

# **The role of acoustic periodicity in perceiving speech**

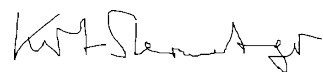
*Kurt Steinmetzger*

A dissertation submitted in fulfillment of the requirements for the  
degree of Doctor of Philosophy of University College London.

Department of Speech, Hearing and Phonetic Sciences

January 23, 2017

I, Kurt Steinmetzger, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

A handwritten signature in black ink, appearing to read "Kurt Steinmetzger". The signature is written in a cursive, flowing style with some capitalization.

## **Abstract**

This thesis investigated the role of one important acoustic feature, periodicity, in the perception of speech. In the context of this thesis, periodicity denotes that a speech sound is voiced, giving rise to a sonorous sound quality sharply opposed to that of noisy unvoiced sounds. In a series of behavioural and electroencephalography (EEG) experiments, it was tested how the presence and absence of periodicity in both target speech and background noise affects the ability to understand speech, and its cortical representation. Firstly, in quiet listening conditions, speech with a natural amount of periodicity and completely aperiodic speech were equally intelligible, while completely periodic speech was much harder to understand. In the presence of a masker, however, periodicity in the target speech mattered little. In contrast, listeners substantially benefitted from periodicity in the masker and this so-called masker-periodicity benefit (MPB) was about twice as large as the fluctuating-masker benefit (FMB) obtained from masker amplitude modulations. Next, cortical EEG responses to the same three target speech conditions were recorded. In an attempt to isolate effects of periodicity and intelligibility, the trials were sorted according to the correctness of the listeners' spoken responses. More periodicity rendered the event-related potentials more negative during the first second after sentence onset, while a slow negativity was observed when the sentences were more intelligible. Additionally, EEG alpha power (7–10 Hz) was markedly increased before the least intelligible sentences. This finding is taken to indicate that the listeners have not been fully focussed on the task before these trials. The same EEG data were also analysed in the frequency domain, which revealed a distinct response pattern, with more theta power (5–6.3 Hz) and a trend for less beta power (11–18 Hz), in the fully periodic condition, but again no differences between the other two

conditions. This pattern may indicate that the subjects internally rehearsed the sentences in the periodic condition before they verbally responded. Crucially, EEG power in the delta range (1.7–2.7 Hz) was substantially increased during the second half of intelligible sentences, when compared to their unintelligible counterparts. Lastly, effects of periodicity in the perception of speech in noise were examined in simulations of cochlear implants (CIs). Although both were substantially reduced, the MPB was still about twice as large as the FMB, highlighting the robustness of periodicity cues, even with the limited access to spectral information provided by simulated CIs. On the other hand, the larger absolute reduction of the MBP compared to normal-hearing also suggests that the inability to exploit periodicity cues may be an even more important factor in explaining the poor performance of CI users than the inability to benefit from masker fluctuations.

## Acknowledgements

This thesis documents my ongoing drift away from the humanities towards hard science within the field of sound. I am thankful to everybody who has been involved along the way preparing me for this experience: Wolfgang Schönflug, Elena Ungeheuer, Diana Deutsch, Stefan Koelsch, Martin Rohrmeier, and of course my dear parents Ines and Ulrich Steinmetzger. Most importantly, I am very grateful to my partner Natalie Berger for agreeing to leave our cherished environment in Berlin to start over in London.

I have spent four full years on this thesis, which was only possible with the generous funding I have received from the European Commission as a member of the Marie Curie Initial Training Network INSPIRE (*Investigating speech in realistic environments!*). Many thanks to Paul Iverson for setting up this opportunity and thanks to all INSPIRE members for their ideas and company during our meetings. I am handing in this thesis shortly after the Brexit referendum, sadly making this type of funding a phased-out model in the UK. Part of the INSPIRE package was a brilliant secondment at Torsten Dau's group at Danish Technical University in autumn 2014. I am grateful to Torsten and Johannes Zaar for their supervision and advice during this time, and our ongoing collaboration. In particular, I thank them for introducing me to the world of amplitude modulations, which feature in pretty much everything I have been working on since then.

I am of course much obliged to my supervisor Stuart Rosen, who has been incredibly responsive, supportive, and knowledgeable throughout these years, but at the same time happy to keep me on a long leash. He also convinced me that, with the right attitude, our line of research is not nerdy, but something groovy that keeps you young.

Thanks also to the numerous people who have supported this work in some form or other: Steve Nevard for excellent technical support in the lab, Alan O Cinneide, whose software was used to generate the Liljencrants-Fant glottal pulses, Hideki Kawahara and Jeanne Clarke for code and helpful advice concerning TANDEM-STRAIGHT, as well as Martin Cooke, Tim Green, Jyrki Tuomainen, and several anonymous reviewers for valuable comments. I also thank Hannah Nash and David Morris, with whom I have had productive collaborations alongside the PhD.

Many thanks also go to the SHaPS community for providing such a *gemütlich* environment: Mauricio Figueroa, Paula Sandoval, and Nico, José Joaquín Atria, Gisela Tomé Lourido, Faith Chiu, Jieun Song, Petra Hödl, Tim Schoof, Daniel Friedrichs, Nada Alsari, Wafaa Alshangiti, Yue Zhang, Michèle Pettinato, Laurianne Cabrera, Axelle Calcus, Gaston Hilkhuisen, Outi Tuomainen, Bronwen Evans, and Paul Iverson. Finally, I thank the Roehampton football gang for all the kicking and rushing over the past few years.

## Contents

<b>1</b>	<b>General introduction</b>	<b>11</b>
1.1	Periodicity: a disambiguation and introduction . . . . .	11
1.2	Theoretical background . . . . .	12
1.3	Chapter overview . . . . .	17
<b>2</b>	<b>The role of periodicity in perceiving speech in quiet and in background noise</b>	<b>20</b>
2.1	Introduction . . . . .	20
2.2	Experiment 1. Short introduction and rationale . . . . .	23
2.3	Experiment 1. Methods . . . . .	24
2.4	Experiment 1. Results and discussion . . . . .	27
2.5	Experiment 2. Short introduction and rationale . . . . .	29
2.6	Experiment 2. Methods . . . . .	31
2.7	Experiment 2. Results and discussion . . . . .	34
2.8	Experiment 3. Short introduction and rationale . . . . .	40
2.9	Experiment 3. Methods . . . . .	41
2.10	Experiment 3. Results and discussion . . . . .	42
2.11	General discussion . . . . .	44
2.12	Summary and conclusion . . . . .	49
<b>3</b>	<b>Effects of acoustic periodicity, intelligibility, and pre-stimulus alpha power on the event-related potentials in response to speech</b>	<b>51</b>
3.1	Introduction . . . . .	51
3.2	Methods . . . . .	54

3.3	Results . . . . .	60
3.4	Discussion . . . . .	66
3.5	Conclusion . . . . .	69
<b>4</b>	<b>Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech</b>	<b>71</b>
4.1	Introduction . . . . .	71
4.2	Methods . . . . .	75
4.3	Results . . . . .	81
4.4	Discussion . . . . .	87
4.5	Conclusion . . . . .	91
<b>5</b>	<b>The role of periodicity in perceiving speech in background noise with simulated cochlear implants</b>	<b>92</b>
5.1	Introduction . . . . .	92
5.2	Methods . . . . .	94
5.3	Results . . . . .	99
5.4	Discussion . . . . .	101
5.5	Conclusion . . . . .	104
<b>6</b>	<b>General Discussion</b>	<b>105</b>
6.1	Summary of the main results and their implications . . . . .	105
6.2	Psychophysical data . . . . .	108
6.3	Electrophysiological data . . . . .	109
6.4	Perspectives . . . . .	110
6.5	Conclusions . . . . .	113
<b>7</b>	<b>References</b>	<b>114</b>



## List of Figures

2.1	Experiments 1–3. Target speech conditions . . . . .	25
2.2	Experiment 1. Proportion correct scores . . . . .	29
2.3	Experiments 2 and 3. Background maskers . . . . .	33
2.4	Experiments 2 and 3. Speech reception thresholds . . . . .	35
2.5	Experiments 2 and 3. Fluctuating-masker benefits . . . . .	36
2.6	Experiments 2 and 3. Masker-periodicity benefits . . . . .	37
2.7	Experiment 2. Psychometric functions . . . . .	39
2.8	Experiment 3. Psychometric functions . . . . .	45
3.1	Stimuli . . . . .	56
3.2	Behavioural data . . . . .	60
3.3	Periodicity . . . . .	62
3.4	Intelligibility . . . . .	64
3.5	Pre-stimulus alpha power . . . . .	66
4.1	Stimuli . . . . .	75
4.2	Behavioural data . . . . .	81
4.3	Periodicity . . . . .	82
4.4	Periodicity: pairwise comparisons . . . . .	83
4.5	Periodicity: post-hoc . . . . .	84
4.6	Intelligibility . . . . .	85
4.7	Intelligibility: pairwise comparisons . . . . .	86
4.8	Intelligibility: acoustic characteristics . . . . .	87
5.1	Signal processing . . . . .	96
5.2	Stimuli . . . . .	98

5.3	Speech reception thresholds . . . . .	99
5.4	Fluctuating-masker benefits . . . . .	100
5.5	Masker-periodicity benefits . . . . .	101
5.6	F0 contours . . . . .	102
5.7	Modulation spectrograms . . . . .	103

# Chapter 1

## General introduction

### 1.1 Periodicity: a disambiguation and introduction

Periodicity, the repetition of events at regular intervals, is ubiquitous in both nature and society. Examples are bodily functions, the seasons of the year, political elections, sports events, and many more. Even within the field of acoustics, there are numerous examples for things that can recur periodically, such as waveforms, amplitude and frequency modulations, phase angles, etcetera. Consequently, the vague term periodicity requires clarification: In the context of the current thesis, periodicity solely means that a given sound has a pitch in the range of the human voice. In other words, periodicity here denotes that speech sounds are voiced as opposed to unvoiced, that is aperiodic.

The alternation of periodic and aperiodic segments is a crucial acoustic feature of speech across languages. However, its role in the perception of speech in quiet and in background noise has not been thoroughly investigated and remains poorly understood. One reason for this is that the most common way to describe the temporal properties of speech sounds, which focusses on the signal processing taking place in the inner ear, only distinguishes between slow envelope modulations and faster modulations, referred to as temporal fine-structure (e.g. [Lorenzi et al., 2006](#); [Moore, 2008](#)). However, a threefold distinction that emphasises the linguistic aspects of temporal information in speech has also been proposed and explicitly includes periodicity information as an additional factor ([Rosen, 1992](#)). As laid out in that paper, (quasi-)periodic fluctuations caused by the vibrations of the vocal folds (at rates between about 50–500 Hz) are critical for transmitting prosodic

features and phonetic manner information. Furthermore, the obvious acoustic contrast between sonorous periodic speech sounds and noisy aperiodic ones (with much higher fluctuation rates in the range of 5–10 kHz) is highlighted. Importantly, the most fundamental limitation of current cochlear implants (CIs) is that spectral cues, and voice pitch information in particular, are hardly available to the listener (e.g. [Shannon et al., 1995](#); [Wilson and Dorman, 2008](#)). This deficit is thought to be the main reason for the limited speech recognition performance of CI users, particularly in noisy environments. Several studies have attempted to improve the transmission of periodicity cues either by changing the devices themselves (e.g. [Green et al., 2004](#); [Green et al., 2005](#)) or by exploiting residual hearing at low frequencies with combined acoustic and electric hearing (reviewed in [Turner et al., 2008](#)).

As sounds with a pitch tend to be perceived as separate auditory streams (e.g. [Bregman, 1990](#); [Oxenham, 2008](#)), the main hypothesis of this thesis is that periodicity should aid the segregation of target speech and masker. Secondly, this thesis is based on the idea that periodicity is such a central feature of human speech that it should not be overlooked when investigating its neural correlates. In order to test these two assumptions, synthesised speech and background noises are introduced that both vary regarding their amount of periodicity.

## 1.2 Theoretical background

### 1.2.1 Psychophysical studies

An abundance of behavioural studies investigating the perception of speech in noise has focussed on how small gaps or amplitude fluctuations of a background noise affect the ability of different groups of listeners to understand speech (see section [2.1](#) for a selection). The central finding of these studies was that any form of hearing impairment impedes the ability to benefit from these masker interruptions. One hypothesis, which has in part motivated the current work, is that the access to temporal fine-structure and periodicity information is crucial in order to benefit from such masker fluctuations (e.g. [Hopkins et al., 2008](#); [Hopkins and Moore, 2009](#); [Lorenzi et al., 2006](#); [Moore, 2008](#)). However, this claim has been a topic of de-

bate (Moore, 2012) and one of the few studies (Freyman et al., 2012) explicitly focussing on the contrast of periodicity and aperiodicity has found no support for it. In this study, aperiodic whispered speech and unprocessed speech were presented in steady or modulated speech-shaped noise and the fluctuating-masker benefit did not substantially differ across the two target speech conditions.

Other previous work that resembles the one presented in the current thesis to some extent, is the series of studies by de Cheveigné and colleagues (de Cheveigné, 1993, 1998; de Cheveigné et al., 1995; de Cheveigné et al., 1997a; de Cheveigné et al., 1997b), leading to the formulation of the theory of harmonic cancellation. Using artificial vowels that were presented concurrently, it was found that it was easier to correctly perceive the target vowel if the other one was harmonic, that is periodic. On the other hand, it did not affect performance if the target vowel itself was harmonic or inharmonic. This effect was also observed when whole sentences and harmonic or inharmonic complex tone maskers with a constant fundamental frequency were used (Deroche and Culling, 2011). In contrast to these studies, all the materials used in this thesis will be derived from recordings of real speech, with the exception of the maskers based on white noise, enabling a more realistic investigation.

Previous work has also shown that periodic maskers are generally less effective than aperiodic ones, which has been attributed to the fact that they allow the listeners to spectrally glimpse portions of the target speech in between the lower masker harmonics that are resolved in the auditory periphery (Deroche et al., 2014a, 2014b). Lastly, recent work by Stone and colleagues (Stone et al., 2011; Stone et al., 2012) has demonstrated that the random amplitude modulations of aperiodic maskers, such as speech-shaped noise, are the primary reason for why the mask a competing speech signal so effectively. This has challenged the traditional view that a masker either interferes with the target speech because of an overlap of spectral energy or, in the case of a speech masker, its linguistic information (Brungart, 2001). Note that even though the term ‘steady’ noise is thus imprecise, it is used in the current thesis to be consistent with earlier research.

In summary, the current thesis seeks to add to the established finding that masker periodicity aids its segregation from a target speech signal, by using more realistic periodic maskers with a dynamically varying pitch contour derived from speech. Secondly, it will be tested whether the intelligibility of a target speech signal embedded in background noise indeed does not depend on whether it is periodic or aperiodic, particularly when the masker is amplitude modulated.

### 1.2.2 Neurophysiological studies

Magneto- and electroencephalographic (M/EEG) signals recorded from the human scalp have traditionally been analysed in the time domain, by comparing the average waveforms of neural activity across experimental conditions. In the auditory modality, investigations of these event-related responses have mostly focussed on the period early after stimulus onset (~0–300 ms; e.g. [Picton et al., 1974](#); [Pratt, 2011](#)), where the responses are dominated by the acoustic properties of the stimuli. Regarding the effect of periodicity, it has been found that non-speech sounds that possess a pitch generally lead to greater response amplitudes (e.g. [Chait et al., 2006](#); [Gutschalk et al., 2002](#); [Gutschalk et al., 2004](#)). Importantly, these studies have shown that this difference is not only present after stimulus onset but can persist for several hundred milliseconds. In line with this, a series of studies by Yrttiaho and colleagues ([Yrttiaho et al., 2008](#); [Yrttiaho et al., 2010](#); [Yrttiaho et al., 2011](#)) has shown that auditory cortical responses to periodic vowels are stronger and emanate from a more anterior cortical source, when compared to aperiodic versions of these.

However, effects of periodicity have not been examined over the course of whole sentences and the same holds true for effects of intelligibility. Regarding the latter, a particular problem is that the intelligibility of the speech materials is usually lowered by an acoustic degradation of them, meaning that these two factors are hard to disentangle (e.g. [Becker et al., 2013](#); [Ding et al., 2014](#); [Obleser and Kotz, 2011](#); [Wöstmann et al., 2015b](#)). In the current thesis, it will be attempted to overcome this limitation by lowering the intelligibility of the stimuli and then sorting the individual trials according to the correctness of the listeners' spoken responses.

It is hypothesised that after controlling for systematic acoustic confounds, slow negative potentials, which are thought to reflect domain-general cognitive processing such as increased working-memory load and attention ([Birbaumer et al., 1990](#); [He and Raichle, 2009](#); [Wöstmann et al., 2015b](#)), will be larger in response to more intelligible speech.

In recent years, the focus of neurophysiological investigations of speech perception has increasingly shifted towards more complex time-frequency analyses of M/EEG data ([Giraud and Poeppel, 2012](#); [Lakatos et al., 2016](#); [Weisz and Obleser, 2014](#); [Wöstmann et al., 2016](#)). Depending on whether the single-trial waveforms are averaged before or after the time-frequency transform, such analyses allow the separate estimation of evoked neural activity, which is time- and phase-locked to a given stimulus event and thought to be mainly driven by the stimulus in a bottom-up fashion, and total neural activity, time- but not necessarily phase-locked and taken to also reflect cognitive processes emanating from the cortex in a top-down manner, across frequency ([Tallon-Baudry and Bertrand, 1999](#)). In contrast to traditional waveform analyses, where non-phase-locked activity is mostly cancelled out during the averaging process, thereby isolating the evoked response, this enables a more fine-grained analysis of the data. When measured at the scalp, a power increase in a given neural frequency band indicates that the firing pattern of a large number of neurons temporarily synchronises, resulting in a periodic fluctuation of neural excitability, a so-called neural oscillation (e.g. [Buzsáki and Draguhn, 2004](#); [Klimesch, 2012](#)).

Accordingly, two lines of research have emerged, one investigating how temporal stimulus properties affect the evoked neural response, the other focussing on attention-related processing, dominated by the induced response. Firstly, it has been shown that low-frequency neural oscillations, particularly in the theta band (~4–7 Hz), entrain to the broadband amplitude envelope of a speech signal (for reviews see [Giraud and Poeppel, 2012](#); [Peelle and Davis, 2012](#)), which has the strongest amplitude modulations in this frequency range, and it has even been claimed that this effect is stronger if the speech is intelligible ([Peelle et al., 2013](#)).

Secondly, several studies have investigated modulations of alpha power (~7–13 Hz) in speech perception experiments. In demanding tasks, such as listening to speech in background noise, alpha power has been shown to increase (e.g. [Obleser et al., 2012](#); [Wilsch et al., 2014](#); [Wöstmann et al., 2015a](#)), which has been explained with the notion that high alpha activity inhibits task-irrelevant brain regions ([Klimesch, 2007](#); [Jensen and Mazaheri, 2010](#); [Strauß et al., 2014b](#)). Strong support for this idea also comes from studies showing that when subjects are asked to attend speech played to one ear and ignore sounds played to the other, alpha power increases in the ipsilateral side of the brain, relative to the contralateral side (e.g. [Kerlin et al., 2010](#); [Wöstmann et al., 2016](#)). Importantly, recent evidence has shown that increased alpha power goes along with decreased envelope entrainment, suggesting that these two neural mechanisms complement each other ([Lakatos et al., 2016](#); [Wöstmann et al., 2016](#)). On the other hand, alpha power in response to speech presented in quiet was found to be suppressed, the less degraded and consequently more intelligible it was ([Becker et al., 2013](#); [Obleser and Weisz, 2012](#)). However, as previously mentioned, these studies are somewhat confounded by the fact that acoustic characteristics and intelligibility of the stimuli varied together.

Furthermore, low alpha power preceding stimulus presentation has repeatedly been shown to be a predictor of good performance in visual and somatosensory detection tasks (e.g. [Hanslmayr et al., 2007](#); [Romei et al., 2010](#); [Schubert et al., 2009](#); [Van Dijk et al., 2008](#)), which is in line with the idea that decreased power in the lower alpha band (7–10 Hz) is associated with heightened alertness and expectancy ([Klimesch, 1999](#)). However, it has not yet been tested whether the intelligibility of whole sentences is similarly affected by the amount of pre-stimulus alpha power.

Apart from the alpha band, however, it is to date largely unknown how the total neural response across frequency is affected by the presentation of speech and the current thesis seeks to explore this further. In particular, and akin to the time domain, the time-frequency analyses in this thesis will focus on slow changes of neural excitability, corresponding to the delta band (1–4 Hz), that are hypothesised to be larger when the speech is more intelligible. Responses in this frequency band



have so far not received much attention, in part because their analysis requires relatively long stimuli durations. To date, auditory studies reporting effects in the delta range have mostly been concerned with how temporal regularities of the stimuli lead to neural entrainment (Lakatos et al., 2005; Lakatos et al., 2016; Ding et al., 2014; Ding et al., 2016) or how the relationship of pre-stimulus delta power and phase affects the subsequent detection of tones (Herrmann et al., 2016). However, it has not been examined how delta activity changes in response to more or less intelligible speech, although it has been suggested that increased delta power in the frontal cortex may be an indicator of concentration and attention (Harmony, 2013).

### 1.3 Chapter overview

Chapter 2 functions as the base of this thesis and explores how the presence and absence of periodicity in both target speech and background masker affects speech perception in normal-hearing listeners. Three behavioural experiments are reported, the first of which examines how performance in quiet listening conditions changes as the amount of periodicity in the stimulus sentences varies. A new channel vocoder software is introduced that allows to either use white noise as source excitation (i.e. typical noise-vocoding) or a pulse train that follows the natural F0 contour of the recordings. These two sound sources are then employed to synthesise speech that is either completely aperiodic, preserves the natural mix of periodicity and aperiodicity, or is completely periodic. Additionally, the number of channels in the vocoder is varied over a wide range. In experiments 2 and 3, the same three types of target speech are presented in the presence of different background maskers that also vary regarding their periodicity and, additionally, have either steady-state or 10-Hz modulated envelopes. The results of experiment 1 are used to identify target speech conditions with equal intelligibility rates in quiet. Speech reception thresholds (SRTs) are measured for all combinations of targets and maskers to estimate effects of periodicity and to test whether the ability to benefit from masker envelope modulations depends on the presence of periodicity.

Chapter 3 extends the first experiment in the previous chapter by using the same stimuli, but recording the continuous electroencephalogram (EEG) along with the spoken responses. The aim of this experiment is twofold: Firstly, it examines how the amount of periodicity in the target speech is reflected in cortical EEG signals analysed in the time domain. Secondly, it attempts to identify effects of intelligibility in the same evoked EEG signals. In order to achieve this, the individual trials are sorted according to their intelligibility, allowing to analyse trials with different amounts of periodicity but similar intelligibility, and vice versa. In addition, it is tested whether the magnitude of alpha power before sentence onset affects the ability to correctly repeat it. This analysis builds upon findings in the visual domain that have associated decreased pre-stimulus alpha power with better task performance, and the idea that pre-stimulus power in the lower alpha band is negatively correlated with attention.

Chapter 4 is based on the same data and paradigm as chapter 3, but here the EEG signals are analysed in the frequency domain. The aim of this approach is to also consider non-phase locked (i.e. induced) changes in neural activity, which are lost when EEG signals are averaged in the time domain. As whole sentences are used as stimulus materials, the individual trials are several seconds long, which allows to analyse neural frequencies as low as 1 Hz. This distinguishes this study from earlier ones, but given that the meaning of a sentence unfolds over time, slow power changes in the delta range (1–4 Hz) are hypothesised to be particularly important when investigating speech intelligibility.

Chapter 5 investigates the role of periodicity in perceiving speech in noise after simulated CI signal processing. Current CIs provide the listener with very limited access to spectral information and this behavioural experiment examines whether temporal cues are sufficient to benefit from periodicity when attempting to segregate target speech and masker. The materials used are similar to those introduced in chapter 2, but in addition maskers with envelopes that are the inverse of a given target sentence envelope are used. The benefit obtained from masker amplitude modulations has repeatedly been shown to be very small in CI users and CI

simulations. By maximising opportunities to glimpse portions of the target speech, these maskers are intended to show whether this deficit is due to the susceptibility to energetic masking alone or if other factors need to be considered as well.

## Chapter 2

### The role of periodicity in perceiving speech in quiet and in background noise<sup>1</sup>

#### 2.1 Introduction

The production of any speech sound can be described by the interplay of a sound source and a vocal tract filter (e.g. [Fant, 1960](#)). Normally, either the periodically vibrating vocal cords (voiced speech) or aperiodic noise arising from constrictions in the vocal tract (voiceless speech) serve as source, although the two may occasionally overlap, such as in voiced fricatives. Clearly, the regular periodic pattern of voiced sounds stands in sharp acoustic contrast to noisy unvoiced sounds, and this contrast is also linguistically relevant since only the complex tones of voiced speech possess a pitch and thus allow the unambiguous signaling of intonation ([Rosen, 1992](#)). The component tones of voiced speech sounds stand in a harmonic relation and are not perceived individually. ‘All components point to a single source and meaning’ ([Rasch and Plomp, 1999](#), p. 95) and hence harmonicity can be said to add coherence to a sound stream (e.g. [Oxenham, 2008](#)). It thus seems reasonable to posit that periodicity in both target and masker helps to segregate a speech target from a background noise or an interfering talker.

