Obleser and Kotz 2009; Obleser and Kotz 2011).

These studies also found that semantic predictability facilitated comprehension of degraded speech at a moderate level of spectral degradation, i.e., at 4 channels noise vocoding. At the moderate degradation level, most speech is intelligible enough for listeners to understand and form meaning representation of the context to generate predictions about upcoming word in the sentence. In sum, xxxx

7.1.2 Comprehension of fast and slow speech

A change in speech rate manipulates the speech signal without producing any spectral degradation. Compared to normal and slow speech, understanding fast speech is more effortful (e.g., Müller et al., 2019; Winn & Teece, 2021; see also, Simantiraki & Cooke, 2019, 2020), with reduced intelligibility and comprehension (Fairbanks & Kodman Jr., 1957; Garvey, 1953; Goldstein, 1941; Konkle et al., 1977; Liu & Zeng, 2006; Peelle & Wingfield, 2005; Schlueter et al., 2014). The

comprehension deficit in fast speech has been linked to speed of processing (Gordon-Salant & Fitzgibbons, 1995; Tun, 1998; see also, Rönnberg et al., 2013). Decoding speech signal and processing language does not occur in vacuum; they take place within the limited cognitive resources and memory that we have (Riggs et al., 1993). When the speech rate is fast, the flow of information is also fast. There is a limited time available to decode and understand the information in the fast speech. Decoding and identification of incoming information in the fast speech puts a high demand on available cognitive resources (e.g., Rodero, 2016). Processing rapidly flowing information exhausts the cognitive resource that is required for language processing (Gordon-Salant & Fitzgibbons, 2004; Janse, 2009). Hence, intelligibility and comprehension of fast speech is reduced compared to normal speech.

In contrast, the evidence on the effect of decrease in speech rate on intelligibility and language comprehension is mixed. **Shobha2009** showed that expansion of speech signal benefits word recognition in noise at different signal to noise ratio (SNR) ranging from 6dB SNR to 12 dB SNR. In a sample of younger adults and middle aged adults, slowing down the speech rate facilitated word recognition when listeners had to segregate the speech from background talkers, i.e., in speech masking (**Brungart2007**).

7.1.3 Predictive processing, degraded speech, and different rates of presentation of peech

7.1.4 Current study

7.2 Experiment 3A

7.3 Methods

7.3.1 Participants

We recruited one group of participant (n=101; $\bar{x} \pm$ SD = 23.14 $\pm$ XX years; age range = 18-31 years; 66 females, 1 preferred not to say) online via Prolific Academic. All participants were native speakers of German residing in Germany. Exclusion

criteria for participating in this study were self-reported hearing disorder, speech-language disorder, or any neurological disorder. All participants received monetary compensation for their participation. The German Society for Language Science ethics committee approved the study and participants provided an informed consent in accordance with the declaration of Helsinki.

7.3.2 Materials

We used the stimuli created by the method described in Section X.X.X in Chapter X.X which consisted of 360 German sentences spoken by a female native German speaker in an unaccented normal rate of speech. Two categories of sentences that differed in the cloze probability of the target words (nouns) appearing the final word of the sentence were created from 120 nouns. Thus, we compared high and low predictability sentences (abbreviated as HP and LP henceforth) which were sentences with low and high cloze target words respectively. This gave 240 sentences that consisted of pronoun, verb, determiner, and object (noun). The mean cloze probabilities of target words for low and high predictability sentences were 0.022 ± 0.027 ($\bar{x}$ ± SD; range = 0.00 - 0.09) and 0.752 ± 0.123 ($\bar{x}$ ± SD; range = 0.56 - 1.00) respectively.

These 240 sentences were passed compressed by a factor of 0.65 using PSOLA built-in in Praat to create fast and slow speech respectively. Speech degradation of all normal, slow and fast speech was achieved by noise vocoding through 4 channels.