On the other hand, de Cheveigné and colleagues ([de Cheveigné et al., 1995](#); [de Cheveigné et al., 1997b](#)) found that listeners benefit from harmonicity in the masker, but not the target speech. In these studies artificial steady-state vowels were used as both target and masker. Inharmonic vowels were much more effective

---

<sup>1</sup>This chapter has been published as: Steinmetzger, K. and Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *Journal of the Acoustical Society of America* 138, 3586–3599.

in masking the target vowel than harmonic ones, while harmonicity of the target vowel did not significantly affect performance. The results were taken to show that the auditory system seems to be able to cancel a harmonic masker out of the signal mixture. This so-called harmonic cancellation was also observed when unprocessed IEEE sentence materials were used as targets and the harmonicity of complex tone maskers was either blurred by modulating the masker F0 or further compromised by additionally reverberating the maskers (Deroche and Culling, 2011). Furthermore, Deroche and colleagues also provided evidence for spectral glimpsing in between resolved masker harmonics as an additional mechanism explaining the masking release found with harmonic complex maskers (Deroche et al., 2014a, 2014b). In sum, these findings emphasize the importance of periodicity in the masker, but not the target speech. However, these studies have computationally manipulated the harmonicity of the materials and so have not investigated the role of periodicity by contrasting voiced and unvoiced sounds as they occur in natural speech.

Although a lot of research in recent years has been devoted to the study of speech perception in noise and in particular the ability of listeners to ‘glimpse’ small sections of target speech in the troughs of an amplitude-modulated masker (Miller and Licklider, 1950), the role of periodicity information in this context has not been investigated thoroughly. It has been claimed that the ability to perceive the temporal fine-structure (TFS) in a target speech signal (i.e. any temporal information in speech, including periodicity information, apart from the slower envelope modulations) is essential in order to benefit from the dips of a fluctuating masker (Gnansia et al., 2009; Hopkins and Moore, 2009; Hopkins et al., 2008; Lorenzi et al., 2006). However, it is unclear to date whether TFS information plays a special role in glimpsing or is just as important for steady maskers (Moore, 2012).

Generally, normal-hearing listeners have been found to show rather large benefits in response to fluctuating maskers such as amplitude-modulated noise (e.g. Festen and Plomp, 1990; Bacon et al., 1998; Nelson et al., 2003; Fastl and Zwicker, 2007, p. 352) or interfering talkers (e.g. Festen and Plomp, 1990; Cullington and Zeng, 2008). Studies with hearing-impaired subjects (Festen and Plomp,

1990; Bacon et al., 1998; Peters et al., 1998) or spectrally degraded stimuli (Peters et al., 1998; Oxenham and Simonson, 2009) on the other hand, tend to find reduced fluctuating masker-benefits (FMBs), while studies with cochlear implant (CI) users and CI simulations find hardly any FMB (Nelson et al., 2003; Fastl and Zwicker, 2007, p. 352; Cullington and Zeng, 2008), or even a worsening of performance (referred to as ‘modulation interference’; Stickney et al., 2004; Kwon et al., 2012).

However, an important confound that has been pointed out by Bernstein and Grant (2009) is that the FMB is generally smaller at higher signal-to-noise ratios (SNRs). Freyman et al. (2012) illustrate this point with the typical shape of the psychometric functions, which are steeper for steady as compared to fluctuating maskers but converge at higher SNRs. Since any form of hearing-impairment or stimulus degradation will lead to increased SNRs generally, the ability to glimpse in these contexts might have been significantly underestimated in previous experiments.

Few studies to date have explicitly investigated the role of periodicity in the perception of speech in noise. Freyman et al. (2012) compared unprocessed speech to naturally produced ‘whispered’ speech and found no substantial differences in terms of the FMB obtained in steady and fluctuating speech-shaped noise, although the intelligibility of whispered speech was much lower. The authors concluded that for normal-hearing listeners, periodicity in the target speech has little effect on the ability to glimpse. However, due to the acoustic distinctiveness of whispered speech, which includes an altered consonant-vowel intensity ratio, it remains unclear whether the role of periodicity is similarly limited in normally articulated speech. Vestergaard and Patterson (2009), using artificially created ‘whispered’ speech, report that only the absence of periodicity cues in both target and masker (i.e. a combination of whispered targets and maskers) negatively affects performance. Thirdly, a study by Rosen et al. (2013) has recently compared speech reception thresholds (SRTs) of unprocessed and noise-vocoded target speech obtained in the presence of multi-talker babble, noise-vocoded babble, and speech-modulated noise. The most effective masker was in both cases the one that most closely re-

sembled the target speech, which again argues against the hypothesis that periodicity helps to segregate competing speech signals. The present study attempted to go beyond previous work by systematically investigating the role of periodicity using normally articulated speech only. Possible confounding factors such as the spectral resolution and intelligibility of the target speech were controlled for and informational masking effects were ruled out by using non-speech maskers only.

The amount of periodicity in the target speech was varied using different types of vocoders. While unvoiced speech can be reproduced adequately using a noise-vocoder that uses noise as source ([Shannon et al., 1995](#)), vocoders with periodic sources have been used less often in the literature ([Faulkner et al., 2000](#)). However, as originally described by [Dudley \(1939\)](#) and more recently by [Loizou \(2013, p. 54\)](#), voiced speech can be simulated efficiently with a vocoder using a pulse train carrier whose frequency follows the natural F0 contour of the original speech. The effects of periodicity in the masker were assessed by comparing aperiodic speech-shaped noise maskers to harmonic complex maskers with dynamically varying F0-contours based on real speech.

Experiment 1 tested whether the intelligibility of speech presented in quiet is affected by the amount of periodicity. In experiments 2 and 3 the amount of periodicity in both target and masker was varied and SRTs were measured in steady and fluctuating maskers. Experiments 2 and 3 differed only regarding the intelligibility of the target speech materials in quiet, so that the results can be presented in the same figures.

## **2.2 Experiment 1. Short introduction and rationale**

Experiment 1 investigated the role of periodicity in the perception of speech in quiet testing conditions by parametrically varying the amount of periodicity in the target speech along with the spectral resolution (i.e. the number of bands in the vocoder).

Aperiodic noise-vocoded speech has been used extensively in simulations of cochlear implants (e.g. [Shannon et al., 1995](#); [Fu and Nogaki, 2005](#); [Whitmal III et al., 2007](#)) and has become a popular tool for reducing the intelligibility of speech signals in neuroscience (e.g. [Scott et al., 2000](#); [Obleser and Weisz, 2012](#)). How-

ever, it has never been examined whether the absence of periodicity itself leads to a decrease in intelligibility. More generally, despite its salience it is unclear to date whether periodicity information is a beneficial cue in the absence of competing talkers or maskers.

In addition to completely unvoiced noise-vocoded speech and vocoded speech with a natural mix of voiced and unvoiced sections, the current experiment included completely voiced vocoded speech. The latter condition sounds very unnatural and is expected to be less intelligible in quiet. However, since periodicity is assumed to aid stream segregation, this condition will be of particular interest in the presence of background noise (experiments 2 and 3). An additional purpose of the current experiment was to identify conditions with similar intelligibility rates across the three processing conditions.

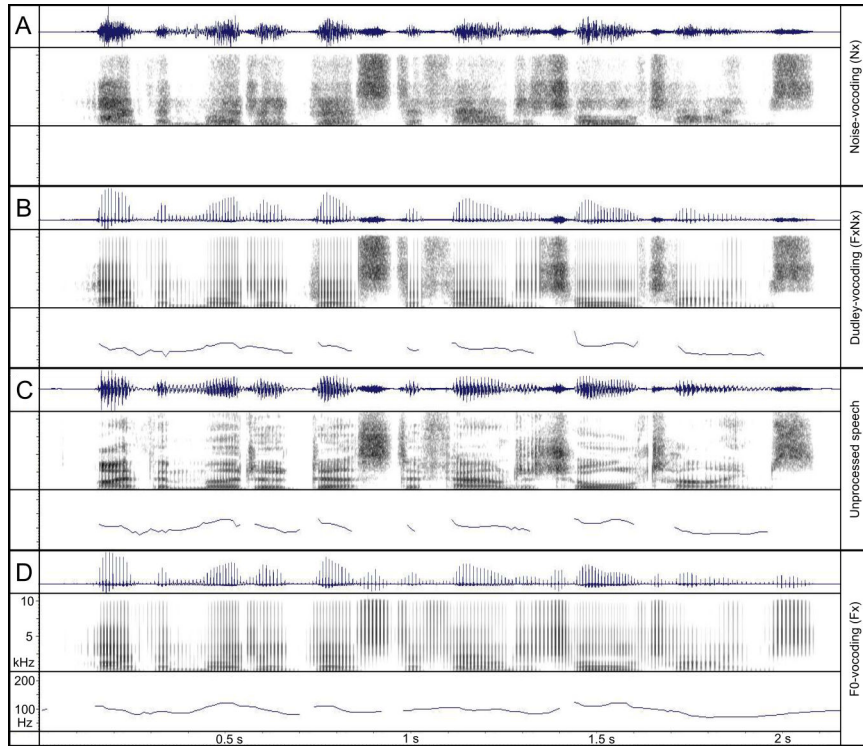
Experiment 1 consisted of 18 processed speech conditions as well as unprocessed speech as an additional condition. Participants were presented with noise-vocoded speech (henceforth referred to as the Nx), Dudley-vocoded speech ([Dudley, 1939](#)) with a natural mix of periodicity and aperiodicity (FxNx), and completely periodic F0-vocoded speech (Fx) with an F0 contour interpolated through unvoiced segments. These three types of stimuli also varied in the number of frequency bands used in their synthesis (6, 7, 8, 10, 12, or 16), and hence their intelligibility. An example sentence with 8 bands for all three processing conditions is shown in Fig. 2.1, along with the unprocessed version of the same sentence.

## 2.3 Experiment 1. Methods

### 2.3.1 Participants

Eleven normal-hearing listeners (six females) were tested. Their ages ranged from 19 to 35 with a mean of 27.3 years. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz.





**Figure 2.1:** Target speech conditions. Waveforms, wide-band spectrograms, and F0 contours for one example sentence (*Either mud or dust are found at all times.*) processed to have A) an aperiodic (noise-vocoding, Nx), B) mixed (Dudley-vocoding, FxNx), or D) periodic source excitation (F0-vocoding, Fx). Panel C) shows the unprocessed version of the same sentence for the purpose of comparison. The three processed sentences were all vocoded with eight frequency bands.

### 2.3.2 Stimuli

The targets used in this experiment were recordings of the IEEE sentences (Rothauser et al., 1969) spoken by an adult male Southern British English talker with a mean F0 of 121.5 Hz that were normalised to a common root-mean-square (RMS) level. The IEEE sentence corpus consists of 72 lists with 10 sentences each and is characterised by similar phonetic content across the lists and overall low semantic predictability. Every sentence contains five key words.

### 2.3.3 Signal processing

All stimulus materials were processed prior to the experiment using a channel vocoder implemented in MATLAB R2012b (Mathworks, Natick, MA). For all three processing conditions (Nx, FxNx, and Fx) the original recordings of the IEEE sen-

tences were first band-pass filtered into 6, 7, 8, 10, 12 or 16 bands using zero phase-shift sixth-order Butterworth filters. The filter spacing was based on equal basilar membrane distance ([Greenwood, 1990](#)) across a frequency range of 0.1 to 11 kHz. The output of each filter was full-wave rectified and low-pass filtered at 30 Hz (zero phase-shift fourth-order Butterworth) in order to extract the amplitude envelope. The low cutoff value was chosen in order to ensure that no temporal periodicity cues were present. The final waveforms were low-pass filtered at 10 kHz (sixth-order elliptic).

For the noise-vocoded condition (Nx), the envelope from each band was then multiplied with a wide-band noise carrier. The resulting signal was again band-pass filtered using the same sixth-order Butterworth filters as in the first stage of the process. Before the signal was summed together, the output of each band was adjusted to the same RMS level as found in the original bands. For the Dudley-vocoded condition (FxFxNx), the envelope from each band was multiplied with either a wide-band noise carrier where the original speech was unvoiced, or a pulse train following the natural F0 contour when the original speech was voiced.

The F0 contours of each sentence were generated using ProsodyPro version 4.3 ([Xu, 2013](#)) implemented in PRAAT ([Boersma and Weenink, 2013](#)). The F0 extraction sampling rate was set to 100 Hz. The results were hand-corrected and the resulting values used to generate the pulse trains for the vocoder software described above. Based on these pulse files, additional F0 contours were created by interpolation through unvoiced sections and periods of silence in order to synthesise fully periodic vocoded speech (Fx). The interpolation was done using piecewise cubic Hermite interpolation in logarithmic frequency. The start and end points of each contour were anchored to the median frequency of the sentence.

#### 2.3.4 Procedure

Every participant listened to two full IEEE lists (i.e. 20 sentences) per processing condition and was asked to repeat as many words as possible after every sentence. The verbal responses were logged by the experimenter before the next sentence was played (in terms of which of the roots of the 5 key words in each sentence were

correctly identified, so-called loose key word scoring). No feedback was given following the responses. The presentation and logging of the responses was carried out using locally developed MATLAB software. The experiment consisted of 19 conditions (3 vocoding conditions  $\times$  6 degrees of spectral resolution, and one additional condition with unprocessed target speech). Hence every participant was presented with 380 sentences in total. The order of the 19 processing conditions was fully randomised using a Latin Square design and the order of the IEEE lists was also randomised. Before being tested the subjects were familiarised with the materials by listening to 2 example sentences of each of the 18 processed conditions. Here every sentence was directly followed by its unprocessed counterpart. The total testing time, including hearing screening and familiarization, was about 1 hour and the subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating booth, with the computer signal being fed through the wall onto a separate monitor. The stimuli were converted with 24-bit resolution and a sampling rate of 22.05 kHz using an RME Babyface sound-card (Haimhausen, Germany) and presented over Sennheiser HD650 headphones (Wedemark, Germany) at a level of about 80 dB SPL over a frequency range of 75 Hz to 10.0 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

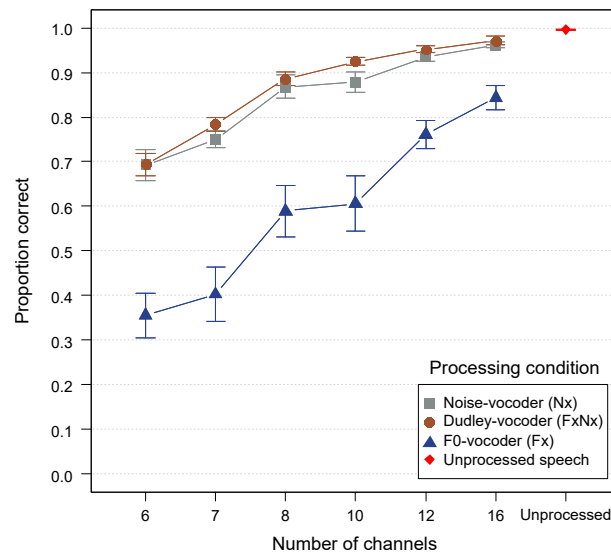
## 2.4 Experiment 1. Results and discussion

The proportion correct scores obtained are shown in Fig. 2.2. The Dudley-vocoded condition (FxNx) with a natural mix of periodicity and aperiodicity led to the highest percentage of correctly repeated key words but is closely followed by the noise-vocoded (Nx) condition, irrespective of the number of frequency bands. Fully periodic F0-vocoded speech (Fx) on the other hand was found to result in much lower intelligibility rates, with only 84% correctly repeated key words even with as much spectral detail as 16 frequency bands. Unprocessed speech was found to have an almost perfect intelligibility level with 99.6% correct key words, proving that the IEEE materials as such do not impose excessive memory demands despite their complexity.

The data were analysed using a generalised linear mixed effects model with a logistic link function that included target periodicity and spectral resolution as fixed factors and subjects as a random factor. The main effects of target periodicity [ $F(2,180) = 114.0, p < 0.001$ ] and spectral resolution [ $F(5,180) = 113.5, p < 0.001$ ] were found to be highly significant, but there was no interaction of the two [ $F(10,180) = 0.5, p = 0.89$ ]. The fixed coefficients furthermore showed that performance with F0-vocoded speech ( $-1.5, p < 0.001$ ), but not Dudley-vocoded speech ( $0.4, p = 0.24$ ), was significantly different from performance with noise-vocoded speech.

The fact that performance with noise-vocoded speech was not significantly worse than that with Dudley-vocoded speech suggests that the absence of any periodicity information, and hence also any prosody cues, is of minor importance in quiet testing conditions. Although voice pitch information is not essential for understanding English declarative sentences, it is still surprising that a cue as salient as periodicity transmits mostly redundant information. However, despite some important acoustic differences, noise-vocoded speech to some extent resembles whispered speech. Hence, listeners are likely to be at least implicitly familiar with this type of speech. Noise-vocoded speech also enables the listeners to use weaker correlated cues like intensity to distinguish between voiced and unvoiced consonants. Additionally, the spectral shape, which is well coded in a vocoder, gives strong cues to voicing, even in the absence of periodicity. Voiced speech is heavily weighted towards low frequencies, while voiceless excitation is typically weighted to the high frequencies.

The unnatural periodic energy in the F0-vocoded condition, especially in the frequency region above 4 kHz, on the other hand, might have substantially interfered with the listener's ability to correctly identify the individual sounds of the presented sentences. Since periodicity is such a dominant cue, weaker cues like intensity differences may not have been noticed. Similarly, for unvoiced fricatives like /s/ and /f/, for example, aperiodic energy at high frequencies is missing as a cue for identification and replaced by periodic energy in the F0-vocoded condition, mak-



**Figure 2.2:** Proportion correct scores in experiment 1 plotted as a function of the number of frequency bands for the three different vocoding conditions: Noise-vocoding (Nx, aperiodic source), Dudley-vocoding (FxFx, mixed source), and F0-vocoding (Fx, periodic source). The score for unprocessed speech is included for the purpose of comparison. The error bars show the standard error of the mean.

ing the information transmitted contradictory. In addition, listeners are confronted with ‘false’ intonation contours due to the interpolation of the natural F0-contours, which is likely to have lowered intelligibility rates even further.

Taken together, the results of experiment 1 show that in quiet testing conditions listeners did not benefit from natural periodicity information, while additional unnatural periodicity cues lead to substantially poorer speech intelligibility rates.

## 2.5 Experiment 2. Short introduction and rationale

Experiment 2 presented the three classes of target speech described in experiment 1 in a variety of background noises. The maskers used were either aperiodic speech-shaped noises or fully periodic harmonic complexes with a dynamically varying F0 contour (similar to those used in [Green and Rosen, 2013](#)). Both types of maskers were presented in a steady or 10 Hz sinusoidally amplitude-modulated version. This design allowed for a systematic variation of periodicity in both target and masker, and also allowed the examination of the role of amplitude fluctuations in the masker.

Performance was assessed via an estimation of the speech reception threshold (SRT, [Plomp and Mimpen, 1979](#)). Importantly, recent studies have emphasised that the difference in SRTs between conditions with steady and amplitude-modulated maskers (i.e. the fluctuating-masker benefit, FMB) is highly dependent on the signal to noise ratios (SNRs) at which they are measured. As [Bernstein and Grant \(2009\)](#) show, there is a strong negative relationship between the SNR found in a steady noise background and the FMB, both for normal-hearing and hearing-impaired listeners. To control for this confound, [Bernstein and Brungart \(2011\)](#) introduced a technique that adjusts the word-set size in each experimental condition in order to equate the performance levels in steady noise. However, an equalization procedure that is based on similar performance levels in steady noise would itself be biased by a possible effect of periodicity in the target speech. Since it appears likely that, for instance, the absence of any periodicity cues makes it particularly difficult to segregate noise-vocoded speech from a steady noise masker, we took a different approach and used the results obtained in experiment 1 to adjust for the different performance levels in quiet.

This approach is based on the assumption that varying the spectral resolution of the target speech in the presence of a masker has no other effect than to determine its intelligibility. While a degraded spectrum is likely to interfere with the segregation of target and masker when both signals are processed together, as is the case in CI simulations, the spectrum of the maskers in experiments 2 and 3 was always intact. As demonstrated by [Apoux et al. \(2015\)](#), it is differences in TFS *per se* that appear to be crucial for segregating target and masker. Thus, the critical point in the current experiments is that two separate carriers were present throughout. Nevertheless, it should be noted that degrading the spectrum of the target speech with a vocoder also introduces changes to the modulation spectrum, such as a greater similarity of the individual channel envelopes with fewer channels in the vocoder.

Conditions which were found to have very similar intelligibility rates in quiet were: Nx7, FxNx7, and Fx12, as well as Nx12, FxNx10, and Fx24 (see Table 2.1).

**Table 2.1:** Target speech conditions in experiment 2. Two sets of three processing conditions with similar percentage correct scores were chosen. The numbers following the abbreviation of the processing conditions indicate the number of frequency bands.

Processing condition	Nx7	FxNx7	Fx12	Nx12	FxNx10	Fx24
Percentage correct score	75.0	78.4	76.1	93.5	92.5	91.2

These 6 target conditions were combined with the 4 different maskers, adding up to 24 conditions. Note that the Fx24 condition was not part of experiment 1, but included in the current one. For convenience, results are presented together with those of experiment 3 that had a similar design but in which the intelligibility of the target speech in quiet was at ceiling.

## 2.6 Experiment 2. Methods

### 2.6.1 Participants

Twelve normal-hearing listeners (five females) were tested. Their ages ranged from 18 to 45 years with a mean age of 25.9. All participants were native speakers of British English, had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz, and did not participate in experiment 1.

### 2.6.2 Stimuli

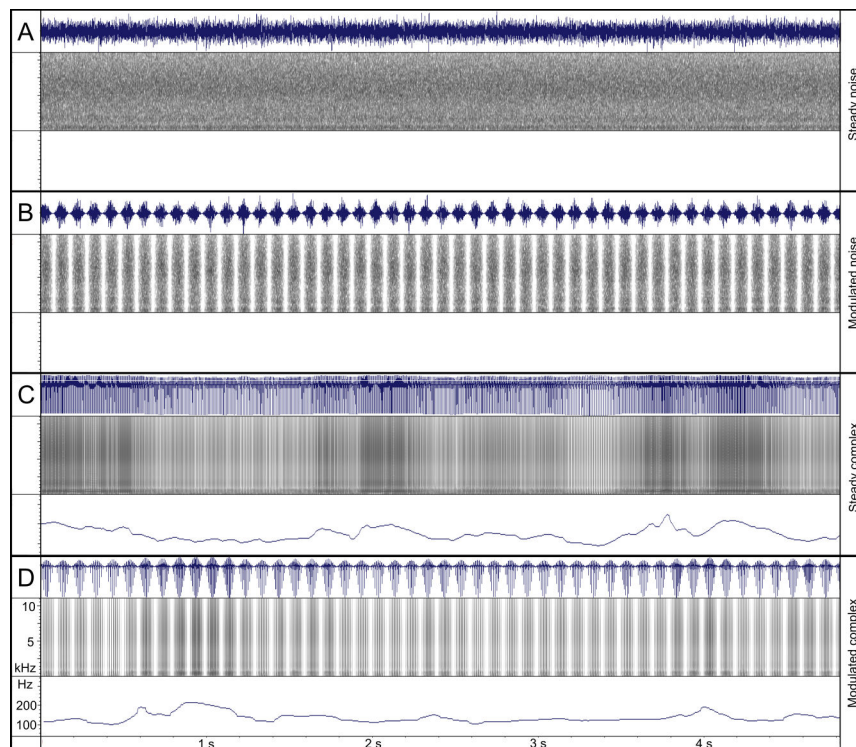
The target materials used in experiment 2 were the same recordings of the IEEE sentence corpus as in experiment 1. The harmonic complex maskers were based on F0 contours extracted from recordings in the EUROM database of English speech in which different speakers read five- to six-sentence passages ([Chan et al., 1995](#)). Sixteen different male talkers with Southern British English accents, and a similar speaking rate and voice quality to that of the target talker were chosen. The median F0 frequency of these 16 passages was 122.9 Hz and the first and third quartiles ranged from 107.0 Hz to 144.1 Hz. The median F0 of the IEEE target sentences was 117.2 Hz with the first and third quartiles ranging from 103.4 Hz to 136.1 Hz. Thus, the median F0 frequency of the target sentences was about 6% lower, but due to the large interquartile range of the F0 contours of both masker complexes and target speech, frequent F0 contour crossings are guaranteed.

Both the noise and harmonic complex maskers were presented either in a steady-state version or were sinusoidally amplitude-modulated at a rate of 10 Hz with a modulation depth of 100%. For each trial of the experiment, a random portion of the noise or complex maskers was picked and presented along with the target sentence. For the harmonic complex maskers, the order of the talkers on which the contour was based was also randomised so that all 16 were used before any of them was repeated. The onset of all the maskers was 600 ms before that of the targets and they continued for another 100 ms after the end of the target sentence. An onset and offset ramp of 100 ms was applied to the mixture of target and masker. Waveforms, wide-band spectrograms, and F0 contours of an example of all four maskers are shown in Fig. 2.3.

### 2.6.3 Signal processing

All target stimulus materials were again processed prior to the experiment. The same channel vocoder software as described in the first experiment was used to create the six target speech conditions. The noise maskers were based on a 24-second passage of white noise that was filtered (FIR filter, Greenwood filter spacing, 1-octave smoothing, filter order 1024, fft window size of 512 samples) to have the same long-term average speech spectrum (LTASS) as the target speech. The LTASS of the unprocessed target speech was determined by computing the power spectral density of the concatenated waveforms using Welch's method (window size 512 samples, 50% overlap, dft length 512 samples). The resulting spectrum was smoothed by 1 octave. F0 contours for the harmonic complex maskers were created by interpolating through unvoiced and silent periods using a piecewise cubic Hermite interpolation in logarithmic frequency. The waveforms were synthesised on a period-by-period basis using the Liljencrants-Fant model (Fant et al., 1985), which closely approximates a typical adult male glottal pulse [see Green and Rosen (2013) for details], and matched in spectrum to the long-term average of the target using the same filtering procedure as for the noise maskers.





**Figure 2.3:** Waveforms, wide-band spectrograms, and F0 contours of examples of the four maskers used in experiments 2 and 3. A) an aperiodic steady-state speech-shaped noise, B) an aperiodic speech-shaped noise with a 10 Hz sinusoidal amplitude modulation, C) a periodic steady-state harmonic complex with a dynamically varying F0 contour, and D) a periodic harmonic complex with a dynamically varying F0 contour and a 10 Hz sinusoidal amplitude modulation.

#### 2.6.4 Procedure

The experimental setting and general procedure were the same as in experiment 1. The current experiment consisted of 24 processing conditions presented in background noise (3 vocoding conditions x 2 degrees of spectral resolution x 4 maskers) and 1 additional condition presented in quiet (F0-vocoded speech with 24 bands, Fx24). Each condition consisted of 20 sentences, adding up to 500 trials in total. Participants were familiarised with the materials by listening to five sentences of each of the six target speech conditions and two additional example sentences in each of the four background noises.

The SRT for every processing condition was computed by tracking the SNR necessary in order to repeat 50% of the key words in a sentence correctly. The initial

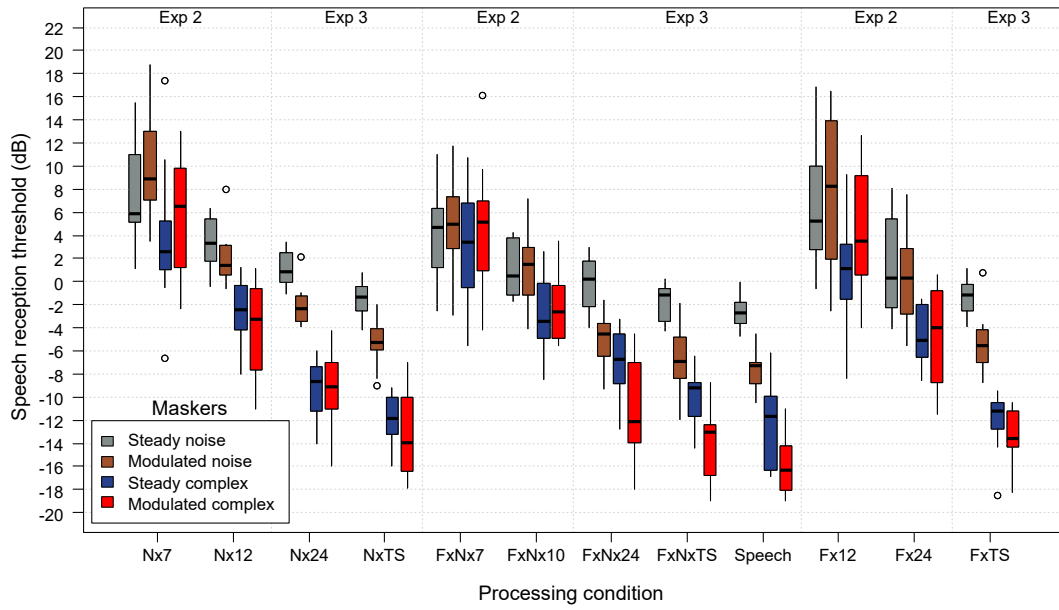
SNR was set to +10 dB and adjusted up or down by 11 dB before the first reversal, 7 dB before the second reversal, and 3 dB after that. If the subject got less than half of the key words correct in the first sentence, the SNR was set to +24 dB and the procedure started over again. The SRT was calculated by taking the mean of the largest even number of reversals with 3-dB step size. Throughout the experiment the level of the target and masker together was fixed at about 80 dB SPL over a frequency range of 75 Hz to 10 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

Psychometric functions were obtained by fitting a single logistic function to the averaged responses of all listeners for each combination of target and masker following the procedure described by [Wichmann and Hill \(2001\)](#). While intercept and slope were estimated without any restrictions, the lapse rate (which sets an upper limit to the performance) was estimated with the constraint to be the same within the set of target speech conditions with a lower intelligibility (Nx7, FxNx7, and Fx12), as well as that with a higher intelligibility (Nx12, FxNx10, and Fx24). The guessing rate was set to zero throughout, since the low semantic predictability and high complexity of the open-set IEEE sentences precludes successful guessing.

## 2.7 Experiment 2. Results and discussion

Figure 2.4 shows the SRTs obtained in experiment 2, together with those of experiment 3. For the three target speech conditions with lower intelligibility (Nx7, FxNx7, and Fx12) SRTs on a group level were positive throughout. The targets with higher intelligibility (Nx12, FxNx10, and Fx24) led to substantially lower SRTs and there was a trend for lower SRTs with more periodicity in the targets.

The data were analysed using a mixed effects model with target intelligibility, target periodicity, masker fluctuations, and masker periodicity as fixed factors, and subjects as a random factor. The main effects of target intelligibility [ $F(1,266) = 275.2, p < 0.001$ ] and masker periodicity [ $F(1,266) = 110.4, p < 0.001$ ] were highly significant. The main effect of target periodicity [ $F(2,264) = 3.1, p = 0.047$ ] was just significant, but there was no significant main effect of masker fluctuations [ $F(1,264) = 3.0, p = 0.09$ ]. Furthermore, the interactions of target intel-

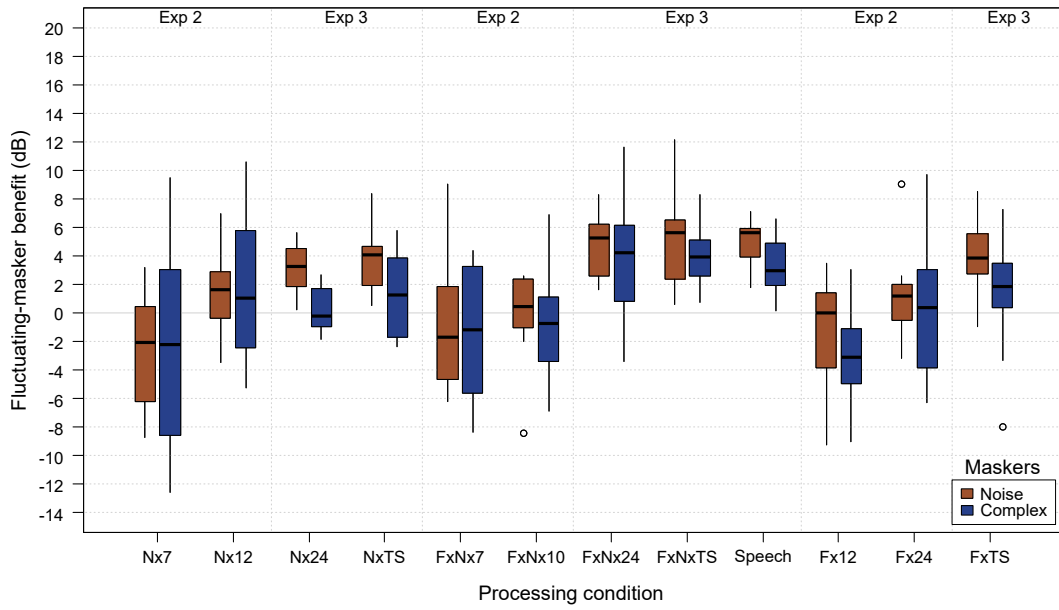


**Figure 2.4:** Boxplots of the speech reception thresholds (SRTs) obtained in experiments 2 and 3. Each of the twelve target speech conditions on the  $x$ -axis was tested in combination with the four different maskers shown in the legend. Nx stands for noise-vocoding, FxNx for Dudley-vocoding, and Fx for F0-vocoding. The numbers affixed to the processing conditions indicate the number of frequency bands in the vocoder. Conditions with the appendix ‘TS’ were produced using TANDEM-STRAIGHT and ‘Speech’ stands for unprocessed speech. The black horizontal lines in the boxplots indicate the median value.

ligibility and masker fluctuations [ $F(1,266) = 11.1, p = 0.001$ ], target intelligibility and masker periodicity [ $F(2,266) = 8.2, p < 0.01$ ], and target periodicity and masker periodicity [ $F(2,266) = 6.0, p < 0.01$ ] were significant.

As can be seen in Fig. 2.4, the SRTs for the four maskers in the FxNx7 condition are closer together than in the other target speech conditions. *Post hoc* pairwise comparisons using Bonferroni-corrected  $t$ -tests confirmed this observation and showed no significant differences between these four conditions, indicating that neither masker fluctuations nor masker periodicity substantially affected the SRTs in this condition. This result is likely to be one of the main reasons for the significant interactions of target intelligibility and masker periodicity as well as target periodicity and masker periodicity.

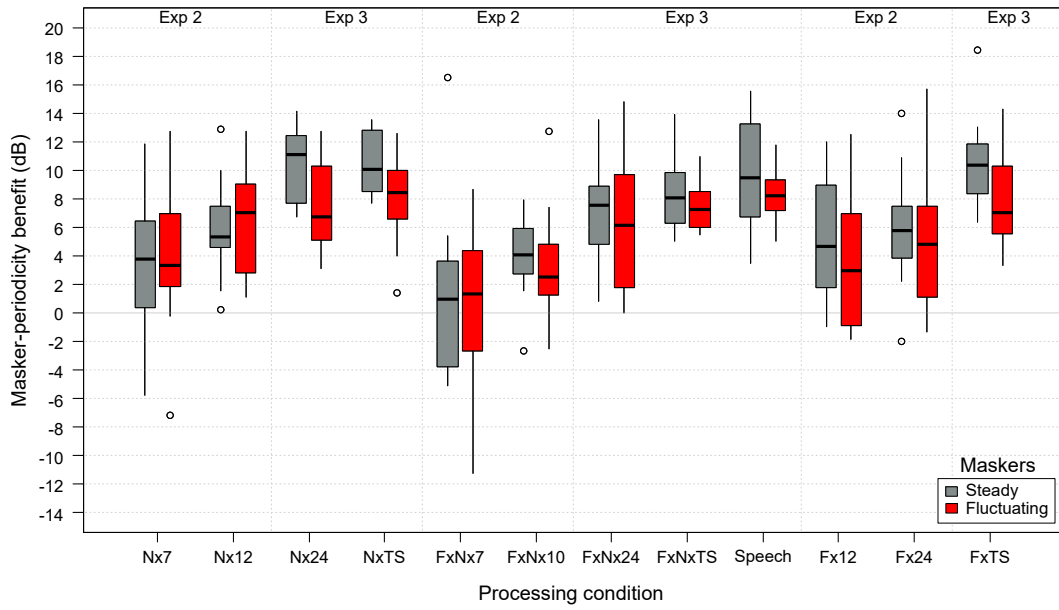
In order to enable a more fine-grained examination of the effects of amplitude fluctuations in the masker, Fig. 2.5 plots the FMB, which is the difference in SRT



**Figure 2.5:** Boxplots of the fluctuating-masker benefits (FMBs) obtained in experiments 2 and 3. For each of the twelve target speech conditions on the  $x$ -axis, the difference between the steady and amplitude-modulated version of the noise and harmonic complex maskers is plotted. Positive numbers on the  $y$ -axis indicate a benefit. Target speech conditions are the same as in Fig. 2.4. The black horizontal lines in the boxplots indicate the median value.

of a steady compared to a fluctuating masker for each target and masker type. The FMBs of experiment 2 are again plotted together with those of experiment 3. Positive FMBs indicate that listeners were able to benefit from masker fluctuations. *Post hoc t*-tests showed that there were no significant differences between the steady and amplitude-modulated versions of the noise and complex maskers in any of the six target speech conditions. It can, however, be seen in Fig. 2.5 that there is a trend for more FMB with the more intelligible targets. While we observed a small but consistent fluctuating-masker *interference* of up to 3 dB for the targets with lower intelligibility (Nx7, FxNx7, and Fx12), this effect disappears when the intelligibility of the targets is higher (Nx12, FxNx10, and Fx24), which also explains the significant interaction of target intelligibility and masker fluctuations.

Figure 2.6 plots the difference between aperiodic and periodic maskers, termed the *masker-periodicity benefit* (MPB), in experiments 2 and 3. In stark contrast to the FMB, subjects did benefit from periodicity in the masker across all target speech



**Figure 2.6:** Boxplots of the masker-periodicity benefits (MPBs) obtained in experiments 2 and 3. For each of the twelve target speech conditions on the  $x$ -axis, the difference between the noise and harmonic complex version of the steady and amplitude-modulated maskers is plotted. Positive numbers on the  $y$ -axis indicate a benefit. Target speech conditions are the same as in Fig. 2.4. The black horizontal lines in the boxplots indicate the median value.

conditions, with effects of up to about 7 dB. As for the FMB, the MPB increased with the intelligibility of the targets, explaining the significant interaction of target intelligibility and masker periodicity.

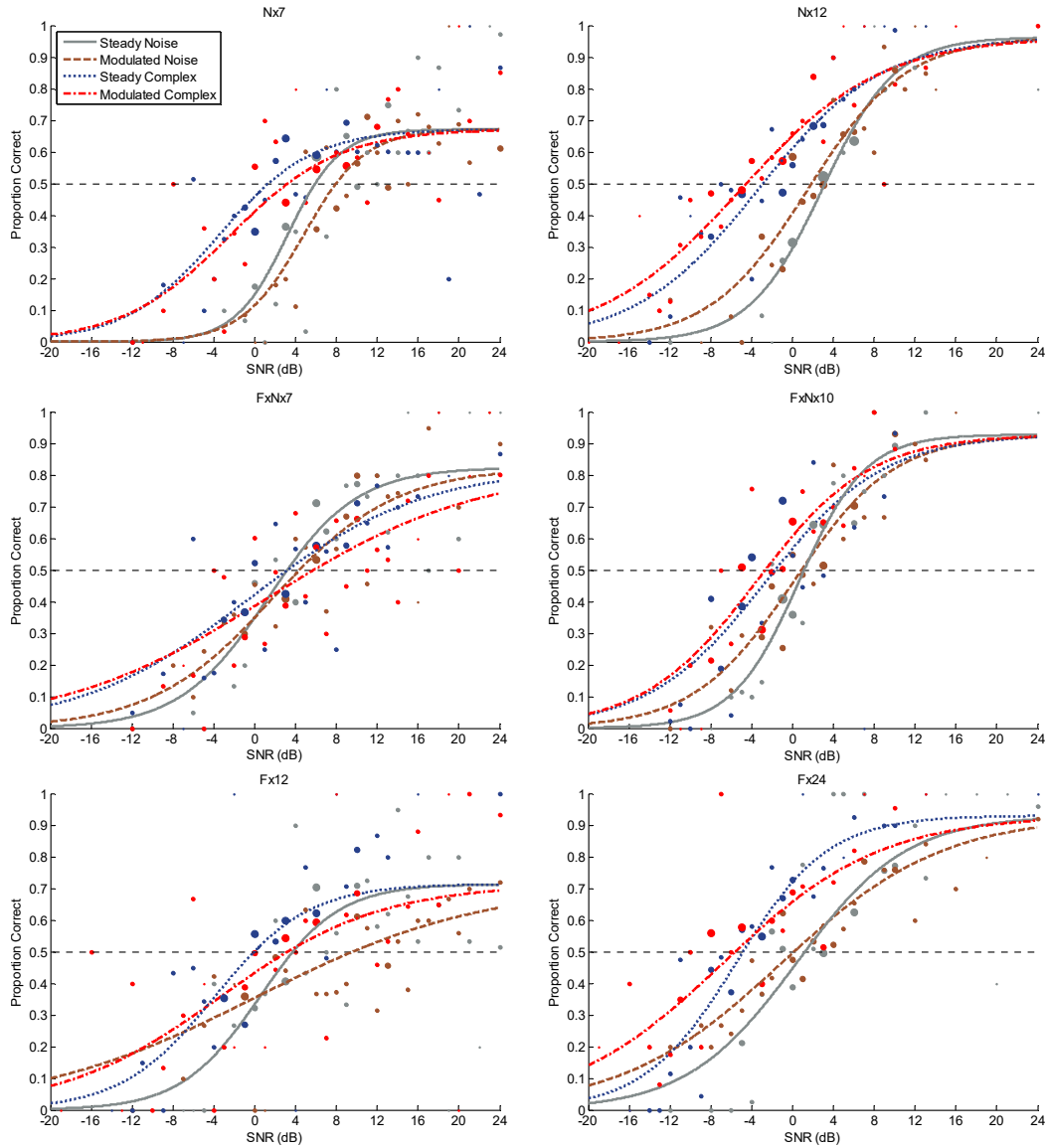
As the SRT results show, performance with the FxNx targets was least affected by the differences between the four maskers. This observation is also evident in the pattern of the MPB results, where the smallest benefits were found with the FxNx targets. *Post hoc t*-tests comparing the periodic and aperiodic maskers in all 6 target speech conditions showed that only in the FxNx7 condition was there no significant difference between these, no matter if they were steady [ $t(11) = 0.16$ ,  $p = 0.88$ ] or fluctuating [ $t(11) = 0.60$ ,  $p = 0.56$ ].

The FMB is known to be strongly influenced by the SNR at which a test is carried out (Freyman et al., 2012; Smits and Festen, 2013). Our results suggest that the same is true for the MPB, but the exact relation of the factors involved is difficult to grasp from the snapshot-like SRT data. In order to obtain a broader

picture of the results we fitted psychometric functions (PFs) to the pooled data of each of the 24 target-masker combinations (Fig. 2.7). On average, the measured SRTs and the estimated 50%-correct values extracted from the PFs were about 0.9 dB apart, indicating a reasonably good fit. As reported previously (Freeman et al., 2012; Smits and Festen, 2013) we found that steady maskers generally led to steeper slopes, as indicated by a significant  $t$ -test comparing the slopes of all conditions with steady maskers to all conditions with modulated maskers [ $t(11) = 4.8, p < 0.001$ ]. A significant  $t$ -test also showed that slopes were steeper for noise maskers when compared to harmonic complex maskers [ $t(11) = 3.3, p < 0.01$ ].

These data are also consistent with the idea that the size of the FMB depends on the SNR, with glimpsing observed almost exclusively at negative SNRs. This effect is particularly strong for the two Fx conditions where the slopes of the functions for steady and fluctuating maskers differ a lot, resulting in large fluctuating-masker interference at positive SNRs and similarly large FMBs at negative SNRs. Increasing the intelligibility of the target speech independently enhanced the likelihood of glimpsing, but only the combination with a negative SNR proved to be both necessary and sufficient to enable some degree of FMB.

Importantly, PFs were found to show three distinct patterns depending on the amount of periodicity in the target speech. These patterns are observable for the targets with lower as well as those with higher intelligibility, pointing to common underlying mechanisms involving aspects of periodicity. In both the Nx7 and Nx12 conditions, for example, the functions for steady and modulated maskers are aligned fairly closely, while the distance between the noise and harmonic complex maskers is much larger, confirming the finding that the MPB is greater than the FMB. Similarly the close alignment of the boxplots in the FxNx conditions is reflected in the shapes of the respective PFs, which remain relatively close together across the whole range of SNRs. Finally, in the Fx conditions, as already mentioned, the effect of masker fluctuations, but not masker periodicity, depended heavily on the SNR.



**Figure 2.7:** Psychometric functions fitted to the aggregated results of each of the 24 processing conditions (6 targets x 4 maskers) in experiment 2. The target speech condition is indicated above each of the six panels, and labels are the same as in Fig. 2.4. The horizontal line in each panel indicates the 50%-level that was tracked in the adaptive SRT procedure. The size of the points corresponds to the number of trials at a particular SNR.

Another observation worth mentioning is that the upper performance limits (i.e. the lapse rates) of the targets with lower intelligibility (Nx7, FxNx7, and Fx12) differ considerably, with the FxNx7 condition leading to much better performance rates at higher SNRs. As masker levels were very low at these SNRs, the unnatural acoustic properties of the Nx7 and Fx12 targets would have been quite evident. Since the listeners were only presented with a few example sentences before the main experiment, their unfamiliarity with these materials may have affected performance.

## 2.8 Experiment 3. Short introduction and rationale

A key finding of experiment 2 was that, on average, listeners always benefitted from periodicity in the masker, but not from masker fluctuations, even when the intelligibility of the target speech was as high as about 90% in quiet. Additionally, there was a clear trend for more MPB and FMB (or less fluctuating-masker interference) when the intelligibility of the targets was higher and the resulting SRTs lower. In order to further investigate this relation, we kept the general design of experiment 2, but used target speech with intelligibility rates approaching ceiling level in order to enable testing at lower SRTs.

An initial obstacle of experiment 3 was that the band-vocoder software used in experiments 1 and 2 cannot be employed to produce noise-vocoded stimuli with a very high number of bands. With more than 24 bands the individual harmonics begin to be resolved, which leads to a clear percept of the F0 and an overall less noise-like sound quality, thereby undermining the idea central to noise-vocoding.

An alternative vocoder that does not filter the input speech into separate frequency bands but instead separates the periodic and aperiodic components of the source from the spectral filter is TANDEM-STRAIGHT ([Kawahara et al., 2008](#)). By default TANDEM-STRAIGHT produces very natural-sounding speech with a mixed source excitation, but the source estimation procedure can be adapted to produce fully aperiodic or fully periodic speech as well.

Apart from 24-band noise- and F0-vocoded speech (Nx24, Fx24), experiment 3 thus also included noise-vocoded, Dudley-vocoded, and F0-vocoded speech pro-



duced with TANDEM-STRAIGHT (henceforth referred to as NxTS, FxNxTS, and FxTS). Extending the idea of maximizing the spectral detail in the targets, we also used unprocessed speech (referred to as ‘Speech’). All six target speech conditions in experiment 3 should lead to near perfect intelligibility in quiet. As the results of experiment 1 show (see Fig. 2.2), the Nx16 and FxNx16 conditions already led to over 95% of correctly repeated key words. Adding another eight frequency bands was therefore hypothesised to raise the performance levels in quiet to those of unprocessed speech. The even higher spectral resolution of the stimuli produced with TANDEM-STRAIGHT is assumed to result in similarly high scores.

## 2.9 Experiment 3. Methods

### 2.9.1 Participants

Twelve normal-hearing listeners (seven females) were tested. Their ages ranged from 18 to 30 years with a mean of 22.3 years. All participants were native speakers of British English, had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz, and did not participate in experiments 1 or 2.

### 2.9.2 Stimuli

The target materials were the same recordings of the IEEE sentence corpus as in experiments 1 and 2, and the maskers were the same as in experiment 2.

### 2.9.3 Signal processing

For the Nx24 and FxNx24 conditions the same channel vocoder software as in experiments 1 and 2 was used. TANDEM-STRAIGHT was used to produce noise-vocoded speech (NxTS) by keeping the default settings, but fixing the F0 to 0 Hz throughout. In order to synthesise Dudley-vocoded speech with TANDEM-STRAIGHT (FxNxTS), the default settings were used, but the values of the sigmoid parameter in the source estimation routine were fixed to 1 and -40, in order to minimise the level of the aperiodic component. This avoids the possibility that higher harmonics are noisier than lower ones, as is the case in natural speech, and ensures comparability with the Dudley-vocoded speech produced with a channel vocoder. The same technique was used to produce F0-vocoded speech with

TANDEM-STRAIGHT (FxTS), but here the same interpolated F0 contours as for the channel vocoder were used as input for the source extraction routine. Additionally, the unprocessed IEEE recordings were used as a sixth target speech condition (Speech).

#### 2.9.4 Procedure

The experimental setting and procedure was generally the same as in experiment 2. Before being tested, the participants were familiarised with the materials by listening to five example sentences of each of the three target conditions with an unnatural source (Nx24, NxTS, and FxTS) in quiet, followed by two unprocessed example sentences combined with each of the four maskers at an SNR of 0 dB. For the analyses of the PFs, the lapse rate was set to 0.

### 2.10 Experiment 3. Results and discussion

The SRTs are shown in Fig. 2.4, along with the SRTs of experiment 2. As expected, unprocessed speech led to the lowest SRTs with all four maskers. Most importantly, the SRTs in experiment 3 show a stepwise descending pattern for each of the six target speech conditions, indicating that listeners benefitted from amplitude fluctuations in the masker, but even more so from periodicity in the masker.

The data were analysed using a mixed effects model with the fixed effects target condition, masker periodicity, and masker fluctuations, and subjects as a random factor. The main effects of target condition [ $F(5,264) = 26.6, p < 0.001$ ], masker periodicity [ $F(1,264) = 978.4, p < 0.001$ ], and masker fluctuations [ $F(1,264) = 144.4, p < 0.001$ ] were all highly significant. There were also significant interactions of target condition and masker periodicity [ $F(5,264) = 2.6, p < 0.05$ ], target condition and masker fluctuations [ $F(5,264) = 3.6, p < 0.01$ ], and masker periodicity and masker fluctuations [ $F(1,264) = 16.4, p < 0.001$ ].

The SRTs of the three conditions produced with TANDEM-STRAIGHT were almost as low as those of unprocessed speech as indicated by non-significant fixed coefficients [NxTS (1.1,  $p = 0.23$ ), FxNxTS (0.9,  $p = 0.34$ ), and FxTS (1.3,  $p = 0.16$ )]. The fixed coefficients of the 24-channel vocoded targets, on the other

hand, indicate that they led to significantly higher SRTs than unprocessed speech [Nx24 (3.7,  $p < 0.001$ ) and FxNx24 (2.4,  $p < 0.01$ )]. Furthermore, a separate mixed model that was similar to the previous one but included only the three TANDEM-STRAIGHT conditions showed no significant main effect of target condition [ $F(2,132) = 0.48$ ,  $p = 0.62$ ], indicating that target periodicity in these conditions did not affect the SRTs.

The FMBs of experiment 3 (Fig. 2.5) show that the largest benefits were obtained for target speech conditions with a natural mixed source (FxNx24, FxNxTS, and Speech). Additionally, the FMB was consistently found to be lower for harmonic complex maskers. These two findings are likely to have caused the significant interactions of target condition and masker periodicity as well as masker periodicity and masker fluctuations, respectively. Furthermore, *post hoc* Bonferroni-corrected *t*-tests showed that for the completely voiced or unvoiced target speech (Nx24, NxTS, and FxTS), the FMB for complex maskers was not significantly different from zero. Thus, only target speech with a natural mixed source seems to enable substantial glimpsing in the presence of harmonic complex maskers.