Each participant was presented with 120 unique sentences: 60 HP and 60 LP sentences. Speech rate was also balanced across each predictability level. The participants received 30 sentences with normal speed and 30 with fast speed in each of the predictability conditions resulting into 4 experimental lists. The sentences in each list were pseudo-randomized, that is, not more than 3 sentences of same speed, or same predictability condition appeared consecutively.

7.3.3 Procedure

Participants were asked to use headphones or earphones. A sample of noise vocoded speech not used in the practice trial and the main experiment was provided so that the participants could adjust the loudness to a preferred level of comfort at the beginning of the experiment. The participants were instructed to listen to the sentences and to type in the entire sentence by using the keyboard. The time for typing in the response was not limited. They were also informed at the beginning of the experiment that some of the sentences would be 'noisy' and not easy to understand, and in these cases, they were encouraged to guess what they might have heard. They were not informed about the speed of speech being slow/fast or normal. Eight practice trials with different levels of speech degradation were given to familiarize the participants with the task before presenting all 120 experimental trials with an inter-trial interval of 1000 ms.

7.4 Analyses

We have already conceded in the previous chapter X.X.X that "the effect of predictability (evoked by the verb) on language comprehension can be rightfully measured if we consider only those trials in which participants identify the verbs correctly." Therefore, we discarded the trials in which verbs were identified incorrectly – XXXX out of XXXX trials.

We preprocessed and analysed data in R-Studio (Version 4.1.1; R Core Team, 2021) following the procedure described in Chapter 4.4.4.

Response accuracy was analyzed with Generalized Linear Mixed Models (GLMMs) with lme4 (Bates, Mächler, et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages. Binary responses (correct/incorrect) for all participants were fit with a binomial logistic mixed-effects model(**Jaeger2006**; Jaeger 2008). Target word predictability (categorical; low and high), speech rate, or speed (categorical; xxx and xxx), and the interaction of predictability and speed were included in the fixed effects.

We fitted a model with maximal random effects structure that included random intercepts for each participant and item (Barr et al. 2013). Both, by-participant and by-item random slopes were included for target word predictability, speed and their interaction.

We applied treatment contrast for target word predictability (XX as a baseline; factor levels: XX, XX) and speed (XXX as a baseline; factor levels: XX, XX). The results from the maximal model are shown in Table X.X.X, and are reported below in the Results section.

7.5 Results and Discussion

7.6 Experiment 3B

7.7 Methods

7.7.1 Participants and Materials

We recruited 48 participants (n=101; ± SD = 23.6 ± 3.2 years; age range = 18-30 years; 14 females) online via Prolific Academic. Same procedure as Experiment 3A was followed.

Stimuli were created following the same procedure described in Experiment 3A, the difference being instead of fast speech, here we created slow speech. Slow speech was created by expanding the 240 sentences (120 HP and 120 LP sentences) by a factor of 1.35 using PSOLA built-in in Praat. All resulting slow speech recordings were then passed through 4 channels noise vocoding.

We followed the same steps as Experiment 3A to balance speech rate and predictability conditions and to pseudo-randomize these experimental conditions.

7.7.2 Procedure

Same procedure as Experiment 3A was followed. We asked participants to report the entire sentence typing in what they heard. Guessing was encouraged too.

7.8 Analyses

We followed the same data analyses procedure as in Experiment 3A. Only the trials with verb-correct responses were considered in the analyses of accuracy of sentence-final target-words (i.e., nouns); 5495 out of 12120 trials were removed before the final analyses.

Target word predictability (categorical; low and high), speech rate, or speed (categorical; normal and slow), and the interaction of predictability and speed were included in the fixed effects. Treatment contrast was applied to both target word predictability (LP as a baseline; factor levels: LP, HP) and speed (slow as a baseline; factor levels: slow, normal). The results from the maximal model are shown in Table X.X.X, and are reported below in the Results section.