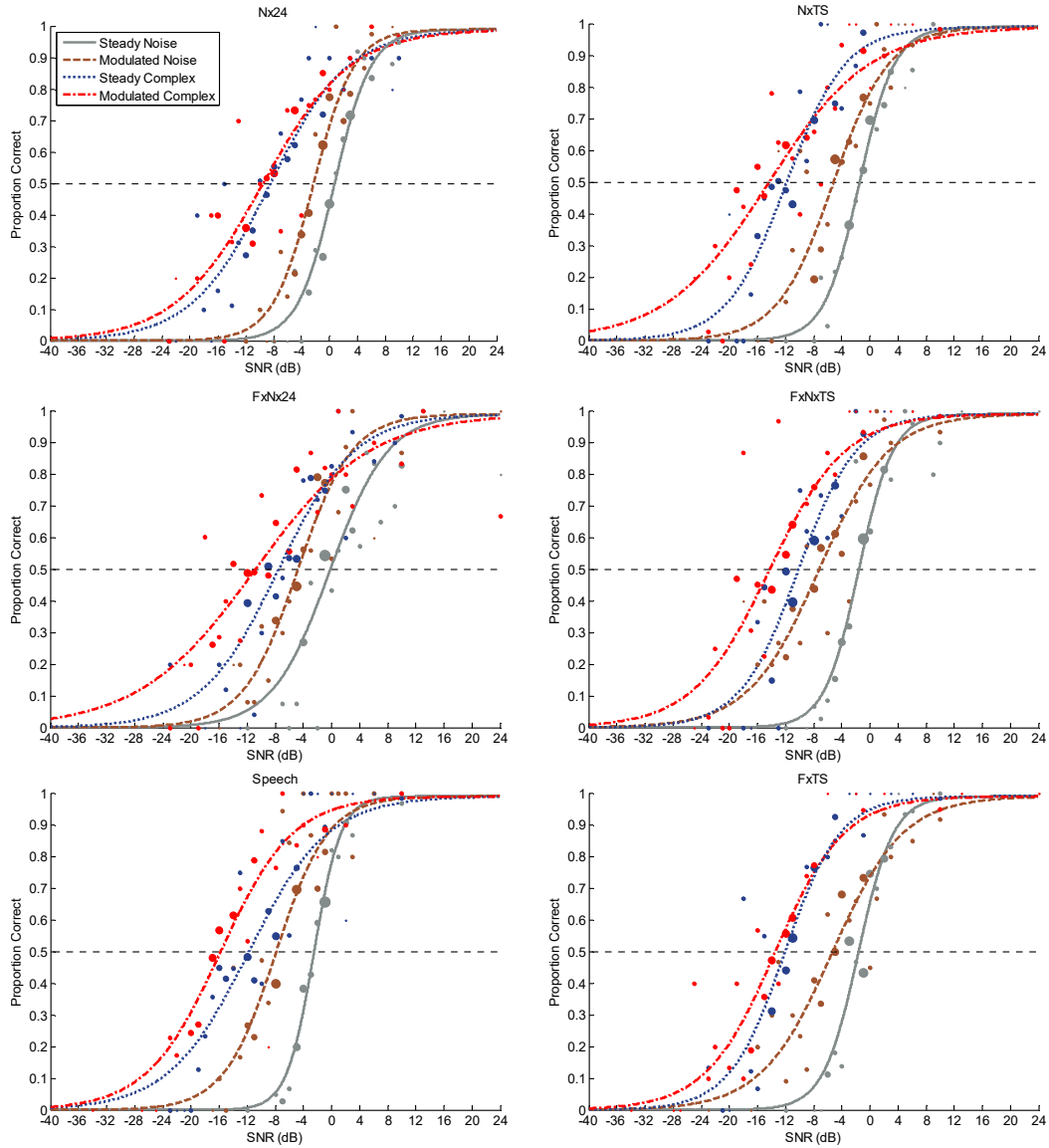
Figure 2.6 shows the masker-periodicity benefits (MPBs) obtained in experiment 3, added to those of experiment 2. Listeners again strongly benefited from periodicity in the masker across all six target speech conditions. Importantly, with a maximum of about 11 dB in the Nx24 condition, the MPB was almost twice as large as the maximum FMB (about 6 dB, see Fig. 2.5). The MPB was also consistently larger for steady maskers, which is another reason for the significant interaction of masker periodicity and masker fluctuations. Additionally, *post hoc t*-tests showed that for steady maskers, the FxNx24 condition showed significantly less MPB than the Nx24 condition [ $t(11) = 5.1$ ,  $p < 0.001$ ], and that the same was true for the FxNxTS condition when compared to NxTS [ $t(11) = 3.1$ ,  $p < 0.05$ ] and FxTS [ $t(11) = 2.6$ ,  $p < 0.05$ ]. When the masker was steady, targets with a natural mixed source thus led to smaller MPBs than aperiodic or periodic target speech. This result also explains the significant interaction of target condition and masker periodicity.

As in experiment 2, we again fitted psychometric functions to the pooled data of each of the 24 target-masker combinations (see Fig. 2.8). The measured SRTs and the estimated 50%-correct values extracted from the PFs were this time about 0.25 dB apart, indicating a good fit. *T*-tests again showed that steady maskers had steeper slopes than modulated maskers [ $t(11) = 3.5$ ,  $p < 0.01$ ] and that noise maskers had steeper slopes than harmonic complex maskers [ $t(11) = 5.0$ ,  $p < 0.001$ ]. The PFs in the current experiment are mostly located in the negative SNR region, but it is again evident that FMB and MPB diminish, or in the case of the FMB even turn into an interference effect, once they approach positive SNRs. Additionally, the three target conditions with a mixed source (FxNx24, FxNxTS, and Speech) all show a more even spacing of the PFs across the four maskers. The latter observation corresponds well with the FMBs of experiment 3 (Fig. 2.5), which show that only targets with a mixed source enabled the listeners to substantially benefit from fluctuations in both the noise and the harmonic complex maskers.

## 2.11 General discussion

### 2.11.1 Target periodicity in background noise

Generally speaking, the amount of periodicity in the target speech affected the SRTs in experiments 2 and 3 relatively little. The main effect of target periodicity was just significant in experiment 2, but the direct comparison of three conditions produced with TANDEM-STRAIGHT in experiment 3 revealed no effect of target periodicity. This is somewhat surprising, since one might expect that, for instance, the combination of an aperiodic target with an aperiodic masker would be particularly difficult due to a lack of cues that aid stream segregation. Yet, as the SRTs in Fig. 2.4 show, performance with the fully voiced Fx targets was in no case more than about 2 dB better than with the aperiodic Nx targets for the two aperiodic noise maskers. The patterns of the psychometric functions for the Nx and Fx targets in experiment 2 (see Fig. 2.7) in particular, however, reveal that while at the 50%-correct level differences between these target conditions are relatively small, the performance with the Nx targets at lower SNRs is indeed much poorer when the masker is aperiodic.



**Figure 2.8:** Psychometric functions fitted to the aggregated results of each of the 24 processing conditions (6 targets x 4 maskers) in experiment 3. The target speech condition is indicated above each of the six panels, and the labels are the same as in Fig. 2.4. The horizontal line in each panel indicates the 50%-level that was tracked in the adaptive SRT procedure. The size of the points corresponds to the number of trials at a particular SNR.

The shapes of the psychometric functions thus confirm that periodicity is important in segregating competing auditory streams, making it clear that SRTs alone are not sufficient in obtaining a complete picture of the patterns in the data. In contrast, this issue does not arise when evaluating the performance in the FxNx conditions. Here the results vary much less between the different maskers across SNRs, suggesting that speech with a natural mix of periodicity and aperiodicity leads to a much more robust percept.

### 2.11.2 Masker fluctuations

The effect of masker fluctuations was found to strongly depend on the intelligibility of the target speech, with interference effects of about 2 dB observed in experiment 2 and maximum benefit of almost 6 dB in experiment 3. This trend is in line with previous studies reporting a strongly reduced ability to glimpse for hearing-impaired listeners and CI users. A recent attempt to model SRTs in fluctuating noise by [Smits and Festen \(2013\)](#) also supports these results by predicting reduced or even negative FMBs at very high SNRs.

Based on the findings of Stone and colleagues ([Stone et al., 2011](#); [Stone et al., 2012](#)) this trend could also be explained with reference to the concept of modulation masking. While the 10 Hz sinusoidal amplitude-modulations of the maskers potentially enabled the glimpsing of sections of target speech, they also introduced additional amplitude modulations to the masker envelope that could interfere with informative modulations in the targets. The benefits of glimpsing seem to outweigh the modulation masking at lower SNRs, but not at higher SNRs, where the target speech is already audible when the masker is steady.

The psychometric functions of experiment 2, however, again show that examining the results only through SRTs can be deceptive. While the small effects of masker fluctuations in the Nx and especially the FxNx conditions are fairly stable across different SNRs, much larger and more variable effects were found in the Fx condition. Here masker fluctuations led to considerable benefits at low SNRs, but also particularly large interference effects at high SNRs.

A less well-established result of the current study is that, apart from the targets with lower intelligibility in experiment 2, there appears to be more glimpsing when the masker is aperiodic. This difference is particularly pronounced for the Nx and Fx targets, and might be due to the fact that complex maskers are inherently more coherent and thus easier to segregate from the target speech, no matter if steady or fluctuating.

The largest FMBs of about 6 dB have been found for target speech with a mixed source and a high intelligibility (FxBx24, FxBxTS, and Speech). In conjunction with the small differences in FMB between the noise and complex maskers for these targets, this suggests that a natural mix of periodicity and aperiodicity in the target speech aids glimpsing. Although the maximum FMBs obtained with the Nx targets are only about 2 dB smaller, this finding hence does support the notion that TFS information in the target speech is important in order to benefit from masker fluctuations (Gnansia et al., 2009; Lorenzi et al., 2006).

### 2.11.3 Masker periodicity

The large and consistent masker-periodicity benefits of up to about 11 dB (see Fig. 2.6) suggest that periodicity in the masker is even more important than masker fluctuations in attempting to segregate target speech from background noise. This finding is in close agreement with the harmonic cancellation theory (de Cheveigné et al., 1995; de Cheveigné et al., 1997b) which states that harmonicity in the masker enables the auditory system to effectively subtract the masking sound from the signal mixture.

There is, however, an additional explanation of the masker-periodicity benefit that does not rely on harmonicity but instead the glimpsing opportunities that arise in between the individual harmonics of the complex maskers. A recent study by Derroche et al. (2014b) refers to this mechanism as ‘spectral glimpsing’ and provides evidence that spectral glimpsing and harmonic cancellation contribute independently in explaining the masker-periodicity benefit. First, they showed that due to the increasing size of spectral dips, both harmonic and inharmonic complexes were less effective in masking the target speech as their F0 frequencies increased. In

addition, they report that even after controlling for the generally greater spectral glimpsing opportunities in inharmonic maskers, the harmonic complexes still led to consistently lower SRTs and that this effect is independent of the F0 frequencies of the complexes.

Another factor explaining the reduced effectiveness of periodic maskers is that, apart from fluctuations at the rate of the F0, the envelopes of harmonic complexes with a stationary F0 hardly fluctuate, particularly not at the low modulation rates essential for speech intelligibility (Deroche et al., 2014b). As Stone and colleagues (Stone et al., 2011; Stone et al., 2012) have shown, envelope fluctuations, rather than envelope energy, are the primary reason for the effectiveness of aperiodic noise maskers. Contrary to the maskers used by Deroche et al. (2014b), the harmonic complexes in the current study had varying F0-contours in order to make them more speech-like and thus more ecologically valid. These changes in F0, however, also introduce additional slow modulations to the envelopes of the lower auditory filters and it remains to be determined whether this has a substantial effect on performance.

The pattern in the SRTs as well as the psychometric functions show that the MPB is smallest for targets with a mixed source (F<sub>x</sub>N<sub>x</sub>). One possible explanation for this could be that the gaps in the F0 contours of these targets made it slightly more difficult to form two separate auditory streams. For the aperiodic and periodic targets in contrast this is likely to be easier since in the former case the harmonic background can be cancelled out (de Cheveigné, 1998), while in the latter case, two F0 contours are present throughout. Furthermore, the MPB tended to be larger for steady than for fluctuating maskers, which seems intuitive given the fact that in fluctuating maskers there are sections with little or no masker energy, while for steady maskers energy is present throughout.

Crucially, the harmonic complex maskers used in the current study were not only meant to provide a periodic counterpart to the more commonly used aperiodic noise maskers, but also designed in an attempt to better match the acoustic characteristics of speech. Connected stress-timed speech, such as English, is voiced about



50% of the time, while unvoiced sections and pauses only amount to about 25% each (Dellwo et al., 2007; Fourcin, 2010). A harmonic complex masker is thus *per se* more speech-like than an aperiodic noise masker.

As mentioned before, the F0s of the IEEE targets and complex maskers differed by about a semitone. It has been shown that even F0 differences of this order can help to tell apart signal and noise, but these findings are restricted to artificial stationary vowels (Culling and Darwin, 1993; de Cheveigné et al., 1997a). As described by Darwin (2008), natural speech is too variable for such small differences in F0 to matter much. The mechanism for segregating stationary vowels with similar F0 frequencies relies on beats caused by the close spacing of the harmonics, which oscillate at relatively slow rates. Studies using real speech as targets have consequently reported hardly any benefit for F0 differences of about one semitone and gradual changes as the difference was increased (Bird and Darwin, 1998; Brokx and Nootboom, 1982).

## 2.12 Summary and conclusion

The present study found that in quiet testing conditions, aperiodic noise-vocoded speech and vocoded speech with a natural amount of source periodicity were equally intelligible, while fully periodic vocoded speech with an interpolated F0 contour is much harder to understand. In the presence of a masker, periodicity in the target speech had a surprisingly small effect. Performance was slightly better with more target periodicity, but only when SRTs were relatively high. Periodicity in the masker, on the other hand, was found to strongly aid speech intelligibility, and this effect was much larger than the FMBs observed. Generally, the higher the intelligibility of the target speech in quiet, the larger were the observed MPBs and FMBs, and a substantial FMB, in particular, required the target speech intelligibility in quiet to be close to ceiling.

In summary, our results show that periodicity in the masker, but surprisingly not the target speech, is an important factor in tracking a speech signal through a background noise. Factors that are thought to underlie the masker-periodicity bene-

fit include the presence of discrete spectral components, the relatively sparse modulation spectrum, and the harmonic relation of the individual components. Further research is needed to identify the respective contributions of these factors.

## Chapter 3

### Effects of acoustic periodicity, intelligibility, and pre-stimulus alpha power on the event-related potentials in response to speech<sup>1</sup>

#### 3.1 Introduction

Acoustically degraded noise-vocoded speech has been used extensively to investigate the neural correlates of speech intelligibility in both magneto- and electroencephalographic (M/EEG) studies (e.g. [Becker et al., 2013](#); [Ding, et al., 2014](#); [Obleser and Weisz, 2012](#); [Peelle et al., 2013](#)) and imaging work (e.g. [Davis and Johnsrude, 2003](#); [Evans et al., 2014](#); [Scott et al., 2000](#)). Noise-vocoding has proven a very useful tool because it allows the parametric reduction of the intelligibility of speech signals by reducing the number of channels in the analysis/synthesis process. However, this signal manipulation alters the acoustic properties of the stimuli as well as their intelligibility, and these two factors have so far not been considered independently.

Furthermore, while the reduction in intelligibility can mainly be attributed to the lowered spectral resolution of the vocoded speech signals, other acoustic properties are affected by the signal processing as well. Most notably, due to the use of a broadband noise as sound source, noise-vocoded speech is completely aperiodic (i.e. unvoiced), making it sound like an intense version of a whisper. In natural speech, on the other hand, voiced and unvoiced segments alternate. Importantly, only voiced speech possesses a pitch. Previous studies that have investigated pitch perception reliably found increased neural responses for stimuli that possess

---

<sup>1</sup>This chapter has been published as: Steinmetzger, K. and Rosen, S. (2017). Effects of acoustic periodicity, intelligibility, and pre-stimulus alpha power on the event-related potentials in response to speech. *Brain and Language* 164, 1–8.

a pitch, when compared to a spectrally matched control condition (e.g. [Griffiths et al., 2010](#); [Norman-Haignere et al., 2013](#)) or a broadband noise ([Chait et al., 2006](#)). In particular, studies analysing MEG signals in the time domain ([Chait et al., 2006](#); [Gutschalk et al., 2004](#)) have shown that following a transient pitch onset response peaking after around 150 ms, a sustained neural response can be observed for several hundred milliseconds. Thus, it appears likely that the neural response elicited by noise-vocoded speech is per se attenuated due to the absence of voicing.

In order to address these issues, we have used a vocoding technique that allows the choice between a periodic (voiced) or an aperiodic (unvoiced) source excitation. This technique was used to synthesise speech that is either completely unvoiced (i.e. noise-vocoded, henceforth referred to as the *aperiodic* condition), preserves the natural mix of voiced and voicelessness (henceforth the *mixed* condition; [Dudley, 1939](#)), or is completely voiced (henceforth the *periodic* condition). Previous behavioural work ([Steinmetzger and Rosen, 2015](#); i.e. chapter 2) has shown that the intelligibility of the aperiodic and mixed conditions is very similar, while the unnatural-sounding fully periodic condition was found to be considerably less intelligible. In order to analyse effects of acoustic periodicity while controlling for differences in intelligibility, the individual trials in the current study were sorted according to the listeners' spoken responses (i.e. the number of correctly repeated key words) obtained after every sentence, and only fully intelligible trials were considered. In summary, the first hypothesis was that speech with more periodicity would lead to more negative event-related potentials (ERPs), reflecting the increased neural sensitivity to auditory input that possess a pitch. This effect was expected to be observed during an early time window following sentence onset, including the auditory evoked potentials (AEPs) and the acoustic change complex (ACC; [Pratt, 2011](#)).

Sorting the individual trials according to the behavioural responses was also intended to enable the separate analysis of more or less intelligible trials in the periodic condition. This second analysis additionally included spectrally-rotated speech, a completely unintelligible non-speech analogue that has been used in a

number of the previously mentioned studies (Becker et al., 2013; Peelle et al., 2013; Scott et al., 2000), as a baseline condition (henceforth the *rotated* condition). In contrast to several recent M/EEG studies that have investigated the perception of noise-vocoded (Becker et al., 2013; Obleser and Weisz, 2012; Obleser et al., 2012) and unprocessed speech (e.g. Kerlin et al., 2010; Müller and Weisz, 2012; Wilsch et al., 2015) by analysing neural activity in the frequency domain, the current study focusses on time-domain responses. Few studies to date have investigated ERPs in response to degraded speech (for exceptions see Becker et al., 2013; Obleser and Kotz, 2011; Wöstmann et al., 2015b) and it is hence not well understood how they are affected by both the acoustic characteristics and the intelligibility of the speech signals, particularly over the course of whole sentences.

Based on the notion that slow cortical potentials reflect the degree of cortical excitability (Birbaumer et al., 1990; He and Raichle, 2009), it was hypothesised that ERP amplitudes over the course of the sentences would be larger in response to more intelligible speech. More specifically, slow negative potentials with an anterior scalp distribution have been associated with both working memory load (e.g. Guimond et al., 2011; Lefebvre et al., 2013) and increased attention (e.g. Teder-Sälejärvi et al., 1999; Woods et al., 1994) in auditory tasks. A typical slow negative potential is the contingent negative variation (CNV), which emerges in between a warning stimulus and a task-relevant second stimulus, and is larger when subjects expect and prepare to respond to the latter stimulus (McCallum and Walter, 1968; Tecce and Scheff, 1969). Importantly, the second stimulus may also be a response to the first stimulus (Birbaumer et al., 1990; Kononowicz and Penney, 2016), and hence the design of the current experiment, in which subjects are supposed to verbally repeat the stimulus sentence, fits into the CNV framework too.

In order to further investigate differences between intelligible and unintelligible trials, we additionally analysed the amount of alpha power in the silent baseline interval preceding the stimulus sentences. Decreased alpha power in the pre-stimulus window has been shown to be a predictor of successful target identification in studies using low-level visual and somatosensory stimuli (e.g. Hanslmayr, et al.,

2007; Romei et al., 2010; Schubert et al., 2009; Van Dijk et al., 2008). Strauß et al. (2015) have recently also reported alpha phase differences before correctly and incorrectly perceived words in a lexical decision task, but no study to date has reported alpha power differences in the baseline window using speech materials presented auditorily. As reviewed by Klimesch (1999, see also Klimesch et al., 1998), slower alpha frequencies (~7–10 Hz) in particular have been associated with alertness and expectancy, and may thus serve as a measure of the attentional state in the period before sentence onset. We thus additionally hypothesised to observe enhanced slow alpha power, indicating that subjects have not been fully focussed on the upcoming task, before sentences that would turn out to be unintelligible to them.

## 3.2 Methods

### 3.2.1 Participants

Eighteen normal-hearing right-handed subjects (8 females, mean age = 21.6 years, SD = 2.3 years) took part in the study. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB Hearing Level at octave frequencies from 125 and 8000 Hz. All subjects gave written consent and the study was approved by the UCL ethics committee.

### 3.2.2 Stimuli

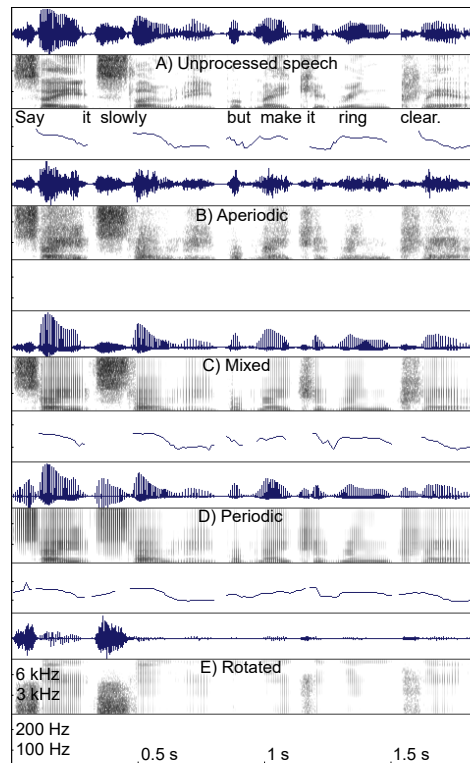
The stimulus materials used in this experiment were recordings of the IEEE sentences (Rothauser et al., 1969) spoken by an adult male Southern British English talker with a mean F0 of 121.5 Hz that were cut at zero-crossings right before sentence onset and normalised to a common root-mean-square (RMS) level. The IEEE sentence corpus consists of 72 lists with 10 sentences each and is characterized by similar phonetic content and difficulty across lists, as well as an overall low semantic predictability (e.g. *Say it slowly but make it ring clear.*). The individual lists are thus supposed to be equally intelligible. Every sentence contains five key words.

All stimulus materials were processed prior to the experiment using a channel vocoder implemented in MATLAB (Mathworks, Natick, MA). For all three vocod-

ing conditions (aperiodic, mixed, and periodic) the original recordings of the IEEE sentences were first band-pass filtered into eight bands using zero phase-shift sixth-order Butterworth filters. The filter spacing was based on equal basilar membrane distance ([Greenwood, 1990](#)) across a frequency range of 0.1 to 8 kHz (upper filter cut-offs in Hz: 242, 460, 794, 1307, 2094, 3302, 5155, 8000; filter centre frequencies in Hz: 163, 339, 609, 1023, 1658, 2633, 4130, 6426). The output of each filter was full-wave rectified and low-pass filtered at 30 Hz (zero phase-shift fourth-order Butterworth) to extract the amplitude envelope. The low cut-off value was chosen in order to ensure that no temporal periodicity cues were present in the aperiodic condition.

In order to synthesise aperiodic speech, the envelope of each individual band was multiplied with a broadband noise carrier. In the mixed condition, the envelope of each band was also multiplied with a broadband noise source, but only in time windows where the original speech was unvoiced. Sections that were voiced in the original recordings were synthesised by multiplying the envelopes with a pulse train following the natural F0 contour. The individual pulses had a duration of one sample point, i.e. about 23  $\mu$ s at a sampling rate of 44.1 kHz. The F0 contours of the original sentences were extracted using ProsodyPro version 4.3 ([Xu, 2013](#)) implemented in PRAAT ([Boersma and Weenink, 2013](#)), with the F0 extraction sampling rate set to 100 Hz. The resulting F0 contours were corrected manually where necessary and then used to determine the distance between the individual pulses of the pulse train sources. Based on the original intermittent F0 contours, we also produced artificial continuous F0 contours by interpolation through unvoiced sections and periods of silence. These continuous F0 contours were used to produce the pulse train sources for the periodic condition.

Finally, in all three vocoding conditions, the eight sub-band signals were again band-pass filtered using the same filters as in the analysis stage of the process. Before the individual bands were summed together, the output of each band was adjusted to the same RMS level as found in the original recordings.



**Figure 3.1:** Stimuli. Waveforms, wide-band spectrograms, and F0 contours for one example sentence (*Say it slowly but make it ring clear.*). A) The unprocessed version of the sentence. B) The same sentence processed to have an aperiodic source, C) a mixed source, D) a periodic source, or E) a mixed source and spectrally rotated. The four processed conditions (B–E) were all vocoded with eight frequency bands. The unprocessed version of the sentence in panel A) is shown for the purpose of comparison only.

Spectrally-rotated speech was produced using a technique introduced by [Blessner \(1972\)](#) and implemented in MATLAB. Here, the waveforms of the mixed condition described above were first multiplied with an 8 kHz sinusoid, resulting in a spectral rotation around the midpoint frequency of 4 kHz. Note, that this procedure also renders the rotated speech inharmonic, since the frequencies of the component tones will not be multiples of a particular F0 anymore. The rotated waveforms were then filtered (FFT-based FIR filter, order 256) to have the average long-term speech spectrum ([Byrne et al., 1994](#)) and, finally, scaled to the same RMS level as the original waveforms in the mixed condition.

Fig. 3.1 shows an unprocessed example sentence along with the same sentence processed in the four ways described.



### 3.2.3 Procedure

Each participant listened to 80 aperiodic, 80 mixed, 160 periodic, and 80 rotated sentences. There were twice as many trials in the periodic condition because we wanted to ensure a sufficient number of unintelligible trials. All 4 conditions were presented in blocks of 10 sentences (i.e. 1 complete IEEE sentence list) and the order of the conditions and IEEE lists was randomised. Only the first 40 IEEE lists were used in the main experiment and none of the sentences was presented more than once. Participants were asked to repeat as many words as possible after every sentence. The verbal responses were logged by the experimenter before the next sentence was played and no feedback was given following the responses. The presentation of the stimuli and the logging of the responses was carried out using Presentation version 17.0 (Neurobehavioral Systems, Berkeley, USA). Throughout this study, the term intelligibility will be defined simply as the average number of correctly repeated key words per condition.

Single trials consisted of a silent pre-stimulus interval with random duration (1.5–2.5 s), a stimulus sentence (average duration = 2.04 s, SD = 0.24 s) followed by a silent interval of 0.25 s, a short beep sound signalling the participants to start responding, the spoken responses, and the subsequent logging of the responses by the experimenter.

Before being tested, the subjects were familiarised with the materials by listening to 10 aperiodic, mixed, and periodic examples sentences each (IEEE lists 41–43). During the familiarisation phase, every sentence was directly followed by its unprocessed counterpart, and again followed by the processed sentence.

The main part of the experiment took about 70 minutes to complete and subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating and electrically shielded booth, with the computer signal being fed through the wall onto a separate monitor. Participants sat in a comfortable reclining chair during EEG acquisition and told to not move their eyes during sentence presentation. The stimuli were converted with 16-bit resolution and a sampling rate of 22.05 kHz using a Creative Sound Blaster SB

X-Fi sound card (Dublin, Ireland) and presented over Sennheiser HD650 headphones (Wedemark, Germany). The presentation level was about 71 dB SPL over a frequency range of 0.1 to 8 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

### 3.2.4 EEG recording and analysis

The continuous EEG was recorded using a Biosemi ActiveTwo system (Amsterdam, Netherlands) with 61 Ag-AgCl scalp electrodes mounted on a cap according to the extended international 10-20 system. Four additional external electrodes were used to record the vertical and horizontal electrooculogram (EOG) by placing them on the outer canthus of each eye as well as above and below the left eye. Two more external electrodes were used to record the reference signal from the left and right mastoids. EEG signals were recorded with a sampling rate of 512 Hz and an analogue anti-aliasing low-pass filter with a cut-off frequency of 200 Hz.

EEG data were processed offline using EEGLAB 12.0.2.5b ([Delorme and Makeig, 2004](#)). The continuous waveforms were first down-sampled to 256 Hz, re-referenced to the mean of the two mastoids, and then filtered using zero-phase shift Hamming-windowed sinc FIR filters (EEGLAB firfilt plugin version 1.5.3.; high-pass cut-off 0.01 Hz, transition bandwidth 0.1 Hz; low-pass cut-off 20 Hz, transition bandwidth 0.5 Hz). An independent component analysis (ICA) was used to remove eye artefacts. Epochs ranging from -1000 to 2500 ms were extracted and rejected if amplitudes  $\pm 200$   $\mu$ V, if linear trends exceeded 200  $\mu$ V in a 1000 ms gliding window, or if the trial was lying outside a  $\pm 6$  SD range (for single channels) and  $\pm 3$  SD (for all channels) of the mean voltage probability distribution or the mean distribution of kurtosis values. On average 81% (324/400, SD = 48.3) of the total number of trials passed the rejection procedure.

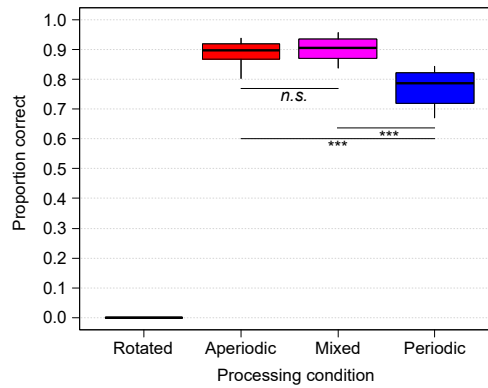
EEG power spectra were computed by estimating the power spectral density (PSD) using Welch's method. The PSD was calculated with a 256-point Hamming window, an oversampling factor of 40, and a window overlap of 50%, resulting in a frequency resolution of 0.025 Hz. The EEG power spectra were computed for the

single trials and averaged afterwards in order to estimate the total spectral power (i.e. time- but not necessarily phase-locked).

The processed EEG data were sorted according to the spoken behavioural responses. For the analysis of periodicity, only trials with all five key words correct were considered, in order to control for the effect of intelligibility. This resulted in an average of 44.2 trials (SD = 8.2) in the aperiodic condition, 44.2 trials (SD = 9.7) in the mixed condition, and 57.9 trials (SD = 17.7) in the periodic condition.

For the analysis of intelligibility, trials in the periodic condition with different numbers of correctly repeated key words and the completely unintelligible rotated condition were separately compared. This resulted in the following average numbers of trials per condition: 8.4 (SD = 4.3) for 0 or 1 key words correct, 12.5 (SD = 5.5) for 2 key words correct, 21.4 (SD = 5.1) for 3 key words correct, 28.9 (SD = 5.9) for 4 key words correct, 57.9 (SD = 17.7) for 5 key words correct, and 67.1 (SD = 10.5) for the rotated condition. In order to obtain the final ERPs, the averaged epochs of each subject were baseline corrected by subtracting the mean amplitude in the -50 to 0 ms window before averaging across subjects.

Statistical differences between conditions were examined using non-parametric cluster-based permutation tests ([Maris and Oostenveld, 2007](#)). Firstly, it was tested whether there was a linear relationship between the amount of periodicity in the stimuli and the ERP amplitude by computing separate two-sided regression *t*-tests for dependent samples with linearly spaced regressors (1–3) at each electrode and for each sample point from 0 to 1000 ms after sentence onset. The same procedure was applied to test whether there was a linear relationship between the intelligibility of the sentences and the ERP amplitude, but this time all sample points in the stimulus window (0–2500 ms) were examined and the regressors were set to values ranging from 1 to 6. Secondly, the individual sample points were merged into clusters if the *t*-values of their regression coefficients were significantly different from 0 at an alpha level of 0.05, and if the same was true for temporally adjacent sample points and at least 3 neighbouring channels. This procedure provides a weak control for false positive findings due to multiple



**Figure 3.2:** Behavioural data. Boxplots showing the average proportion of correctly repeated key words in each of the four speech conditions. The black horizontal lines in the boxplots represent the median value. \*\*\* indicates a  $p$ -value  $< 0.001$ , n.s. stands for not significant.

comparisons by only allowing effects that are coherent in time and space. Next, the  $t$ -values within a given cluster were summed to obtain the cluster-level statistic. The significance probability of a cluster was then assessed by comparing this cluster-level statistic to the one obtained after randomly re-allocating the individual trials to the conditions. This step was repeated 1000 times and the proportion of these Monte-Carlo iterations in which the cluster-level statistic was exceeded then determined the final cluster  $p$ -value.

The same statistical technique was applied to test whether there was a linear relationship between pre-stimulus alpha power and sentence intelligibility in the periodic condition, but in this case the EEG power spectrum in the pre-stimulus period (-1000–0 ms) was first averaged over a frequency window from 7 to 10 Hz in each condition. Here, the regressors were set to values from 1 to 5, corresponding to the number of correct key words. Consequently, only a single regression coefficient was computed per electrode, and these were subsequently clustered according to their  $p$ -values and spatial adjacency.

### 3.3 Results

#### 3.3.1 Behavioural data

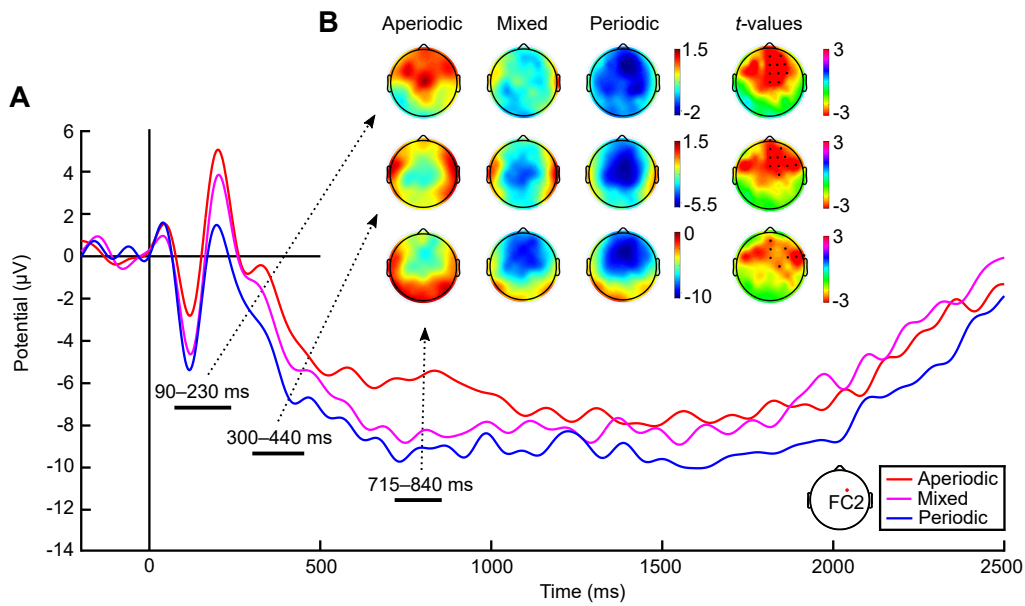
The averaged spoken behavioural responses obtained after each trial (Fig. 3.2) show that the aperiodic and mixed conditions are equally intelligible (88.8% and 90.0%

correct key words on average;  $t(17) = -1.60$ ,  $p = 0.13$ ), while the rotated condition is completely unintelligible (0%), and periodic speech is slightly less intelligible (77.4%) than aperiodic ( $t(17) = -8.42$ ,  $p < 0.001$ ) and mixed speech ( $t(17) = -11.60$ ,  $p < 0.001$ ). Furthermore, we compared the responses to the first and the second half of the trials in the periodic condition and found no significant differences (77.8% and 77.0%;  $t(17) = 0.70$ ,  $p = 0.49$ ), indicating that there were no learning effects over the course of the 160 trials.

### 3.3.2 Periodicity

As shown by the ERP traces recorded at electrode FC2 in Fig. 3.3A, the three conditions varying in acoustic periodicity (aperiodic, mixed, and periodic speech) all elicited an auditory-evoked P1-N1-P2 complex after sentence onset, followed by an acoustic change complex (ACC, consisting of CP1, CN1, and CP2 components) from about 300 to 500 ms in response to the onset of the second syllable (Pratt, 2011). Furthermore, all three conditions showed a sustained negativity from about 300 to 2500 ms past sentence onset.

Crucially, after the initial P1 component, peaking at around 50 ms after sound onset, the ERPs in the three conditions were found to diverge, showing greater negative amplitudes with more periodicity. This parametric effect is observable until about one second after sound onset and thus considerably overlaps with the slow negativity. A cluster-corrected linear regression  $t$ -test including all three conditions confirmed that there was a significant linear negative relationship during this time window by returning three separate significant clusters in the right fronto-central scalp region: the first one was found during the period of the N1 and P2 components between about 90 to 230 ms ( $p = 0.034$ ), the second cluster ranging from about 300 to 440 ms ( $p = 0.028$ ) coincided with the ACC, and the third cluster was observed between about 715 to 840 ms ( $p = 0.03$ ) after sound onset. The average voltage distributions of each condition during the three clusters along with  $t$ -value maps depicting the scalp distributions of statistical differences for each cluster are shown in Fig. 3.3B.



**Figure 3.3:** Periodicity. A) Grand average ERPs recorded at electrode FC2 for fully intelligible trials (all 5 key words correctly repeated) in the aperiodic, mixed, and periodic conditions. The three thick black lines below the ERP traces indicate time windows during which there was a significant linear negative relationship between the amount of periodicity in the stimuli and the ERP amplitude ( $p < 0.05$ ). ERP waveforms were low-pass filtered at 10 Hz for illustration purposes. B) Voltage maps showing the mean activity during the three significant time windows for each condition. In the three  $t$ -value maps on the far right, black dots indicate electrodes whose  $p$ -values were  $< 0.05$  at each sample point during the respective time window.

### 3.3.3 Intelligibility

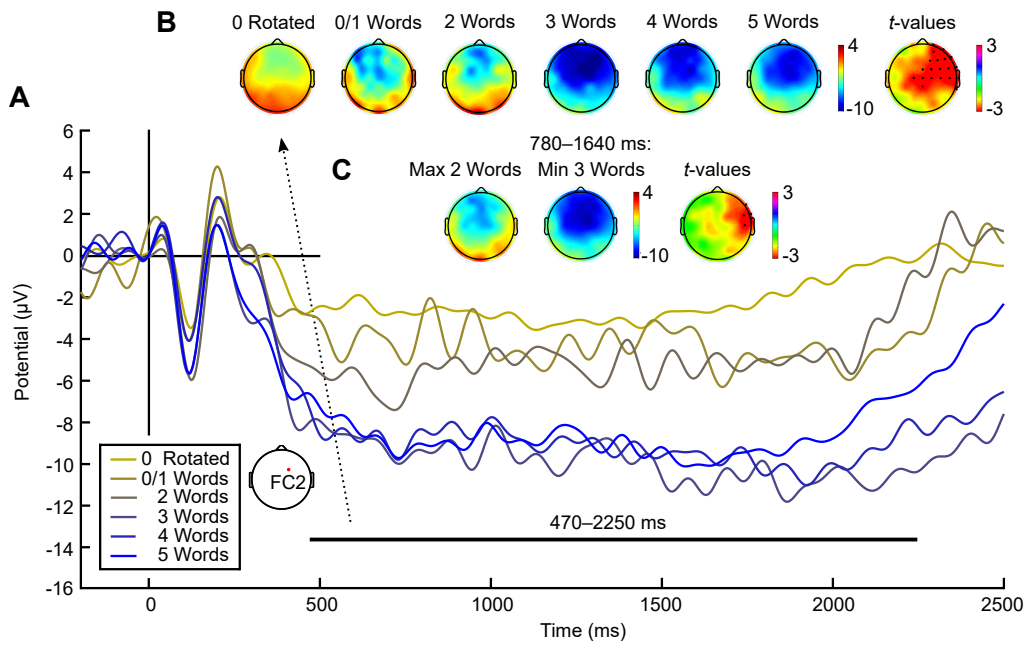
In order to analyse how the ERPs were affected by the intelligibility of the stimulus sentences, trials in the periodic condition were sorted into five categories, according to the spoken responses of the participants (zero or one, two, three, four, and five key words correct). Additionally, spectrally-rotated speech was included as a completely unintelligible control condition.

As illustrated in Fig. 3.4A, which shows the ERPs recorded at electrode FC2, all six conditions elicited a slow negativity from about 300 to 2500 ms after the beginning of the sentences. This slow negativity, taken to be a CNV, had the smallest amplitude in the rotated condition, followed by slightly larger amplitudes for trials in the periodic condition with zero or one and two correct key words, and substantially larger amplitudes for trials in the periodic condition with three, four and five

key words correct. A cluster-corrected regression  $t$ -test including all six conditions returned one large significant cluster ( $p = 0.004$ ) from about 470 to 2250 ms, confirming that there was a linear negative relation between the intelligibility of the sentences and the amplitude of the CNV. The corresponding  $t$ -map shows that this cluster included a large number of electrodes in the central and right fronto-temporal scalp region (see map at far right in Fig. 3.4B). The voltage maps showing the ERP amplitudes averaged over the duration of the whole cluster in each condition confirm that the activity was strongest in the fronto-central scalp region and slightly lateralised to the right, particularly for the more intelligible conditions (three or more correct key words, Fig. 3.4B).

In order to test whether the smaller CNV in the conditions with two or less correct key words were due to the low trial numbers, we computed the Spearman rank correlation coefficients between the number of trials per subject and their CNV amplitudes (averaged over all 61 scalp electrodes and the whole stimulus window). These correlations were in both cases not significant (0/1 words:  $r = -.24$ ,  $p = 0.34$ ; 2 words:  $r = 0.24$ ,  $p = 0.33$ ), indicating that the observed effect was not driven by the subjects with the fewest trials within each condition.

In addition to the finding that CNV amplitudes were larger when the sentences were more intelligible to the subjects, the data in Fig. 3.4 also show that the six conditions appeared to group into three distinct categories (rotated, maximally two key words, and minimally three key words). In order to follow up this observation, we tested if there were any significant differences within these categories. Firstly, trials with zero or one correct key words were compared to trials with two correct key words using a cluster-based  $t$ -test. Secondly, trials with three, four, and five correct key words were compared using a cluster-based ANOVA. Both tests revealed no significant differences at any point during the stimulus window (0–2500 ms). Based on this finding, trials in the periodic condition were pooled into a more and less intelligible category (maximally two versus minimally three correct key words, respectively) and separately compared, leaving out the rotated condition to ensure a test that is free of any acoustic confounds. For this post-hoc analysis, a cluster-



**Figure 3.4:** Intelligibility. A) Grand average ERPs recorded at electrode FC2 for the completely unintelligible rotated condition and trials in the periodic condition with 0/1, 2, 3, 4, or 5 correctly repeated key words. The thick black line below the ERP traces indicates the time window during which there was a significant linear negative relationship between the intelligibility of the stimuli and the ERP amplitude ( $p < 0.01$ ). ERP waveforms were low-pass filtered at 10 Hz for illustration purposes. B) Voltage maps showing the mean activity during the significant time window for each condition. In the  $t$ -value map on the far right, black dots indicate electrodes whose  $p$ -values were  $< 0.01$  at each sample point during the respective time window. C) Voltage distributions and  $t$ -map showing the mean activity during the time window in which the ERP amplitudes of the pooled less (maximally 2 key words) and more (minimally 3 key words) intelligible trials in the periodic condition differed significantly ( $p < 0.05$ ).

corrected regression  $t$ -test including all sample points in the significant time window (470–2250 ms) revealed one cluster with a  $p$ -value of 0.036 from about 780 to 1640 ms. The voltage maps averaged over this significant time window show that the activity is lateralised to the right in the more intelligible condition, which is confirmed by the location of the significant cluster of electrodes in the right temporal scalp region (Fig. 3.4C).

### 3.3.4 Pre-stimulus alpha power

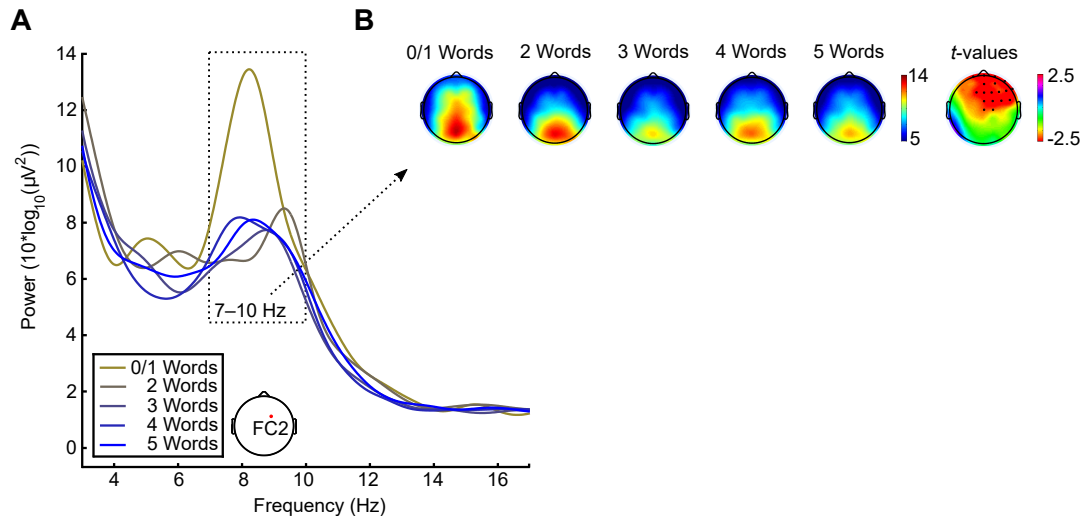
In a final analysis, we tested whether the amount of alpha power in the silent pre-stimulus period before sentence onset stands in relation to the intelligibility of the



stimulus sentences in the periodic condition. As shown by the line plot in Fig. 3.5A, depicting the average EEG power spectra in the pre-stimulus window (-1000–0 ms) recorded at electrode FC2, slow alpha power (7–10 Hz) was markedly increased before the least intelligible trials, with maximally one out of five correctly repeated key words. Furthermore, it can be seen that the differences between the five conditions were indeed confined to the slow alpha range. The scalp distributions of the average spectral power in this frequency window show peaks of activity over the occipital scalp region in all five conditions, along with a widespread power increase extending into the anterior scalp region for the least intelligible trials (Fig. 3.5B). A cluster-based regression *t*-test comparing the averaged pre-stimulus slow alpha power (7–10 Hz/-1000–0 ms) in all five conditions at each electrode revealed a large cluster comprising 18 significant electrodes in the right frontal scalp region ( $p = 0.016$ , see *t*-map at far right of Fig. 3.5B).

Same as for the ERPs, the Spearman rank correlation coefficients between the number of trials per subject and the amount of slow alpha power (averaged over all 61 scalp electrodes, and the whole pre-stimulus window) was not significant for the conditions with the fewest trials (0/1 words:  $r = 0.07$ ,  $p = 0.78$ ; 2 words:  $r = 0.06$ ,  $p = 0.83$ ), showing that the results within these conditions were not biased by the subjects with the lowest numbers of trials.

As the relation between slow alpha power in the pre-stimulus window and intelligibility was not strictly linear, further tests were performed. Firstly, the four conditions with two or more correct key words, which appeared not to differ regarding the amount of slow alpha power, were separately compared using a cluster-based ANOVA. This test did not reveal any significant differences between the four conditions. However, when all trials with two or more correct key words were pooled into a single condition and compared to the least intelligible trials using a one-tailed cluster-corrected *t*-test, the same significant cluster of electrodes as shown in Fig. 3.5B was obtained, which confirms that slow alpha power was increased for the least intelligible trials only.



**Figure 3.5:** Pre-stimulus alpha power. A) Line plot showing the averaged EEG power spectra in the silent pre-stimulus window (-1000–0 ms), recorded at electrode FC2, for trials in the periodic condition with 0/1, 2, 3, 4, or 5 correctly repeated key words. B) Scalp maps of the mean alpha power in the 7 to 10 Hz frequency window for each of the five conditions. In the  $t$ -map on the far right, black dots indicate electrodes with  $p$ -values < 0.05.

### 3.4 Discussion

The purpose of the present study was to tease apart effects of acoustics and intelligibility on the ERPs in response to speech. It was found, firstly, that more acoustic periodicity in the speech signals parametrically rendered the ERP waveforms during the first second after sentence onset more negative. Periodicity thus appears to amplify the evoked cortical response in the early period after sound onset. Secondly, we observed a CNV that was larger when the speech signals were more intelligible to the participants. However, this relationship was not strictly linear, as the amplitude of the negativity differed significantly between trials with less and more than half of the key words correctly repeated, but not within these categories. Additionally, slow alpha power (7–10 Hz) in the silent baseline interval preceding the sentences that turned out to be least intelligible to the participants was found to be markedly increased, while there was no difference between the rest of the trials.

### 3.4.1 Periodicity

The finding that more periodicity leads to larger negative ERP amplitudes is in line with pitch perception studies reporting greater neural responses to sound input that possesses a pitch (e.g. [Chait et al., 2006](#); [Griffiths et al., 2010](#); [Norman-Haignere et al., 2013](#)). As we have controlled for differences in intelligibility across conditions by only including trials with all five key words correctly repeated, and sentence materials as well as the behavioural task were the same throughout the experiment, it seems unlikely that any cognitive process can explain this effect. Furthermore, the effect was significant from as early as 90 ms after acoustic onset, a latency which is generally thought to be dominated by responses to the acoustic properties of a stimulus ([Picton et al., 1974](#); [Pratt, 2011](#)). However, the effect was not confined to the time window of AEPs and ACC, i.e. until about 500 ms post-onset, but present until almost one second after sound onset, classifying as a sustained pitch response ([Gutschalk et al., 2004](#)). The current results thus stress the importance of taking the acoustic properties of the stimuli into account when investigating speech perception, particularly when the duration of the stimuli is relatively short (e.g. single words).

### 3.4.2 Intelligibility

As outlined in the introduction, slow cortical potentials may reflect working memory operations, the level of attention spent on a task, and how prepared to respond a subject is. Regarding the task to verbally repeat relatively long auditorily presented sentences, it appears likely that all three factors play a role. Firstly, larger amounts of verbal material have to be retained in working memory when the sentences are more intelligible. Secondly, when the stimulus sentences were less intelligible to them, subjects were presumably paying less attention to a task they realised they could not accomplish. Similarly, the inability to understand the materials is necessarily going along with failing to prepare for the subsequent verbal response. In line with this interpretation, significant differences in CNV amplitude were not observed right after sentence onset, but started to emerge a few hundred milliseconds after, suggesting that the subjects first had to process the initial part of the sentences before these cognitive processes were triggered.

Although the task used in this study was not typical for eliciting a CNV, the fact that the amplitude of the slow negativity did not increase further when three or more key words per sentence were correctly repeated provides further evidence for this interpretation. CNV amplitudes have often been reported to be limited, or even to have an inverted u-shaped relationship with task demand ([Birbaumer et al., 1990](#); [Kononowicz and Penney, 2016](#)). In turn, however, the monotonic but not strictly linear relation of speech intelligibility and CNV amplitude observed in the current study also suggests that the CNV cannot be used as an accurate predictor of speech intelligibility scores.