7.9 Results and Discussion

Mean response accuracy for different conditions are shown in Table X.X.X. and are presented in Figure X.X.X. It shows that the accuracy increased with an increase in target word predictability from low to high. These observations are confirmed by the results of statistical analyses (Table X.X.X): We again found a main effect of number of noise vocoding channels such that response accuracy at 8 channels was higher than both 4 channels (β = -3.49, SE = .23, z (4320) = -15.29, $p < .001$), and 6 channels noise vocoding (β = -0.61, SE = .20, z (4320) = -3.07, p = .002).

In contrast to Experiment 1A, there was also a main effect of target word predictability: Response accuracy in high predictability sentences was significantly higher than in low predictability sentences (β = 1.25, SE = .28, z (4320) = 4.50, $p < .001$). We also found a statistically significant interaction between speech degradation and target word predictability (β = -.95, SE = .30, z (4320) = -3.14, p = .002). Subsequent subgroup analyses of each channel condition showed that the interaction was driven by the difference in response accuracy between high predictability sentences and low predictability sentences at 8 channels (β = 1.42, SE = .62, z (1440) = 2.30, p = .02), and 6 channels noise vocoding conditions

(β = 1.14, SE = .34, z (1440) = 3.31, $p < .001$); at 4 channels noise vocoding condition, the difference between high and low predictability sentences was not significant (β = .28, SE = .18, z (1440) = 1.59, p = .11).

In contrast to Experiment 1A, these results indicate an effect of target word predictability, that is, response accuracy was higher when the target word predictability was high as compared to low. Also, the interaction between predictability and speech degradation, which was not observed in Experiment 1, showed that semantic predictability facilitated the comprehension of degraded speech already at moderate degradation levels (like, 6 and 8 noise vocoding channels). In line with the findings from Experiment !a, response accuracy was better with a higher number of channels.

7.10 General Discussion

7.11 Conclusions

8

Experiment 4: Older adults rely more on sentence context than on the auditory signal in comprehension of moderately degraded speech

9 General discussion

9.1 Summary of the experiments

9.2 A new framework on the interaction between top-down predictive and bottom-up auditory processes in perception and comprehension of degraded speech

Contents

A conclusion is the place where you got tired of thining.

— Martin H. Fischer (**Darwin1859**)

10

Theoretical and practical implications

10.1 Potential limitations of predictive processing

10.2 Attention, adaptation and processing speech: Moderator, mediator or subsumed factor in rediction?

10.3 Implications for clinical audiology

10.3.1 Materials used in hearing tests

10.3.2 Rehabilitative training of cochlear implantees

Active attention to speech materials

Contents

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

B

The Second Appendix, for Fun

Works Cited

Altmann, Gerry T.M and Yuki Kamide (Dec. 1999). "Incremental interpretation at verbs: restricting the domain of subsequent reference". In: *Cognition* 73.3, pp. 247–264. DOI: 10.1016/s0010-0277(99)00059-1. URL: http://dx.doi.org/10.1016/s0010-0277(99)00059-1.

— (2007). "The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing". In: *Journal of Memory and Language* 57.4, pp. 502–518. DOI: 10.1016/j.jml.2006.12.004.

Amichetti, Nicole M. et al. (Jan. 2018). "Linguistic Context Versus Semantic Competition in Word Recognition by Younger and Older Adults With Cochlear Implants". In: *Ear & Hearing* 39.1, pp. 101–109. DOI: 10.1097/aud.0000000000000469. URL: http://dx.doi.org/10.1097/AUD.0000000000000469.

Atienza, Mercedes, Jose L. Cantero, and Elena Dominguez-Marin (2002). "The time course of neural changes underlying auditory perceptual learning". In: *Learning and Memory* 9.3, pp. 138–150. DOI: 10.1101/lm.46502.

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). "Mixed-effects modeling with crossed random effects for subjects and items". In: *Journal of Memory and Language* 59.4, pp. 390–412. DOI: 10.1016/j.jml.2007.12.005. URL: http://dx.doi.org/10.1016/j.jml.2007.12.005.

Barr, Dale J. et al. (Apr. 2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". In: *Journal of Memory and Language* 68.3, pp. 255–278. DOI: 10.1016/j.jml.2012.11.001. URL: http://dx.doi.org/10.1016/j.jml.2012.11.001.