In a recent study that resembles the current one to some extent, [Wöstmann et al. \(2015b\)](#) have reported a slow negativity, which was also taken to be a CNV, in an auditory number comparison task. In their study, subjects had to remember numbers in the presence of a competing talker in the background, and the signal mixture was furthermore acoustically degraded. Crucially, more severe degradations resulted in larger CNV amplitudes, although the intelligibility of the numbers and the task performance decreased somewhat. Wöstmann et al. thus concluded that the CNV amplitude serves as a measure of expected task difficulty and listening effort. Although it remains to be investigated how the CNV in response to speech presented in background noise varies when the intelligibility fluctuates over a wider range, this suggests that slow cortical potentials may reflect different cognitive processes for speech presented in quiet and in noise. Importantly, in the present study subjects could not know whether they would be able to understand a particular sentence in the periodic condition before it was played to them. Hence, the differences in CNV amplitude for the more or less intelligible trials cannot be explained by the expected task difficulty, which was assumed to be constant.

### 3.4.3 Pre-stimulus alpha power

The slow alpha power before the least intelligible trials was found to have a broad scalp distribution extending into the anterior scalp region. As summarised by [Klimesch \(1999\)](#), slower alpha frequencies generally have a more anterior scalp distribution than faster ones and the distribution found in the current study also cor-

responds well with the example scalp map provided in this review paper. As shown by [Laufs et al. \(2006\)](#), there appear to be two distinguishable alpha networks, one that comprises occipital vision areas and a second one in fronto-parietal areas associated with attention. The scalp location of the cluster of significant electrodes found in the current study corresponds well with that of the right-lateralised ventral fronto-parietal attention network, which is deactivated when subjects focus on a task ([Corbetta et al., 2008](#); [Corbetta and Shulman, 2002](#)). Deactivation of this network has been associated with the prevention of irrelevant task switching ([Shulman et al., 2007](#)) and our data suggest that this deactivation may coincide with a decrease in alpha power. The location of this effect is also well in line with the results of [Strauß et al. \(2015\)](#), who have observed the strongest differences in alpha phase before correct and incorrect trials in a lexical decision task in this region.

As described by [Mazaheri and Jensen \(2008, 2010\)](#), slow ERP deflections may be caused by amplitude fluctuations of induced alpha power because the peaks of alpha oscillations appear to be more strongly modulated than the troughs. However, this explanation does not seem to apply to the current results, since the amplitude of the slow negativity varies independently of the pre-stimulus alpha power. That is, the slow alpha power was only increased before the least intelligible trials (zeros or one correct key words), but the CNV had a similar amplitude for these trials and those with two correct key words. Hence, same as for the CNV, the non-linear relationship of pre-stimulus alpha power and intelligibility does not allow the accurate prediction of speech intelligibility rates.

### 3.5 Conclusion

The current study investigated cortical EEG responses to auditorily presented sentences with a focus on the differential contributions of acoustics and intelligibility. Firstly, more acoustic periodicity in the stimuli was found to render the ERPs during the first second after speech onset more negative. This demonstrates that acoustic factors should not be disregarded in neuroscientific studies investigating speech perception, even when focussing on cognitive processes. Secondly, we observed a CNV from about half a second after sentence onset, the amplitude of which was

larger when the sentences were more intelligible to the participants. Additionally, slow alpha power before the least intelligible sentences was significantly higher than before the rest of the trials. However, as the latter two measures did not vary precisely as a function of the number of correctly repeated key words and did not appear to co-vary, neither appears to be an accurate predictor of speech intelligibility.

## Chapter 4

### Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech<sup>1</sup>

#### 4.1 Introduction

In order to shed light on the underlying neural mechanisms and cognitive processes involved when attempting to understand spoken speech, a growing number of magneto- and electroencephalographic (M/EEG) studies focus on the time-frequency properties of the neural signals rather than traditional waveform analyses (for reviews see [Giraud and Poeppel, 2012](#); [Weisz and Obleser, 2014](#)). The current experiment adds to existing knowledge by investigating effects of acoustics and intelligibility separately, two factors that usually vary together when speech signals are acoustically manipulated. Specifically, we manipulated the amount of acoustic periodicity, while controlling for differences in intelligibility, and vice versa. In the context of speech, periodicity denotes that a sound is produced by the periodic vibrations of the vocal folds, resulting in voiced speech with a pitch corresponding to the vibration rate. Unvoiced speech sounds, in contrast, emanate from constrictions in the vocal tract and have aperiodic fluctuations in energy, leading to a noisy sound quality and the absence of a pitch.

A popular speech processing technique that has been used in the neurosciences (e.g. [Davis and Johnsrude, 2003](#); [Obleser and Weisz, 2012](#); [Peelle et al., 2013](#); [Scott et al., 2000](#)) as well as in psychoacoustic studies concerned with the simulation of

---

<sup>1</sup>This chapter is based on the same EEG data as chapter 3, which is why they bear some resemblance, and has been published as: Steinmetzger, K. and Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia* 95, 173–181.

cochlear implants (e.g. [Qin and Oxenham, 2003](#); [Schoof et al., 2013](#); [Shannon et al., 1995](#)), is noise-vocoding (henceforth referred to as the aperiodic condition). By filtering the unprocessed input speech into a specified number of frequency bands, it allows the spectral resolution of the synthesised output speech to be varied in a controlled manner, a feature that is closely related to speech intelligibility. At the same time, using noise as source results in a loss of the natural mix of voiced and voicelessness, and consequently also any voice pitch information, making it resemble whispered speech.

Nevertheless, our previous behavioural work ([Steinmetzger and Rosen, 2015](#); i.e. chapter 2) has shown that preserving periodicity information in a vocoder (henceforth the *mixed* condition) does not lead to improved intelligibility rates. This suggests that periodicity information, despite its salience, is a redundant cue, at least in non-tonal languages and quiet listening conditions. The first question the current study addresses, is thus whether EEG time-frequency responses are similarly unaffected by the absence of periodicity.

To enable a more comprehensive investigation of the effects of periodicity, we included a third processing condition in which the same speech materials were synthesised with a completely periodic source (henceforth the *periodic* condition). Acoustically this condition is in fact closer to natural speech (which is voiced about 50% of the time – [Dellwo et al., 2007](#); [Fourcin, 2010](#)), than aperiodic noise-vocoded speech. However, because natural speech does not contain periodic sounds with much energy in the frequency region above 4 kHz, it sounds very unnatural. Additionally, periodicity is such a salient cue that it obscures weaker cues such as intensity differences, thereby making the information transmitted contradictory. For unvoiced fricatives like /s/ and /f/, for example, aperiodic high-frequency energy is missing and replaced by periodic energy, which makes it difficult to identify these sounds. Consequently, periodic speech has substantially lower intelligibility rates than the other two conditions ([Ardoint et al., 2014](#); [Steinmetzger and Rosen, 2015](#); i.e. chapter 2).



In order to control for this expected difference in intelligibility, the single trials were sorted according to the spoken responses of the participants. This approach was also chosen to enable a direct comparison of intelligible and unintelligible trials in the periodic condition. Consequently, the current study also provides the opportunity to investigate how the EEG time-frequency responses are affected by the intelligibility of the speech materials after controlling for systematic acoustic differences. This approach is akin to studies generating a pop-out effect by presenting the same stimulus materials twice, first without any additional information and then again after providing a written transcript (Sohoglu et al., 2012) or the unprocessed recording (Millman et al., 2015), but avoids any predictive top-down processing.

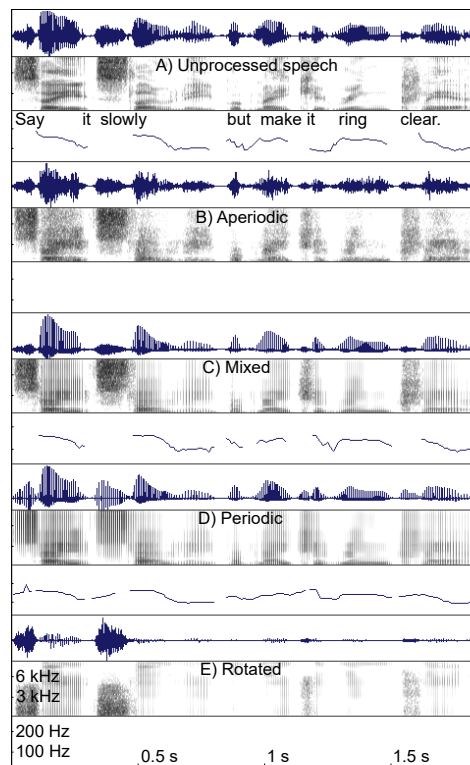
In addition to the acoustically similar unintelligible trials, the current study included completely unintelligible spectrally-rotated speech as a second control condition (henceforth the *rotated* condition). Rotated speech has a similar spectro-temporal complexity as unrotated speech and has been used in several of the studies mentioned above (Becker et al., 2013; Peelle et al., 2013; Scott et al., 2000). Yet, apart from not being a precise acoustic match, the obvious meaninglessness casts doubts on whether it is indeed an adequate control condition. The design of the current experiment thus also serves to directly compare these two control condition types.

Importantly, in contrast to the event-related potentials (ERPs), the EEG time-frequency analyses in the current paper were not assumed to be affected by the perception of a voice pitch. Direct cortical recordings and MEG experiments have shown that the presence of a pitch coincides with increased high gamma power ( $>80$  Hz; e.g. Griffiths et al., 2010; Sedley et al., 2012). However, due to potential muscle artefacts and the low signal strength when recorded with cortical EEG, we did not include these frequencies in the current analysis. Moreover, functional magnetic resonance imaging (fMRI) signals, which fluctuate at rates of less than 0.5 Hz (He and Raichle, 2009), have been shown to be larger for signals with a pitch (e.g. Norman-Haignere et al., 2013; Patterson et al., 2002). Yet, frequencies below 0.5

Hz similarly lie outside the possible frequency range of EEG time-frequency analyses, because they would require excessively long baseline and stimulus windows.

Based on the results of [Strauß and colleagues \(2014a\)](#), who have recently reported increased theta activity (here 3–7 Hz) in a left-lateralised fronto-temporal network for so-called ambiguous pseudo-words in an auditory word recognition task, we expected increased theta power in the periodic condition. The pseudo-words used by [Strauß et al. \(2014a\)](#) were characterised by a wrong core vowel, making them resemble the periodic condition in the current study to some extent. Theta oscillations have also been associated with the storage of sequentially presented verbal information in working memory and the phonological loop ([Roux and Uhlhaas, 2014](#)). Based on this idea, [Strauß et al. \(2014a\)](#) suggested that subjects may have internally rehearsed the unusual pseudo-words in order to classify them as words or non-words. More generally, this effect was taken to indicate an information processing conflict ([Botvinick et al., 2001](#), [Botvinick et al., 2004](#)), although studies eliciting response conflicts in non-speech tasks, for example by using the Stroop paradigm, have usually reported mid-frontal theta power increases ([Cohen and Donner, 2013](#); [Hanslmayr et al., 2008](#)).

A recent theoretical approach has linked increased alpha power (~7–13 Hz) to the selective inhibition of brain areas that are not currently task relevant ([Jensen and Mazaheri, 2010](#)). Applied to speech perception, it has been proposed that alpha oscillations might be actively enhanced in order to cope with a demanding task, particularly listening to speech in the presence of background noise (e.g. [Strauß et al., 2014b](#); [Wöstmann et al., 2015a](#)). For words presented in quiet listening conditions, on the other hand, alpha activity was found to be increasingly suppressed with higher intelligibility levels ([Becker et al., 2013](#); [Obleser and Weisz, 2012](#)). However, in these studies the intelligibility of the mostly noise-vocoded stimuli varied along with their acoustic properties (i.e. the number of frequency bands in the vocoder) and hence also the subjective listening effort, which is similarly thought to depend on the degree of acoustic degradation ([Obleser and Weisz, 2012](#); [Wöstmann et al., 2015a](#)). Sorting the trials in the periodic condition according to



**Figure 4.1:** Stimuli. Waveforms, wide-band spectrograms, and F0 contours for one example sentence (*Say it slowly but make it ring clear.*). A) The unprocessed version of the sentence and the same sentence processed to have B) an aperiodic source, C) a mixed source, D) a periodic source, or E) a mixed source and spectrally rotated. The four processed conditions (B–E) all have eight frequency bands, i.e. the same spectral resolution. The unprocessed version of the sentence in panel A) is shown for the purpose of comparison only.

the spoken behavioural responses provided the opportunity to test whether there is indeed a direct relation between alpha suppression and speech intelligibility.

## 4.2 Methods

### 4.2.1 Participants

Eighteen normal-hearing right-handed subjects (8 females, mean age = 21.6 years, SD = 2.3 years) took part in the study. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz. All subjects gave written consent and the study was approved by the UCL research ethics committee.

### 4.2.2 Stimuli

The stimulus materials used in this experiment were recordings of the IEEE sentences (Rothausen et al., 1969) spoken by an adult male Southern British English talker with a mean F0 of 121.5 Hz that were cut at zero-crossings right before sentence onset and normalised to a common root-mean-square (RMS) level. The IEEE sentence corpus consists of 72 lists with 10 sentences each and is characterized by similar phonetic content and difficulty across the lists, as well as an overall low semantic predictability. Every sentence contains five key words (nouns, verbs, or adjectives; e.g. *Say it slowly but make it ring clear.*).

All stimulus materials were processed prior to the experiment using a channel vocoder implemented in MATLAB (Mathworks, Natick, MA). For all three vocoding conditions (aperiodic, mixed, and periodic) the original recordings of the IEEE sentences were first band-pass filtered into eight bands using zero-phase-shift sixth-order Butterworth filters. The filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of 0.1 to 8 kHz (upper filter cut-offs in Hz: 242, 460, 794, 1307, 2094, 3302, 5155, 8000; filter centre frequencies in Hz: 163, 339, 609, 1023, 1658, 2633, 4130, 6426). The output of each filter was full-wave rectified and low-pass filtered at 30 Hz (zero-phase-shift fourth-order Butterworth) to extract the amplitude envelope. The low cut-off value was chosen in order to ensure that no temporal periodicity cues were present in the aperiodic condition.

In order to synthesise aperiodic speech, the envelope of each individual band was multiplied with a broadband white noise carrier. In the mixed condition, the envelope of each band was also multiplied with a broadband white noise, but only in time windows where the original speech was unvoiced. Sections that were voiced in the original recordings were synthesised by multiplying the envelopes with a pulse train following the natural F0 contour. The individual pulses had a duration of one sample point, i.e. about 23  $\mu$ s at a sampling rate of 44.1 kHz. The F0 contours of the original sentences were extracted using ProsodyPro version 4.3 (Xu, 2013) implemented in PRAAT (Boersma and Weenink, 2013), with the F0 extraction sam-

pling rate set to 100 Hz. The resulting F0 contours were corrected manually where necessary and then used to determine the distance between the individual pulses of the pulse train sources. Based on the original intermittent F0 contours, we also produced artificial continuous F0 contours by interpolation through unvoiced sections and periods of silence. These continuous F0 contours were used to produce the pulse train sources for the periodic condition.

Finally, in all three vocoding conditions, the eight sub-band signals were again band-pass filtered using the same filters as in the analysis stage of the process. Before the individual bands were summed together, the output of each band was adjusted to the same RMS level as found in the original recordings.

Spectrally-rotated speech was produced using a technique introduced by [Blessner \(1972\)](#) and implemented in MATLAB. Here, the waveforms of the mixed condition described above were first multiplied with an 8 kHz sinusoid, resulting in a spectral rotation around the midpoint frequency of 4 kHz. Note that this procedure also renders the rotated speech inharmonic, since the frequencies of the component tones will not be multiples of a particular F0 anymore. The rotated waveforms were then filtered (FFT-based FIR filter, order 256) to have the average long-term speech spectrum ([Byrne et al., 1994](#)) and, finally, scaled to the same RMS level as the original waveforms in the mixed condition.

Fig. 4.1 shows an unprocessed example sentence along with the same sentence processed in the four ways described.

### 4.2.3 Procedure

Each participant listened to 80 aperiodic, 80 mixed, 160 periodic, and 80 rotated sentences. There were twice as many trials in the periodic condition because we wanted to ensure a sufficient number of unintelligible trials. All 4 conditions were presented in blocks of 10 sentences (i.e. 1 complete IEEE sentence list) and the order of the conditions and IEEE lists was randomised. Only the first 40 IEEE lists were used in the main experiment and none of the sentences was presented more than once. Participants were asked to repeat as many words as possible after every sentence. The verbal responses were logged by the experimenter before the

next sentence was played and no feedback was given following the responses. The presentation of the stimuli and the logging of the responses was carried out using Presentation version 17.0 (Neurobehavioral Systems, Berkeley, USA).

Single trials consisted of a silent pre-stimulus interval with random duration (1.5–2.5 s), a stimulus sentence (average duration = 2.04 s, SD = 0.24 s) followed by a silent interval of 0.25 s, a short beep signalling the participants to start responding, the spoken responses, and the subsequent logging of the responses by the experimenter.

Before being tested, the subjects were familiarised with the materials by listening to 10 aperiodic, mixed, and periodic example sentences each (IEEE lists 41–43). During the familiarisation phase every sentence was directly followed by its unprocessed counterpart, and again followed by the processed sentence.

The main part of the experiment took about 70 minutes to complete and subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating and electrically-shielded booth, with the computer signal being fed through the wall onto a separate monitor. Participants sat in a comfortable reclining chair during EEG acquisition and told to not move their eyes during sentence presentation. The stimuli were converted with 16-bit resolution and a sampling rate of 22.05 kHz using a Creative Sound Blaster SB X-Fi sound card (Dublin, Ireland) and presented over Sennheiser HD650 headphones (Wedemark, Germany). The presentation level was about 71 dB SPL over a frequency range of 0.1 to 8 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

#### **4.2.4 EEG recording and pre-processing**

The continuous EEG was recorded using a Biosemi ActiveTwo system (Amsterdam, Netherlands) with 61 Ag-AgCl scalp electrodes mounted on a cap according to the extended international 10-20 system. Four additional external electrodes were used to record the vertical and horizontal electrooculogram (EOG) by placing them on the outer canthus of each eye as well as above and below the left eye. Two more external electrodes were used to record the signal from the left and right mastoids.

EEG signals were recorded with a sampling rate of 1024 Hz and an analogue anti-aliasing low-pass filter with a cut-off frequency of 200 Hz.

The EEG data were processed offline using EEGLAB 13.5.4b (Delorme and Makeig, 2004). The waveforms were first down-sampled to 512 Hz, re-referenced to the mean of the two mastoids, and then filtered (zero-phase shift Hamming-windowed sinc FIR filter, order 3380, using the firfilt plugin version 1.5.3.) with a 0.1 Hz high-pass filter and a 100 Hz low-pass filter. An independent component analysis (ICA) was used to remove artefacts caused by eye blinks, eye movements, and muscular activity. Epochs ranging from -1000 to 3000 ms around sentence onset were extracted and rejected if amplitudes exceeded  $\pm 120 \mu\text{V}$ , if linear trends exceeded  $120 \mu\text{V}$  in a 500 ms gliding window, or if the trial was lying outside a  $\pm 6$  SD range (for single channels) and  $\pm 3$  SD (for all channels) of the mean voltage probability distribution or the mean distribution of kurtosis values. On average 73.5% (294/400, SD = 13%, range = 57–95%) of the total number of trials passed the rejection procedure.

#### 4.2.5 EEG time-frequency analysis

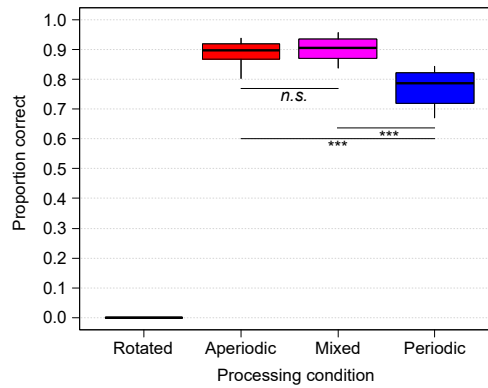
In order to ensure the same signal-to-noise ratio in each condition, all analyses in the present paper are based on matched trial numbers across conditions. For each individual participant, the number of trials was determined by the condition with the fewest trials, with excess trials in the other conditions omitted randomly.

The pre-processed EEG data were first sorted according to the spoken responses. For the analysis of periodicity, only trials with all five key words correct were considered in order to control for differences in intelligibility. This resulted in an average of 35.7 trials in each of the 4 processing conditions (SD = 9.7, range = 23–56). For the analysis of intelligibility, trials in the periodic condition with all five key words correct were compared to trials with maximally two correct key words (i.e. less than half of the sentence correctly repeated). Three participants with less than 10 unintelligible trials were excluded due to the low signal-to-noise ratio of the data. The remaining 15 participants (8 females) had an average number of 21.7 trials per condition (SD = 7.3, range = 14–35).

The time-frequency decomposition of the pre-processed and sorted data was conducted by computing the event-related spectral perturbation (ERSP) as implemented in EEGLAB. The ERSP is a measure of the relative change in power from baseline to stimulus period ([Makeig, 1993](#)). For each time-frequency point in the stimulus window, the spectral power is divided by the average power of the respective frequency bin in the baseline window and transformed into a dB value. The data were analysed from 1 to 30 Hz in 200 log-spaced frequency steps by convolving them with a set of Morlet wavelets, whose widths increased linearly with frequency from 1 to 15 cycles. This resulted in an analysis window ranging from -442.4 to 2442.4 ms around sentence onset. For the sake of simplicity, the rounded values -500 to 2500 ms will be used henceforth. Within this window, the ERSP for each of the 200 frequency bins was calculated 100 times, resulting in a decomposition step size of about 29 ms. In order to limit the overlap between pre- and post-stimulus activity due to the windowing of the time-frequency analysis, the baseline window lasted from -1000 to -100 ms ([Shahin et al., 2009](#), [Shahin et al., 2008](#)). All analyses in the current study are based on estimates of the total EEG power. In order to obtain the total (i.e. time- but not necessarily phase-locked) EEG power, the ERSP was computed for the single trial data and averaged afterwards ([Tallon-Baudry and Bertrand, 1999](#)).

Statistical differences between conditions were examined using non-parametric cluster-based permutation tests ([Maris and Oostenveld, 2007](#)). Firstly, it was tested whether there was a linear relationship between the amount of periodicity (aperiodic vs. mixed vs. periodic) or intelligibility (rotated vs. unintelligible periodic vs. intelligible periodic) and the total EEG power by computing separate two-sided regression t-tests for dependent samples at each electrode. In both cases, the whole stimulus window (0-2500 ms) and the complete array of analysed neural frequencies (1–30 Hz) was included in the test. Individual time-frequency sample points were considered to belong to a cluster if their  $F$ -values fell below the alpha level of 0.05, if the same was true for at least 3 neighbouring channels, and if they





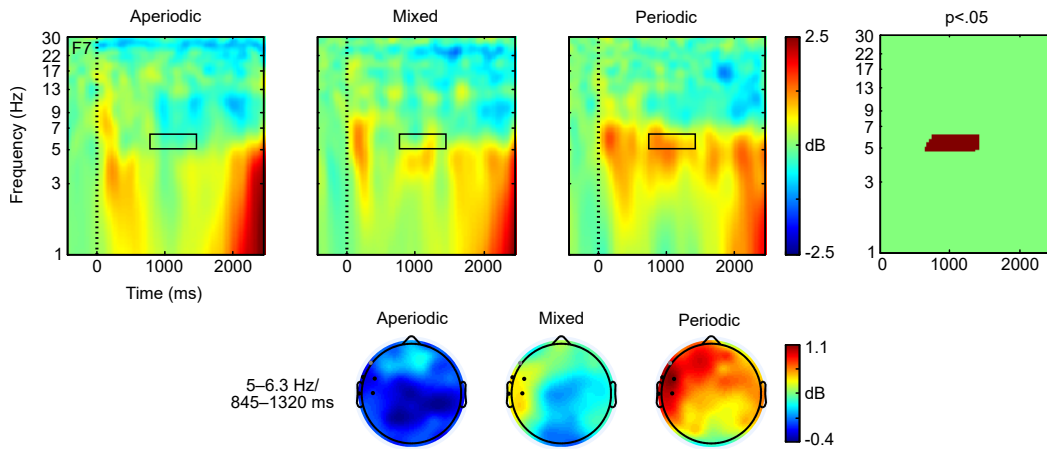
**Figure 4.2:** Behavioural data. Boxplots showing the average proportion of correctly repeated key words in each of the four processing conditions. The black horizontal lines in the boxplots indicate the median value. \*\*\* indicates a  $p$ -value  $< 0.001$ , *n.s.* stands for non-significant.

were connected to other significant sample points surrounding them. This procedure provides a weak control for false positive findings due to multiple comparisons by only allowing effects that are coherent in time, frequency, and space. Next, the individual  $F$ -values within a given cluster were summed to obtain the cluster-level statistic. The significance probability of a cluster was then assessed by comparing this cluster-level statistic to the one obtained after randomly re-allocating the individual trials to the conditions. This step was repeated 1000 times and the final cluster  $p$ -value was then determined by the proportion of these Monte Carlo iterations in which the cluster-level statistic was exceeded. The same statistical technique was also applied when two conditions were compared, but in this case two-sided  $t$ -tests were used in order to determine the  $p$ -values of the individual time-frequency sample points. In all statistical tests reported in this paper, an effect was considered to be significant if the cluster  $p$ -value was smaller than 0.05.