Bates, Douglas, Reinhold Kliegl, et al. (2015). "Parsimonious Mixed Models". In: *arXiv*. arXiv: 1506.04967. URL: http://arxiv.org/abs/1506.04967.

Bates, Douglas, Martin Mächler, et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1. DOI: 10.18637/jss.v067.i01. URL: http://dx.doi.org/10.18637/jss.v067.i01.

Bhandari, Pratik, Vera Demberg, and Jutta Kray (2021). "Semantic predictability facilitates comprehension of degraded speech in a graded manner". In: *Frontiers in Psychology* 3769. DOI: 10.3389/fpsyg.2021.714485.

Charpentier, F. J. and M. G. Stella (1986). "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation." In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2015–2018. DOI: 10.1109/icassp.1986.1168657.

Clark, Catherine, Sara Guediche, and Marie Lallier (2021). "Compensatory cross-modal effects of sentence context on visual word recognition in adults". In: *Reading and Writing* 34.8, pp. 2011–2029. DOI: 10.1007/s11145-021-10132-x. URL: https://doi.org/10.1007/s11145-021-10132-x.

Corps, Ruth E. and Hugh Rabagliati (Aug. 2020). "How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form". In: *Journal of Memory and Language* 113, p. 104114. DOI: 10.1016/j.jml.2020.104114. URL: http://dx.doi.org/10.1016/j.jml.2020.104114.

Dahan, Delphine and James S. Magnuson (2006). "Spoken Word Recognition". In: Elsevier, pp. 249–283. DOI: 10.1016/b978-012369374-7/50009-2. URL: http://dx.doi.org/10.1016/b978-012369374-7/50009-2.

Davis, Matthew H. et al. (2005). "Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences." In: *Journal of Experimental Psychology: General* 134.2, pp. 222–241. DOI: 10.1037/0096-3445.134.2.222. URL: http://dx.doi.org/10.1037/0096-3445.134.2.222.

DeLong, Katherine A, Thomas P Urbach, and Marta Kutas (July 10, 2005). "Probabilistic word pre-activation during language comprehension inferred from electrical brain activity". In: *Nature Neuroscience* 8.8, pp. 1117–1121. DOI: 10.1038/nn1504. URL: http://dx.doi.org/10.1038/nn1504.

Dudley, Homer (1939). "The vocoder". In: *Bell Laboratories Record* 18.4, pp. 122–126.

Dupoux, Emmanuel and Kerry Green (1997). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes." In: *Journal of Experimental Psychology: Human Perception and Performance* 23.3, pp. 914–927. DOI: 10.1037/0096-1523.23.3.914. URL: http://dx.doi.org/10.1037/0096-1523.23.3.914.

Eckert, Mark A., Susan Teubner-Rhodes, and Kenneth I. Vaden (July 2016). "Is Listening in Noise Worth It? The Neurobiology of Speech Recognition in Challenging Listening Conditions". In: *Ear & Hearing* 37.1, 101S–110S. DOI: 10.1097/aud.0000000000000300. URL: http://dx.doi.org/10.1097/aud.0000000000000300.

Erb, J. et al. (June 26, 2013). "The Brain Dynamics of Rapid Perceptual Adaptation to Adverse Listening Conditions". In: *Journal of Neuroscience* 33.26, pp. 10688–10697. DOI: 10.1523/jneurosci.4596-12.2013. URL: http://dx.doi.org/10.1523/JNEUROSCI.4596-12.2013.

Erb, Julia (2014). "The neural dynamics of perceptual adaptation to degraded speech". Doctoral dissertation. Universität Leipzig, p. 211.

Faulkner, Andrew, Stuart Rosen, and Tim Green (2012). "Comparing live to recorded speech in training the perception of spectrally shifted noise-vocoded speech". In: *The Journal of the Acoustical Society of America* 132.4, EL336–EL342. DOI: 10.1121/1.4754432.

Fletcher, Harvey (1929). "Speech and Hearing". In: D. Van Nostrand Company.

Frisson, Steven, Keith Rayner, and Martin J. Pickering (2005). "Effects of contextual predictability and transitional probability on eye movements during reading". In: *Journal of Experimental Psychology: Learning Memory and Cognition* 31.5, pp. 862–877. DOI: 10.1037/0278-7393.31.5.862.

Greenwood, Donald D. (1990). "A cochlear frequency-position function for several species—29 years later". In: *Journal of the Acoustical Society of America* 87.6, pp. 2592–2605. DOI: 10.1121/1.399052.

Grueber, C. E. et al. (Jan. 27, 2011). "Multimodel inference in ecology and evolution: challenges and solutions". In: *Journal of Evolutionary Biology* 24.4, pp. 699–711. DOI:

10.1111/j.1420-9101.2010.02210.x. URL: http://dx.doi.org/10.1111/j.1420-9101.2010.02210.x.

Guediche, Sara et al. (2014). "Speech perception under adverse conditions: Insights from behavioral, computational, and neuroscience research". In: *Frontiers in Systems Neuroscience* 7.Jan, pp. 1–16. DOI: 10.3389/fnsys.2013.00126.

Hartwigsen, Gesa, Thomas Golombek, and Jonas Obleser (July 2015). "Repetitive transcranial magnetic stimulation over left angular gyrus modulates the predictability gain in degraded speech comprehension". In: *Cortex* 68, pp. 100–110. DOI: 10.1016/j.cortex.2014.08.027. URL: http://dx.doi.org/10.1016/j.cortex.2014.08.027.

Heilbron, Micha et al. (2020). "A hierarchy of linguistic predictions during natural language comprehension". In: *bioRxiv*. DOI: 10.1101/2020.12.03.410399.

Jaeger, T. Florian (2008). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models". In: *Journal of Memory and Language* 59.4, pp. 434–446. DOI: 10.1016/j.jml.2007.11.007. URL: http://dx.doi.org/10.1016/j.jml.2007.11.007.

James, William (1890). *The Principles of Psychology*. New York: Henry Holt.

Kaiser, E and J Trueswell (Dec. 2004). "The role of discourse context in the processing of a flexible word-order language". In: *Cognition* 94.2, pp. 113–147. DOI: 10.1016/j.cognition.2004.01.002. URL: http://dx.doi.org/10.1016/j.cognition.2004.01.002.

Kamide, Yuki, Gerry T.M. Altmann, and Sarah L Haywood (July 2003). "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements". In: *Journal of Memory and Language* 49.1, pp. 133–156. DOI: 10.1016/s0749-596x(03)00023-8. URL: http://dx.doi.org/10.1016/S0749-596X(03)00023-8.

Knoeferle, Pia et al. (Feb. 2005). "The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events". In: *Cognition* 95.1, pp. 95–127. DOI: 10.1016/j.cognition.2004.03.002. URL: http://dx.doi.org/10.1016/j.cognition.2004.03.002.

Kutas, Marta and Kara D. Federmeier (Jan. 10, 2011). "Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)". In: *Annual Review of Psychology* 62.1, pp. 621–647. DOI: 10.1146/annurev.psych.093008.131123. URL: http://dx.doi.org/10.1146/annurev.psych.093008.131123.

Kutas, Marta and Steven A. Hillyard (Jan. 1984). "Brain potentials during reading reflect word expectancy and semantic association". In: *Nature* 307.5947, pp. 161–163. DOI: 10.1038/307161a0. URL: http://dx.doi.org/10.1038/307161a0.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13. DOI: 10.18637/jss.v082.i13. URL: http://dx.doi.org/10.18637/jss.v082.i13.

Loizou, Philipos C., Michael Dorman, and Zhemin Tu (Oct. 1999). "On the number of channels needed to understand speech". In: *The Journal of the Acoustical Society of America* 106.4, pp. 2097–2103. DOI: 10.1121/1.427954. URL: http://dx.doi.org/10.1121/1.427954.