### 4.3 Results

#### 4.3.1 Behavioural data

The averaged spoken behavioural responses obtained after each trial (Fig. 4.2) show that the aperiodic and mixed conditions are equally intelligible (88.8% and 90.0% correct key words on average;  $t(17) = -1.60$ ,  $p = 0.13$ ), while the rotated condition is completely unintelligible (0%), and periodic speech is slightly less intelligible



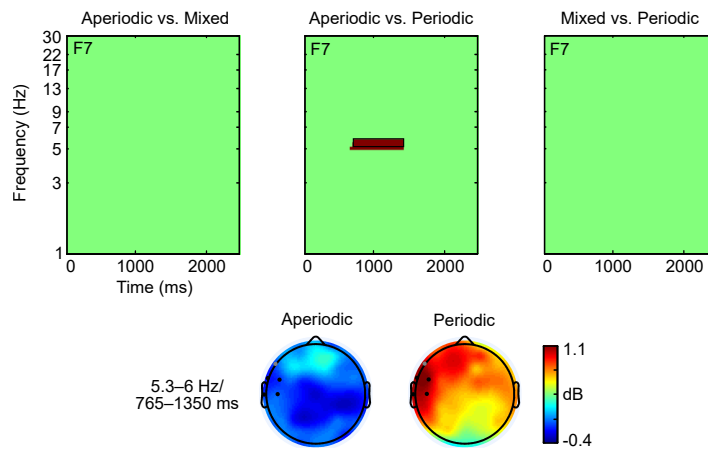
**Figure 4.3:** Periodicity. Total EEG power changes relative to baseline for the fully intelligible trials in the aperiodic, mixed, and periodic conditions. The upper part of the figure shows spectrograms of EEG activity recorded at electrode F7. In the panel on the far right, time-frequency sample points with  $p$ -values  $< 0.05$  are shown in red. The scalp distributions of the significant time-frequency window indicated by the black boxes ( $\sim 5$ – $6.3$  Hz/ $845$ – $1320$  ms) are plotted in the lower part of the figure. Electrodes that are part of the significant cluster are shown as black dots and electrode F7, which showed the strongest effect, is indicated by a grey dot.

(77.4%) than aperiodic ( $t(17) = -8.42$ ,  $p < 0.001$ ) and mixed speech ( $t(17) = -11.60$ ,  $p < 0.001$ ). Furthermore, we compared the responses to the first and the second half of the trials in the periodic condition and found no significant differences (77.8% and 77.0%;  $t(17) = 0.70$ ,  $p = 0.49$ ), indicating that there were no learning effects over the course of the 160 trials.

#### 4.3.2 Periodicity

The total EEG power changes in response to the fully intelligible trials (all five key words correctly repeated) in the aperiodic, mixed, and periodic conditions are shown in Fig. 4.3. The periodic condition was found to substantially deviate from the other two conditions, which had a very similar response pattern.

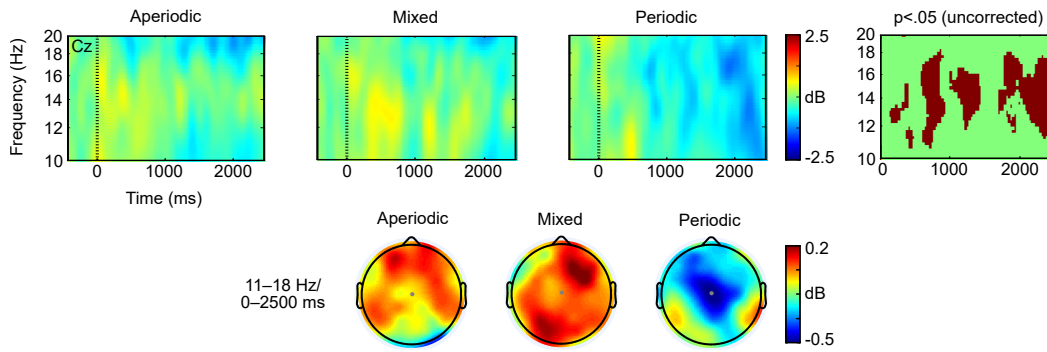
Firstly, as shown by the spectrograms in the upper part of Fig. 4.3, there was a power increase in the upper delta and theta region during the middle part of the stimulus window in the periodic condition. A cluster-based regression  $t$ -test including all three conditions confirmed that there was a linear positive relation between the amount of periodicity in the stimuli and the EEG power in this time-frequency



**Figure 4.4:** Periodicity: pairwise comparisons. Pairwise statistical comparisons of the fully intelligible trials in the aperiodic, mixed, and periodic conditions. Time-frequency sample points with  $p$ -values  $< 0.05$  are shown in red. The scalp distributions of the significant time-frequency window indicated by the black box ( $\sim 5.3$ – $6$  Hz/ $765$ – $1350$  ms) is plotted in the lower part of the figure. Electrodes that are part of the significant clusters are shown as black dots and electrode F7, which showed the strongest effect, is indicated by a grey dot.

region. This effect was most pronounced at electrode F7, but included 4 more electrodes in the left temporal scalp region (FT7, FC5, T7, and C5). These 5 electrodes all showed a consistent significant effect in time-frequency-electrode space from about 5 to 6.3 Hz and 845 to 1320 ms ( $p = 0.045$ ). This time-frequency window is indicated by the black boxes in the spectrogram plots and the corresponding scalp distributions are shown in the lower part of Fig. 4.3. Subsequent pairwise comparisons (Fig. 4.4) revealed a significant difference between the aperiodic and periodic conditions in the same time-frequency-electrode region ( $p = 0.047$ ), but no additional effects.

Secondly, there was a trend for less low beta power (11–18 Hz; see Fig. 4.5) in the periodic condition throughout the stimulus window. This observation was confirmed by a post-hoc analysis in which the whole stimulus window was statistically examined, but the frequencies range was reduced to 10 to 20 Hz. As indicated by the uncorrected  $p$ -values, this trend was strongest over the central scalp region, particularly at electrode Cz, and present throughout the stimulus window. As can be told from both the spectrograms at electrode Cz and the corresponding scalp maps, there was again hardly any difference between the aperiodic and mixed conditions.

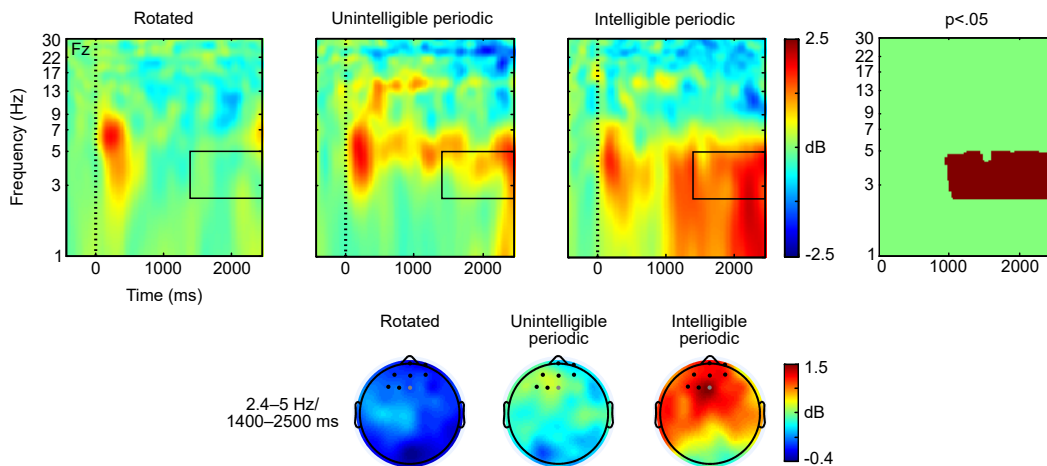


**Figure 4.5:** Periodicity: post-hoc. Total EEG power changes relative to baseline for the fully intelligible trials in the aperiodic, mixed, and periodic conditions. The upper part of the figure shows spectrograms of EEG activity recorded at electrode Cz. In the panel on the far right, time-frequency sample points with uncorrected  $p$ -values  $< 0.05$  are shown in red. The scalp distributions of the time-frequency window in which significant differences were observed (11–18 Hz/0–2500 ms) are plotted in the lower part of the figure. Electrode Cz, which showed the strongest effect, is indicated by a grey dot.

### 4.3.3 Intelligibility

The total EEG power changes for completely unintelligible rotated speech, largely unintelligible periodic speech (trials with maximally two out of five key words correct, henceforth referred to as the unintelligible periodic condition), and fully intelligible periodic speech (all five key words correct, henceforth the intelligible periodic condition) are shown in Fig. 4.6. There was a general trend for greater power changes when the speech was more intelligible. In particular, the neural response in the rotated condition was very small, apart from an initial burst of activity following the acoustic onset of the sentences. Unintelligible and intelligible periodic speech, in contrast, showed sustained activity in the theta band ( $\sim 4$ – $7$  Hz) throughout the stimulus window. Crucially, intelligible periodic speech also led to a substantial increase in delta power (1–4 Hz) during the second half of the stimulus window, which was absent in both the rotated and unintelligible periodic conditions.

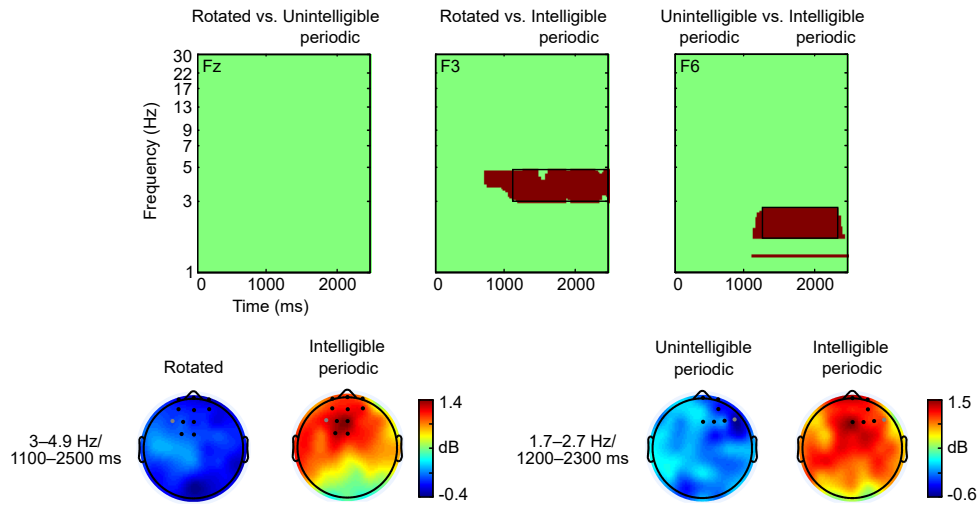
A cluster-based regression  $t$ -test including all three conditions confirmed that there was a linear positive relation between the intelligibility of the stimuli and the EEG power in the delta and theta region during the second half of the stimulus window. This effect was most pronounced at electrode Fz, but included 7 more electrodes in the frontal scalp region (Fpz, Fp2, AF3, AFz, AF4, F3, and F1). These 8



**Figure 4.6:** Intelligibility. Total EEG power changes relative to baseline for the completely unintelligible rotated condition, largely unintelligible trials in the periodic condition (maximally two out of five correctly repeated key words), and intelligible trials in the periodic condition (all five key words correctly repeated). The upper part of the figure shows spectrograms of EEG activity recorded at electrode Fz. In the panel on the far right, time-frequency sample points with  $p$ -values  $< 0.05$  are shown in red. The scalp distributions of the significant time-frequency window indicated by the black boxes ( $\sim 2.4$ – $5$  Hz/ $1400$ – $2500$  ms) are plotted in the lower part of the figure. Electrodes that are part of the significant cluster are shown as black dots and electrode Fz, which showed the strongest effect, is indicated by a grey dot.

electrodes all showed a more or less consistent significant effect in time-frequency-electrode space from about 2.4 to 5 Hz and 1400 to 2500 ms ( $p = 0.01$ ). This time-frequency window is indicated by the black boxes in the spectrograms and the corresponding scalp distributions are shown in the lower part of Fig. 4.6.

Subsequent pairwise comparisons showed that this cluster consisted of two overlapping smaller clusters (Fig. 4.7). Firstly, the direct comparison of the rotated and intelligible periodic conditions returned a cluster from about 3 to 4.9 Hz and 1100 to 2500 ms with a slightly left-lateralised frontal location ( $p = 0.01$ ). This effect was strongest at electrode F3 and in total included 11 electrodes showing a more or less consistent significant difference in time-frequency-electrode space (Fp1, Fpz, Fp2, AF3, AFz, AF4, F3, F1, Fz, FC1, and FCz). Secondly, when comparing the unintelligible and intelligible periodic conditions directly, another cluster with a slightly right-lateralised frontal distribution was obtained. This cluster had a similar temporal extension, but did not overlap in frequency with the previous one

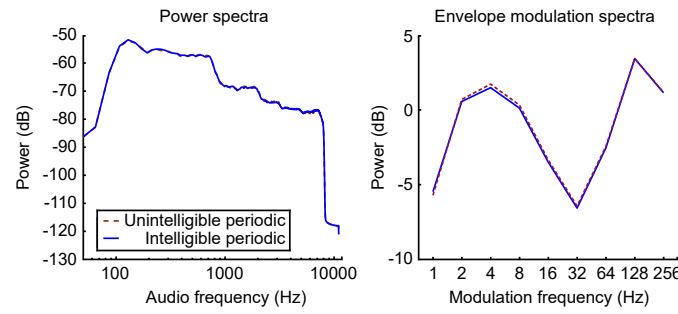


**Figure 4.7:** Intelligibility: pairwise comparisons. Pairwise statistical comparisons of the completely unintelligible rotated condition, largely unintelligible trials in the periodic condition (maximally two out of five correctly repeated key words), and intelligible trials in the periodic condition (all five key words correctly repeated). Time-frequency sample points with  $p$ -values  $< 0.05$  are shown in red. The scalp distributions of the two significant time-frequency windows indicated by the black boxes ( $\sim 3$ – $4.9$  Hz/ $1100$ – $2500$  ms and  $\sim 1.7$ – $2.7$  Hz/ $1200$ – $2300$  ms) are plotted in the lower part of the figure. Electrodes that are part of the significant clusters are shown as black dots and electrodes F3 and F6, which showed the strongest effects, are indicated by grey dots.

( $\sim 1.7$ – $2.7$  Hz/ $1200$ – $2300$  ms;  $p = 0.019$ ). Here, the strongest effect was observed at electrode F6 and in total 7 electrodes showed a consistent significant difference (Fpz, Fp2, AF4, Fz, F2, F4, and F6).

#### 4.3.4 Acoustic comparison of the unintelligible and intelligible periodic conditions

In order to test whether there were any substantial acoustic differences between the unintelligible and intelligible periodic conditions, we ran a number of additional acoustic analyses. Firstly, both the average duration of the sentences (means =  $2.05/2.04$  s, SDs =  $0.22/0.26$ , medians =  $2.04/2.01$ ,  $t(173) = 0.34$ ,  $p = 0.73$ ) and the average F0 frequencies of the concatenated sentences in the two conditions (means =  $119.33/119.01$  Hz, SDs =  $25.42/24.20$ , medians =  $115.46/116.18$ ,  $t(173) = -0.42$ ,  $p = 0.68$ ) were found to show little difference. The F0 frequencies were initially extracted with a sampling rate of 100 Hz, but down-sampled to 0.56 Hz for statisti-



**Figure 4.8:** Acoustic characteristics of the largely unintelligible (maximally two out of five correctly repeated key words) and intelligible (all five key words correctly repeated) trials in the periodic condition. The left panel shows the average power spectra, i.e. the stimulus power plotted as a function of audio frequency, and the right panel the average envelope modulation spectra, i.e. the stimulus power plotted as a function of envelope modulation frequency.

cal testing to obtain the same degrees of freedom as for the comparison of sentence duration. Secondly, we compared the power spectra of the two concatenated sets of sentences (computed using Welch’s method, FFT size = 1024 samples, sampling rate = 22.05 kHz). The left panel of Fig. 4.8 shows that the spectra are virtually identical, which is underlined by a very high Pearson’s correlation coefficient of the frequency bins closest to the centre frequencies of the eight vocoder bands ( $r = 0.99$ ,  $p < 0.001$ ). Lastly, we computed the average modulation spectra of the amplitude envelopes of all the sentence in the two conditions using the front end of the mr-sEPSM speech intelligibility model (Jørgensen et al., 2013). The modulation spectra were averaged over all 22 gammatone audio filters of the model, resulting in the simple line plot shown in the right panel of Fig. 4.8. The high correlation coefficient of the nine modulation filter centre frequencies across conditions ( $r = 0.99$ ,  $p < 0.001$ ) again confirms that there is little acoustic difference between the two conditions.

#### 4.4 Discussion

The present study sought to identify effects of acoustic periodicity and intelligibility in the EEG time-frequency responses to acoustically presented sentences. We thereby attempted to overcome the limitation that acoustic factors and intelligibility have not been examined independently in previous studies. Firstly, it was found

that despite considerable acoustic differences, the total responses in the aperiodic and mixed conditions were almost identical. In contrast, the total EEG power in the periodic condition differed substantially from the other two conditions, even after controlling for the lower intelligibility. Differences were observed in the theta and low beta bands. Secondly, completely unintelligible rotated speech and largely unintelligible as well as intelligible periodic speech we compared. Here, we observed hardly any power changes in the rotated condition, apart from the acoustic onset response, but a substantial increase in delta power during the second half of the intelligible periodic sentences, when compared to their acoustically similar unintelligible counterparts.

#### 4.4.1 Periodicity

The increase in theta power in the periodic condition agrees well with the results of [Strauß et al. \(2014a\)](#), who found significantly more theta activity in a left-lateralised fronto-temporal network for ambiguous pseudo-words. Our results thus corroborate the idea that increased theta activity in this region is an indicator of response conflicts in auditory speech tasks. As suggested by [Roux and Uhlhaas \(2014\)](#), enhanced theta power in the context of speech tasks may indicate that verbal information is kept in the phonological loop, where the materials are sub-vocally rehearsed. In line with this idea, the power decrease in the low beta range in the periodic condition seems to be a mu rhythm de-synchronisation, often observed before imagined or real movements (e.g. [Cohen and Donner, 2013](#); [Pfurtscheller et al., 1997](#); [Wisniewski et al., 2015](#)). It thus appears that both effects that distinguish the periodic speech from the other two conditions stand in relation to each other. Importantly, the participants correctly repeated every stimulus sentence in each of the three conditions included in the current analysis. Hence, the effect cannot be a result of motor preparation per se, but must be due to specific processes associated with the periodic condition.

Previous studies have reported that more intelligible words lead to a greater suppression of alpha activity ([Becker et al., 2013](#); [Obleser and Weisz, 2012](#)). In line with these findings, we have observed no differences in the alpha range between the aperiodic, mixed, and periodic conditions after controlling for differences in



intelligibility. On the other hand, we did also not observe a suppression of alpha activity. The absence of this effect is in fact the only notable difference between the EEG spectrogram of the 8-channel aperiodic (i.e. noise-vocoded) condition in the study of [Obleser and Weisz \(2012\)](#) and the one in the current study. This might be due to the fact that we have used the relatively long and difficult IEEE sentences, and not single words. Although the alpha power level did not appear to decrease towards the onset of the sentences (see Fig. 4.3), it may have been lowered throughout the pre-stimulus window, indicating a state of ‘anticipatory attention’ preceding a demanding task ([Klimesch, 2012](#)). However, since we presented the conditions in blocks of ten sentences (i.e. whole IEEE sentence lists), subjects could also form expectancies regarding the upcoming stimulus, which may have caused the alpha power level to remain relatively stable between the baseline and stimulus windows.

In summary, the very similar EEG responses in the aperiodic and mixed conditions suggests the existence of a default response pattern to speech signals that are relatively easy to understand. A deviation from this pattern, as in the case of the periodic condition, in turn may indicate that a speech signal sounds unnatural and interferes with normal processing.

#### 4.4.2 Intelligibility

Spectrally rotated speech was introduced to neuroscience in an attempt to provide an adequate non-speech analogue for intelligible speech ([Rosen and Iverson, 2007](#); [Scott et al., 2000](#)). However, despite the speech-like acoustic properties, we did not observe any substantial neural activity in the rotated condition, apart from the initial acoustic onset response. The largest part of the signal thus resembled a recording of silence. This suggests that the attempt to mimic the acoustic properties of speech in an unintelligible control condition may in fact be needless, at least in M/EEG studies. In contrast, the neural response in the unintelligible periodic condition, which included trials with up to two correctly repeated key words out of five, resembled its intelligible counterpart much more and additionally provides the benefit of being acoustically more similar. In particular, only the two periodic conditions showed activity in the theta band throughout the stimulus window.

The main finding when comparing the total EEG power in the three conditions was the pronounced increase in delta power in the intelligible periodic condition. Given that there is no acoustic event that could have triggered this effect, it shows a surprisingly sharp onset at around 1000 ms after sentence onset and was present throughout the remainder of the stimulus window. As the average duration of the sentences was 2.04 s, the onset of this effect coincides with the beginning of the second half of the sentences. In line with the rule that the lower the frequency of a neural oscillation, the wider its distribution (e.g. [Buzsáki and Draguhn, 2004](#)), this effect was observed at the vast majority of electrodes, but only reached statistical significance in the frontal scalp region. Although the phase of delta oscillations has been shown to entrain to tone sequences ([Lakatos et al., 2005](#)), and the detection of small loudness differences of tones has been reported to depend on delta power ([Herrmann et al., 2016](#)), the power of delta oscillations has so far not been associated with speech intelligibility. Importantly, delta power increases towards the end of the stimulus window were also observed in all three conditions in the analysis of periodicity (see Fig. 4.3), which demonstrates that this effect is not confined to the unnatural sounding periodic condition. Clearly, further research is needed to explore the exact relation between delta oscillations and speech intelligibility, particularly its time course. It is noteworthy that the average word duration of clearly articulated English is about 400 ms (i.e. 2.5 Hz; [Hazan and Baker, 2011](#)), and thus falls right into the middle of the delta band. However, since the increase in delta power was only observed during the second half of the intelligible sentences, it does not appear to be associated with the intelligibility of the individual words. Instead, one may speculate that this effect reflects the understanding of the meaning of the sentences as a whole.

Finally, we did again not observe any significant differences in the alpha band, although Fig. 4.5 shows that alpha power after sentence onset is slightly increased in the unintelligible periodic condition. This trend resembles the finding that alpha power is enhanced when speech signals are embedded in background noise, which was suggested to reflect the attempt to cope with demanding listening conditions

(e.g. [Strauß et al., 2014b](#); [Wilsch et al., 2015](#)). However, it appears that the target speech needs to be both difficult to understand, as is the case for the unintelligible periodic condition, and be presented in noise in order to lead to pronounced alpha power changes. Our data hence suggest that for speech presented in quiet, there is no strong association between alpha power and speech intelligibility, after controlling for acoustic differences.

#### **4.5 Conclusion**

By manipulating the amount of source periodicity in the materials, the present study has shown that total EEG power changes in response to speech do not reflect acoustic stimulus properties as such, but the perceptual effects of these properties. Even after controlling for differences in intelligibility, responses to fully periodic speech, an artificial condition that makes it difficult to identify the individual speech sounds, deviated markedly from the two other conditions with an entirely aperiodic or mixed source excitation. The neural responses in the latter two conditions, on the other hand, were very similar, despite their acoustic differences. In a second analysis, EEG power changes to unintelligible and intelligible speech were compared. Firstly, the very sparse neural response to spectrally-rotated speech casts strong doubts on whether it is a suitable unintelligible control condition in M/EEG studies. Secondly, the direct comparison of the unintelligible and intelligible trials in the periodic condition revealed an increase in delta power during the second half of the sentences. The current results thus suggest that delta oscillations are a possible neural correlate of successful speech understanding.

## Chapter 5

### The role of periodicity in perceiving speech in background noise with simulated cochlear implants

#### 5.1 Introduction

Previously we ([Steinmetzger and Rosen, 2015](#); i.e. chapter 2) have investigated the ability of normal-hearing (NH) listeners to perceive sentences in a variety of conditions involving the presence and absence of periodicity in both target speech and masker. Listeners were found to substantially benefit from periodicity in the masker, while there was little effect of periodicity in the target speech. The periodic maskers used were harmonic complexes with dynamically varying F0-contours derived from real speech. Moreover, the benefit from masker periodicity was substantially larger than the fluctuating-masker benefit (FMB) obtained from sinusoidal 10 Hz modulations of the masker amplitude envelope. Here, we used a similar design and tested to what extent masker-periodicity benefit (MPB) and FMB were maintained in simulations of cochlear implants (CIs).

Factors that are thought to explain the MPB include the ability to use F0 cues to cancel out harmonic maskers from the signal mixture ([de Cheveigné et al., 1995](#); [de Cheveigné et al., 1997b](#)), the possibility to glimpse sections of the target speech in between the individual masker harmonics ([Deroche et al., 2014a, 2014b](#)), the absence of random envelope modulations in periodic sounds ([Stone et al., 2011](#); [Stone et al., 2012](#)), and the fact that the modulations arising in steady-state periodic sounds in the voice F0 range are not in the low-frequency range crucial for speech intelligibility ([Elliott and Theunissen, 2009](#)). However, the exact contribution of each of these factors remains to be specified.

As the access to spectral information and F0 cues is known to be severely restricted with current CIs (e.g. [Fu and Nogaki, 2005](#); [Green et al., 2002](#); [Wilson and Dorman, 2008](#)), listeners were not assumed to be able to cancel out the periodic maskers from the signal mixture based on their harmonicity. Similarly, they were not expected to spectrally glimpse portions of the target speech signal in between the individual masker harmonics. Although the F0-related envelope modulations of periodic maskers will to some extent be preserved after CI simulation processing, they are most prominent in higher auditory filters where the harmonics are not resolved. On the other hand, these preserved F0-related modulations will still lead to a greater temporal regularity of the periodic maskers, which may help to fuse them together into a single auditory stream, thereby making it easier to segregate them from the target speech.

CI users have consistently been found to show hardly any benefit from masker envelope fluctuations ([Cullington and Zeng, 2008](#); [P. B. Nelson and Jin, 2004](#)), or even a decline in performance ([P. B. Nelson et al., 2003](#); [Stickney et al., 2004](#)), while there tends to be a small FMB in CI simulation studies ([Cullington and Zeng, 2008](#); [P. B. Nelson and Jin, 2004](#); [Qin and Oxenham, 2003](#)). The absence of an FMB for CI users has been attributed to several factors, including reduced spectral cues ([Fu et al., 1998](#)), increased forward masking ([D. A. Nelson and Donaldson, 2001](#)), and the limited access to F0 information ([Stickney et al., 2007](#); [Stickney et al., 2004](#)). At least in part, it can also be explained by the elevated speech reception thresholds (SRTs) compared to NH listeners ([Bernstein and Grant, 2009](#)), as the FMB is generally larger at lower signal-to-noise ratios (SNRs; [Freyman et al., 2012](#)). Importantly, the same pattern has also been found for the MPB (see Fig. 2.6).