Mattys, Sven L. et al. (2012). "Speech recognition in adverse conditions: A review". In: *Language and Cognitive Processes* 27.7-8, pp. 953–978. DOI: 10.1080/01690965.2012.705006.

Moulines, Eric and Francis Charpentier (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". In: *Speech Communication* 9.1990, pp. 453–467.

Nieuwland, Mante S et al. (2018). "Large-scale replication study reveals a limit on probabilistic prediction in language comprehension". In: *eLife* 7, pp. 1–24. DOI: 10.7554/elife.33468.

Obleser, Jonas (Dec. 2014). "Putting the Listening Brain in Context". In: *Language and Linguistics Compass* 8.12, pp. 646–658. DOI: 10.1111/lnc3.12098. URL: http://dx.doi.org/10.1111/lnc3.12098.

Obleser, Jonas and S. A. Kotz (June 26, 2009). "Expectancy Constraints in Degraded Speech Modulate the Language Comprehension Network". In: *Cerebral Cortex* 20.3, pp. 633–640. DOI: 10.1093/cercor/bhp128. URL: http://dx.doi.org/10.1093/cercor/bhp128.

Obleser, Jonas and Sonja A. Kotz (2011). "Multiple brain signatures of integration in the comprehension of degraded speech". In: *NeuroImage* 55.2, pp. 713–723. DOI: 10.1016/j.neuroimage.2010.12.020. URL: https://www.sciencedirect.com/science/article/pii/S1053811910016034.

Obleser, Jonas, R. J. S. Wise, et al. (Feb. 28, 2007). "Functional Integration across Brain Regions Improves Speech Perception under Adverse Listening Conditions". In: *Journal of Neuroscience* 27.9, pp. 2283–2289. DOI: 10.1523/jneurosci.4663-06.2007. URL: http://dx.doi.org/10.1523/JNEUROSCI.4663-06.2007.

Pickering, Martin J. and Chiara Gambi (Oct. 2018). "Predicting while comprehending language: A theory and review." In: *Psychological Bulletin* 144.10, pp. 1002–1044. DOI: 10.1037/bul0000158. URL: http://dx.doi.org/10.1037/bul0000158.

Rayner, Keith et al. (Apr. 2011). "Eye movements and word skipping during reading: Effects of word length and predictability." In: *Journal of Experimental Psychology: Human Perception and Performance* 37.2, pp. 514–528. DOI: 10.1037/a0020990. URL: http://dx.doi.org/10.1037/a0020990.

Richards, Shane A., Mark J. Whittingham, and Philip A. Stephens (Aug. 2011). "Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework". In: *Behavioral Ecology and Sociobiology* 65.1, pp. 77–89. DOI: 10.1007/s00265-010-1035-8. URL: http://dx.doi.org/10.1007/s00265-010-1035-8.

Rosen, Stuart, Andrew Faulkner, and Lucy Wilkinson (Dec. 1999). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants". In: *The Journal of the Acoustical Society of America* 106.6, pp. 3629–3636. DOI: 10.1121/1.428215. URL: http://dx.doi.org/10.1121/1.428215.

Samuel, Arthur G. and Tanya Kraljic (Aug. 2009). "Perceptual learning for speech". In: *Attention, Perception, & Psychophysics* 71.6, pp. 1207–1218. DOI: 10.3758/app.71.6.1207. URL: http://dx.doi.org/10.3758/APP.71.6.1207.

Shannon, R. V., F.-G. Zeng, et al. (Oct. 13, 1995). "Speech Recognition with Primarily Temporal Cues". In: *Science* 270.5234, pp. 303–304. DOI: 10.1126/science.270.5234.303. URL: http://dx.doi.org/10.1126/science.270.5234.303.

Shannon, Robert, Qian-Jie Fu, and John Galvin Iii (Apr. 1, 2004). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation". In: *Acta Oto-Laryngologica* 124.0, pp. 50–54. DOI: 10.1080/03655230410017562. URL: http://dx.doi.org/10.1080/03655230410017562.

Sheldon, Signy, M. Kathleen Pichora-Fuller, and Bruce A. Schneider (Oct. 2008a). "Effect of age, presentation method, and learning on identification of noise-vocoded words". In: *The Journal of the Acoustical Society of America* 123.1, pp. 476–488. DOI: 10.1121/1.2805676. URL: http://dx.doi.org/10.1121/1.2805676.

— (Feb. 2008b). "Priming and sentence context support listening to noise-vocoded speech by younger and older adults". In: *The Journal of the Acoustical Society of America* 123.1, pp. 489–499. DOI: 10.1121/1.2783762. URL: http://dx.doi.org/10.1121/1.2783762.

Sommers, Mitchell S., Lynne C. Nygaard, and David B. Pisoni (Sept. 1994). "Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude". In: *The Journal of the Acoustical Society of America* 96.3, pp. 1314–1324. DOI: 10.1121/1.411453. URL: http://dx.doi.org/10.1121/1.411453.

Staub, Adrian (Aug. 2015). "The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation". In: *Language and Linguistics Compass* 9.8, pp. 311–327. DOI: 10.1111/lnc3.12151. URL: http://dx.doi.org/10.1111/lnc3.12151.

Strauß, Antje, Sonja A. Kotz, and Jonas Obleser (Aug. 2013). "Narrowed Expectancies under Degraded Speech: Revisiting the N400". In: *Journal of Cognitive Neuroscience* 25.8, pp. 1383–1395. DOI: 10.1162/jocn_a_00389. URL: http://dx.doi.org/10.1162/jocn_a_00389.

"The vocoder" (1940). In: *Nature* 145.3665, p. 157. DOI: 10.1038/145157a0.

Vaden, Kenneth I., Stefanie E. Kuchinsky, Jayne B. Ahlstrom, Judy R. Dubno, et al. (2015). "Cortical activity predicts which older adults recognize speech in noise and when". In: *Journal of Neuroscience* 35.9, pp. 3929–3937. DOI: 10.1523/JNEUROSCI.2908-14.2015.

Vaden, Kenneth I., Stefanie E. Kuchinsky, Jayne B. Ahlstrom, Susan E. Teubner-Rhodes, et al. (Dec. 18, 2015). "Cingulo-Opercular Function During Word Recognition in Noise for Older Adults with Hearing Loss". In: *Experimental Aging Research* 42.1, pp. 67–82. DOI: 10.1080/0361073x.2016.1108784. URL: http://dx.doi.org/10.1080/0361073X.2016.1108784.

Vaden, Kenneth I., Stefanie E. Kuchinsky, Stephanie L. Cute, et al. (2013). "The cingulo-opercular network provides word-recognition benefit". In: *Journal of Neuroscience* 33.48, pp. 18979–18986. DOI: 10.1523/JNEUROSCI.1417-13.2013.

Verhelst, Werner and Marc Roelands (1993). "Overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech". In: *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing* 2.1, pp. 2–5. DOI: 10.1109/icassp.1993.319366.

Whitmire, Clarissa J. and Garrett B. Stanley (2016). "Rapid Sensory Adaptation Redux: A Circuit Perspective". In: *Neuron* 92.2, pp. 298–315. DOI: 10.1016/j.neuron.2016.09.046. URL: http://dx.doi.org/10.1016/j.neuron.2016.09.046.

Wlotko, Edward W. and Kara D. Federmeier (Apr. 3, 2012). "Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension". In: *Psychophysiology* 49.6, pp. 770–785. DOI: 10.1111/j.1469-8986.2012.01366.x. URL: http://dx.doi.org/10.1111/j.1469-8986.2012.01366.x.

Xiang, Ming and Gina Kuperberg (2015). "Reversing expectations during discourse comprehension". In: *Language, Cognition and Neuroscience* 30.6, pp. 648–672. DOI: 10.1080/23273798.2014.995679. URL: http://dx.doi.org/10.1080/23273798.2014.995679.