In all previously mentioned studies concerned with FMBs, target and masker envelope varied independently of each other. [Kwon and colleagues \(2012\)](#), in contrast, introduced a set of maskers that maximise (+MR) or minimise (-MR) opportunities to glimpse portions of the target speech by altering the temporal overlap of signal and masker. Masker envelopes were either the inverse (+MR) of the speech envelope or the same (-MR), adjusted in 50 ms steps. The current study included

the +MR maskers in addition to the previously used steady and 10-Hz modulated maskers, with the intention to parametrically vary the amount of energetic masking (steady > 10-Hz modulated > +MR). Contrary to what would be expected in the absence of energetic masking, only the few CI users in [Kwon et al. \(2012\)](#) whose sentence intelligibility in quiet was above 90% showed substantial FMBs when tested with the +MR maskers. The authors concluded that CI users may find it particularly difficult to identify the segmental boundaries between speech and noise. The current experiment aimed to test whether this finding similarly applies to CI simulations and maskers that are not based on speech-shaped noise.

## **5.2 Methods**

### **5.2.1 Participants**

Eleven normal-hearing listeners (6 females) were tested. Their ages ranged from 18–21 years, with a mean of 19.5. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB hearing level (HL) at frequencies between 125 and 8000 Hz. All subjects gave written consent and the study was approved by the UCL ethics committee.

### **5.2.2 Stimuli**

The target speech materials used in this experiment were recordings of the Basic English Lexicon sentences (BEL; [Calandruccio and Smiljanic, 2012](#)) spoken by an adult male Southern British English talker that were normalised to a common root-mean-square (RMS) level. The sentences were slightly modified for appropriate British vocabulary and usage. The BEL sentence corpus consists of 20 lists with 25 sentences each and is characterised by a simple syntactic sentence structure and the use of basic non-native English vocabulary. Every sentence contains four key words.

The masker materials were the same as the ones used in [Steinmetzger and Rosen \(2015\)](#). Harmonic complex maskers were based on F0 contours extracted from recordings in the EUROM database of English speech in which different speakers read 5- to 6-sentence passages ([Chan et al., 1995](#)). Sixteen different male

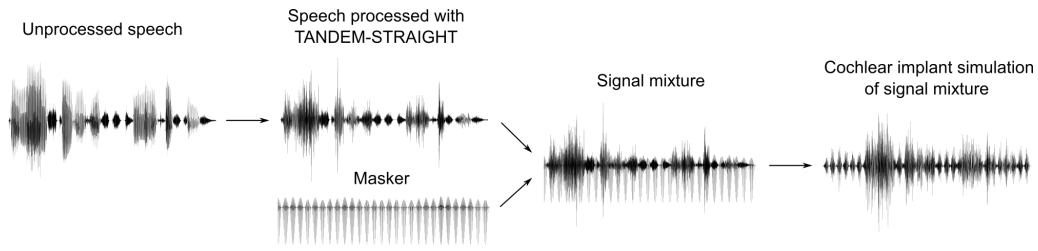
talkers with Southern British English accents, and a similar speaking rate and voice quality to that of the target talker were chosen. The noise maskers were based on a 24-second passage of white noise.

### 5.2.3 **Signal processing**

Three target speech conditions with different amounts of source periodicity were produced prior to the experiment using TANDEM-STRAIGHT ([Kawahara et al., 2008](#)) implemented in MATLAB (Mathworks, Natick, MA). TANDEM-STRAIGHT is a type of vocoder that, unlike a channel vocoder, does not filter the input speech into separate frequency bands but separates the periodic and aperiodic components of the source from the spectral filter. By default, TANDEM-STRAIGHT produces speech with a mixed source excitation that sounds very natural, but the source estimation procedure can be adapted to produce fully aperiodic or fully periodic speech as well.

Aperiodic speech was synthesised by keeping the default settings of TANDEM-STRAIGHT, but setting the F0 to 0 Hz throughout. In order to synthesise speech with a natural mix of periodicity and aperiodicity, the default settings were kept, but the values of the sigmoid parameter in the source estimation routine were fixed to 1 and -40, in order to minimise the level of the aperiodic component in voiced speech segments. This avoids higher harmonics being noisier than lower ones, as is the case in natural speech, and hence emphasises the contrast of voiced and unvoiced speech. The same technique was used to produce fully periodic speech, but here interpolated F0 contours were used as input for the source extraction routine. These interpolated F0 contours were produced by first extracting the original F0 contours using ProsodyPro ([Xu, 2013](#)) implemented in Praat ([Boersma and Weenink, 2013](#)). The F0 extraction sampling rate was set to 100 Hz. Secondly, F0 contours were interpolated through unvoiced sections and periods of silence, using a piecewise cubic Hermite interpolation in logarithmic frequency. The start and end points of each contour were anchored to the median frequency of the sentence.

The same interpolation procedure was used to obtain the F0 contours for the harmonic complex maskers. The waveforms for these maskers were synthesised



**Figure 5.1:** Schematic depiction of the signal processing. Unprocessed target speech was first processed to have an aperiodic, mixed, or periodic source excitation using TANDEM-STRAIGHT. Next, a masker was added to the processed target speech signal at a given signal-to-noise ratio and the signal mixture was then additionally noise-vocoded to yield the final stimulus.

on a period-by-period basis using the Liljencrants-Fant model (Fant et al., 1985), which closely approximates a typical adult male glottal pulse [see Green and Rosen (2013) for details]. Both the harmonic complexes and the noise maskers were matched in spectrum to the long-term average of the targets (LTASS) with finite impulse response filter [Greenwood filter spacing, 1-octave smoothing, filter order 1024, fast Fourier transform (fft) with a window size of 512 samples]. The LTASS of the unprocessed target speech was determined by computing the power spectral density of the concatenated waveforms using Welch’s method (window size 512 samples, 50% overlap, fft length 512 samples). The resulting spectrum was then smoothed over one octave.

Masker envelopes were either steady, sinusoidally amplitude-modulated at a rate of 10 Hz with a modulation depth of 100%, or adjusted to be the inverse of the target sentence envelope in 50 ms steps (+MR; Kwon et al., 2012). As in the paper by Kwon and colleagues (2012), the level of the +MR masker envelope was restricted to vary between -50 and -10 dB relative full scale to generate a noise floor and to avoid clipping, respectively. The stimulus sentences were tightly cut, so that the inverse envelopes were only constructed during the actual sentences and not the silent periods before and after. For the additional portions of the masker inserted before and after the stimulus sentences, the resulting inverse envelopes were then simply extended at the RMS levels where they started and stopped.

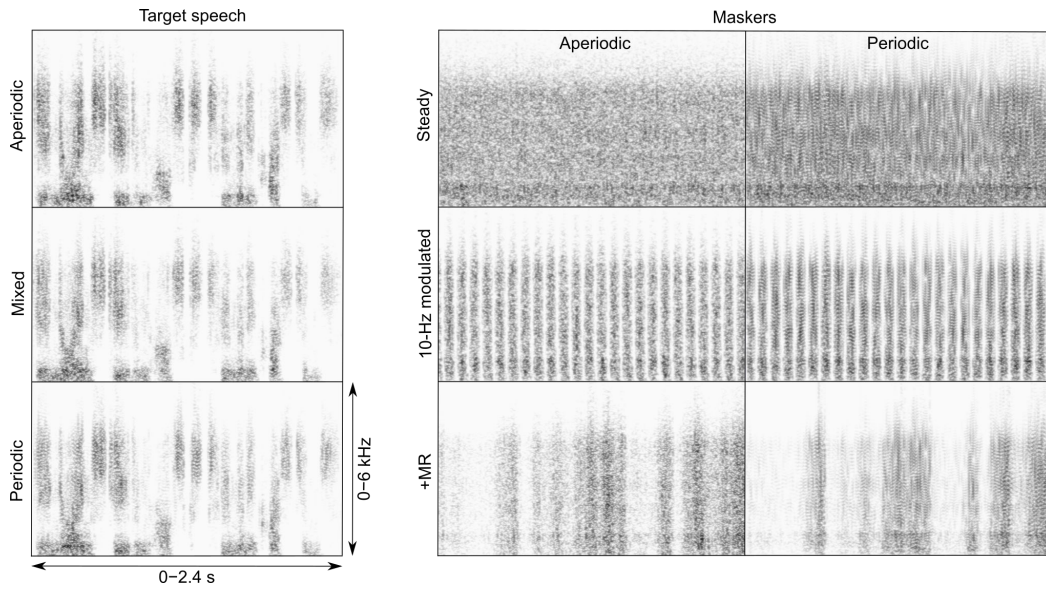


Next, target speech and masker were added together. The masker level was kept constant and the speech level was adjusted to achieve a particular SNR. In order to simulate CI processing, the signal mixture was then additionally noise-vocoded on the fly before each trial, using a channel vocoder implemented in MATLAB. The mixture of target sentence plus masker was first band-pass filtered into eight bands using third-order Butterworth filters. The filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of 70 Hz to 4 kHz. The output of each filter was full-wave rectified and low-pass filtered at 400 Hz (second-order Butterworth) in order to extract the amplitude envelope. The high cut-off value was chosen in order to ensure that temporal periodicity cues were preserved. The envelope from each band was then multiplied with a wide-band noise carrier and the resulting signals were again band-pass filtered using the same third-order Butterworth filters as in the first stage of the process. Finally, before summing the individual bands together, the output of each band was adjusted to the same RMS level as found in the original recording. A schematic depiction of the complete signal processing pipeline is shown in Fig. 5.1 and examples of the resulting stimuli after CI simulation processing are shown in Fig. 5.2.

#### 5.2.4 Procedure

Participants were presented with 1 BEL sentence list in each of the 18 conditions (3 targets x 6 maskers). The SRT for every processing condition was determined by tracking the SNR necessary in order to repeat 50% of the key words in a sentence correctly. The initial SNR was set to +10 dB and adjusted up or down by 11 dB before the first reversal, 7 dB before the second reversal, and 3 dB after that. If the subject got less than half of the key words correct in the first trial, the SNR was set to +24 dB and the procedure started over again. The SRT was calculated by taking the mean of the largest even number of reversals with 3-dB step size.

The verbal responses were logged by the experimenter before the next sentence was played. A so-called loose key words scoring technique was applied, in which the roots of the four key words had to be correctly identified. No feedback was given following the responses. The presentation and logging of the responses

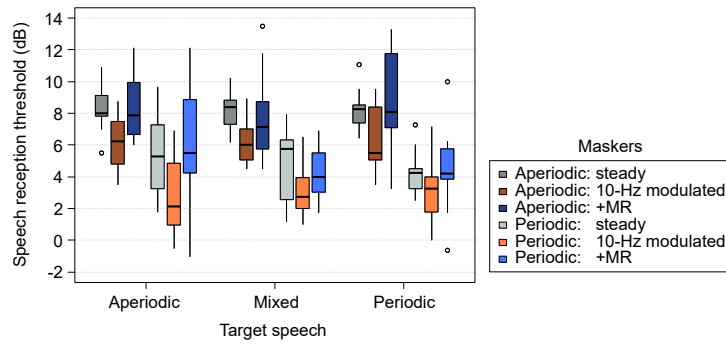


**Figure 5.2:** Stimuli. The left panel shows narrow-band spectrograms for one example sentence (The annoying student asks too many questions.) processed to have an aperiodic, mixed, or periodic source excitation. The right panel shows narrow-band spectrograms of the six different maskers. Maskers sources were either aperiodic or periodic and masker envelopes were either steady, 10-Hz modulated, or the inverse of the target speech (+MR). The +MR masker example is tailored to the example sentence shown on the left. All stimuli are shown after cochlear implant simulation processing (by noise vocoding).

was carried out using locally developed MATLAB software. The order of the 18 processing conditions was fully randomised using a Latin Square design and the order of the BEL lists was also randomised.

For each trial of the experiment, a random portion of the maskers was picked and presented along with the target sentence. For the harmonic complexes, the order of the talkers was also randomised, ensuring that all 16 of them were picked before any of them was repeated. The onset of all the maskers was 600 ms before that of the targets and they continued for another 100 ms after the end of the target sentence. An onset and offset ramp of 100 ms was applied to the mixture of target and masker.

Before being tested, the subjects were familiarised with the materials by listening to 4 example sentences of each of the three target speech conditions in quiet and one example sentence of each of the 18 conditions to be used in the main experiment at an SNR of +10 dB. The first BEL sentence list was reserved for the



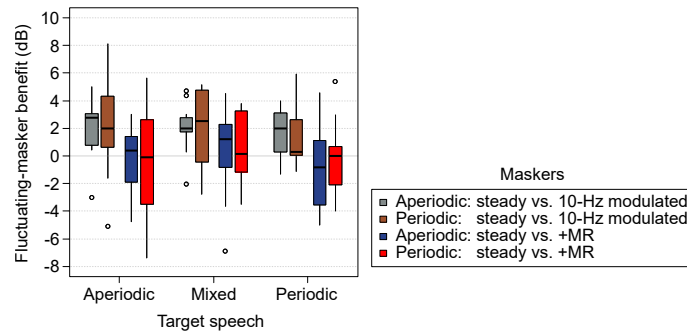
**Figure 5.3:** Boxplots of the speech reception thresholds. Each of the three target speech conditions on the x-axis was tested in combination with the six different maskers shown in the legend. The black horizontal lines in the boxplots indicate the median value.

familiarisation procedure and not used in the main experiment. The total duration of the experiment, including hearing screening and familiarisation procedure, was about 45 minutes and subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating booth, with the computer signal being fed through the wall onto a separate monitor. The stimuli were converted with 24-bit resolution and a sampling rate of 22.05 kHz using an RME Babyface soundcard (Haimhausen, Germany) and presented over Sennheiser HD650 headphones (Wedemark, Germany). The level of the target and masker mixture was set to about 70 dB SPL over a frequency range of 70 Hz to 4 kHz, as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

### 5.3 Results

Fig. 5.3 shows the SRTs obtained in each of the 18 processing conditions, which were positive throughout. Overall, listeners benefitted substantially from masker periodicity and to a lesser extent also from sinusoidal masker modulations. Performance with the +MR maskers, however, was similar to that with steady maskers.

The data were analysed using a mixed effects model with target periodicity (aperiodic, mixed, periodic), masker periodicity (aperiodic, periodic), and masker envelope (steady, 10-Hz modulated, +MR) as fixed factors, and subjects and sentence lists as random factors. The main effects of masker periodicity

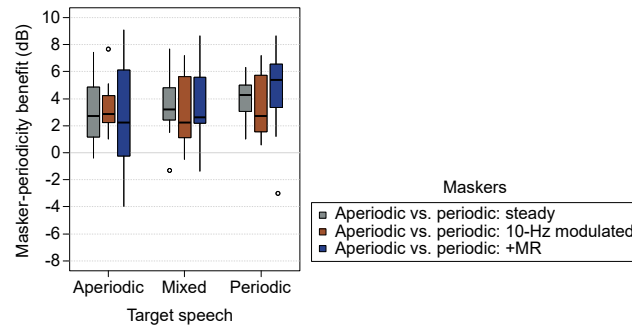


**Figure 5.4:** Boxplots of the fluctuating-masker benefits. For each of the three target speech conditions on the x-axis, the difference between the steady and amplitude-modulated, as well as the steady and +MR version of the noise and harmonic complex maskers is plotted. Positive numbers on the y-axis indicate a benefit. The black horizontal lines in the boxplots indicate the median value.

[ $F(1,180) = 128.1, p < 0.001$ ] and masker envelope [ $F(2,180) = 19.5, p < 0.001$ ] were both highly significant, but neither the main effect of target periodicity [ $F(2,180) = 0.76, p = 0.47$ ], nor any of the interactions reached significance ( $F \leq 0.74, p \geq 0.48$ ).

In Fig. 5.4, the same data are re-plotted as FMBs, i.e. the difference in SRT of a steady compared to a fluctuating masker. Positive FMBs indicate that listeners were on average able to benefit from masker envelope fluctuations. Overall, FMBs were rather small, with the largest effects of about 2 dB observed for the 10-Hz modulated noise maskers. Same as for normal-hearing listeners (see chapter 2), the FMB was smaller for harmonic complex maskers across conditions. The +MR maskers did not enable any FMB across all six combinations of target speech and masker.

In Fig. 5.5 the same data are again re-plotted as MPBs, i.e. the difference in SRT between noise and harmonic complex maskers. The MPB was generally larger than the FMB, with the largest effect of about 5 dB observed for the combination of periodic targets and periodic +MR masker. In addition, for the steady and +MR maskers, there was a trend for larger MPBs with more periodicity in the targets. However, when analysing the SRT data, this threefold interaction of target periodicity, masker periodicity, and masker envelopes was far from significant [ $F(4,180) = 0.18, p = 0.95$ ].



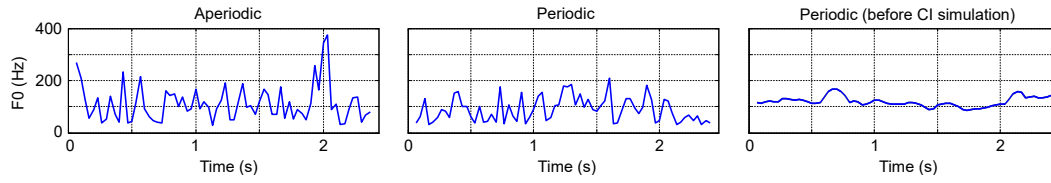
**Figure 5.5:** Boxplots of the masker-periodicity benefits. For each of the three target speech conditions on the x-axis, the difference between the noise and harmonic complex version of the steady and amplitude-modulated, as well as the steady and +MR maskers is plotted. Positive numbers on the y-axis indicate a benefit. The black horizontal lines in the boxplots indicate the median value.

In summary, there was no effect of target periodicity, while FMB and MPB were reduced to about half their size with simulated CIs, compared to normal hearing (on average from about 4 to 2 dB and 9 to 4 dB, respectively; see chapter 2). However, even in the absence of salient pitch cues, the MPB was still twice as large as the FMB. In addition, although they did not energetically mask the targets, the +MR maskers with their speech-like envelopes led to similar SRTs as steady interferers.

## 5.4 Discussion

### 5.4.1 Periodicity

As for normal hearing (Steinmetzger and Rosen, 2015; i.e. chapter 2), the amount of target periodicity had little effect on speech recognition performance. On the other hand, in line with our hypothesis, the MPB was to some extent preserved after CI simulation processing. Due to the spectral smearing introduced by noise-vocoding the signal mixture with eight channels, access to F0 cues was very limited, which precludes an explanation based on harmonic cancellation (de Cheveigné et al., 1995; de Cheveigné et al., 1997b). In order to demonstrate this, F0 contours of examples of steady aperiodic and periodic maskers were computed using YIN (de Cheveigné and Kawahara, 2002; default settings, but F0 sampling rate set to 30 Hz and F0 search range restricted to 30–400 Hz). Fig. 5.6 shows that the F0 contours of both maskers fluctuate randomly. In contrast, before CI simulation processing,

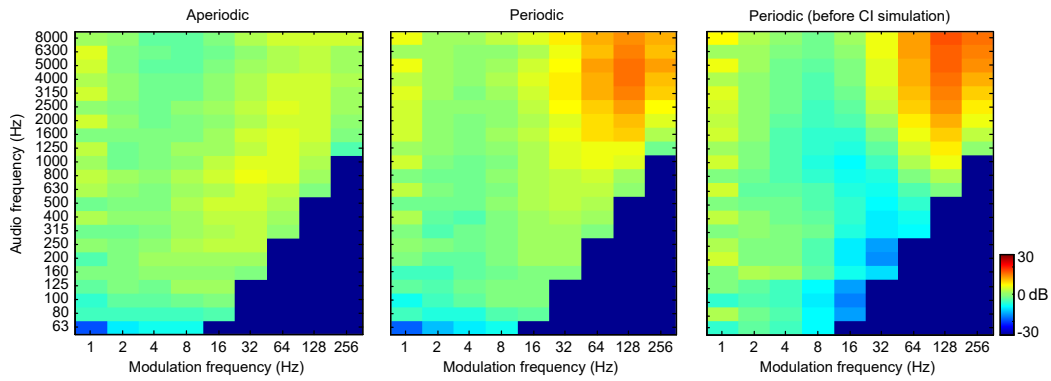


**Figure 5.6:** F0 contours of aperiodic and periodic maskers. The steady aperiodic and periodic maskers shown after cochlear implant (CI) simulation in Fig. 5.2 were used as examples. For the purpose of comparison, the same steady periodic masker is also shown before CI simulation processing. The F0 contour was computed using YIN.

the same periodic masker has a well-defined F0 contour, which does not resemble the one of the periodic masker after CI simulation.

As demonstrated by [Stone and colleagues \(2011; 2012\)](#), the primary reason for the effectiveness of noise maskers arises from random envelope fluctuations, rather than pure energetic masking of the speech signal. In order to estimate the amount of modulation masking, the modulation spectrograms of the maskers used in Fig. 5.6 were computed using the front end of the mr-sEPSM speech intelligibility model ([Jørgensen et al., 2013](#)). Fig. 5.7 shows that after CI simulation the modulation pattern is diffuse for both the aperiodic and periodic maskers, as is typical for noise maskers. Before CI simulation, in contrast, the modulations of the periodic masker are confined to the highest and lowest modulation frequencies, corresponding to the F0-related envelope fluctuations and the variations of F0, respectively.

Crucially, after CI simulation a substantial portion of the F0-related high frequency modulations of the periodic masker is still present. These additional modulations are hypothesised to have helped the listeners to perceive the periodic maskers as a single auditory stream and thereby distinguish them from the target speech. However, as mentioned in the introduction, the preserved F0-related modulations are confined to higher auditory filters, in which the individual harmonics are not resolved. The better performance with periodic maskers hence cannot be explained by the possibility to spectrally glimpse sections of the speech signal in between the individual masker harmonics ([Deroche et al., 2014a, 2014b](#)).



**Figure 5.7:** Modulation spectrograms of aperiodic and periodic maskers. The maskers shown are the same as in Fig. 5.6. The modulation spectrograms were computed using the front end of the mr-sEPSM speech intelligibility model.

#### 5.4.2 Masker fluctuations

In line with the CI data of [Kwon and colleagues \(2012\)](#), listeners in the current study did on average not show an FMB with the +MR maskers. As energetic masking is presumed to be minimal with these maskers, CI simulation processing appears to make it particularly difficult to distinguish segments of target speech and masker. This may in large part be because spectral cues that aid stream segregation are mostly unavailable. However, it has also been shown that in CI simulations listeners have problems fusing auditory information across temporal gaps, even in the absence of background noise ([P. B. Nelson and Jin, 2004](#)). In this study, subjects were presented with sentences interrupted by periods of silence and recognition performance was severely impaired across gating frequencies, which ranged from 1 to 32 Hz. Similar results have also been obtained by [Ardoint et al. \(2014\)](#), who have shown that 5-Hz interruptions affect the intelligibility of vocoded speech much more than that of unprocessed speech. Additionally, whereas the envelopes of the 10-Hz sinusoidally modulated maskers fluctuate periodically, the envelope modulations of the +MR maskers are aperiodic, which may make it harder to identify them as a non-speech signal. More specifically, the listener is confronted with an inverted copy of the target speech envelope itself, and both signals together will then combine to form a continuous speech-like envelope.

## 5.5 Conclusion

In line with the results obtained from normal-hearing listeners, the amount of periodicity in the target speech signal hardly affected speech recognition performance. However, although the perception of spectral information is severely restricted by CI simulation signal processing, masker periodicity was found to aid speech perception to a greater extent than superimposed sinusoidal masker envelope fluctuations at a rate of 10 Hz. On the other hand, the attempt to improve FMBs by tailoring masker envelopes to be the inverse of the target speech envelopes resulted in performance rates as poor as with steady interferers. This suggests that in addition to the greater susceptibility for energetic masking, CI simulation processing also makes it harder to perceive the segmental boundaries between competing signals.



## **Chapter 6**

### **General Discussion**

#### **6.1 Summary of the main results and their implications**

The aim of this thesis was to investigate the role of periodicity (i.e. voicing) in the perception of speech, a crucial acoustic feature of speech sounds across languages, whose role has not been systematically examined yet. In a series of experiments, behavioural and electrophysiological data were obtained to test how the presence or absence of periodicity affects the perception of speech in quiet, whether periodicity aids the perceptual segregation of target speech and background noise, and how periodicity is represented in cortical electroencephalography (EEG) signals recorded in response to speech. With the exception of the maskers based on white noise, all the materials in the present thesis were derived from recordings of real speech in an attempt to make them as realistic as possible, which distinguishes the current body of work from earlier attempts to investigate the issues in question. Additionally, the speech signals introduced in the first part of this thesis were employed to further investigate the neural correlates of speech intelligibility. Here, the objective was to ensure a comparison of intelligible and unintelligible speech that is free of any acoustic confounds.

In chapter 2, the ability of normal-hearing listeners to perceive sentences in quiet and in background noise was investigated in a variety of conditions mixing presence and absence of periodicity in both target and masker. Experiment 1 showed that in quiet, aperiodic noise-vocoded speech and speech with a natural amount of periodicity were equally intelligible, while fully periodic speech was much harder to understand. In experiments 2 and 3, speech reception thresholds

for these targets were measured in the presence of four different maskers: speech-shaped noise, harmonic complexes with a dynamically varying F0 contour, and 10 Hz amplitude-modulated versions of both. For experiment 2, results of experiment 1 were used to identify conditions with equal intelligibility in quiet, while in experiment 3 target intelligibility in quiet was near ceiling. In the presence of a masker, periodicity in the target speech mattered little, but listeners strongly benefitted from periodicity in the masker. Substantial fluctuating-masker benefits required the target speech to be almost perfectly intelligible in quiet. In summary, these results suggest that the ability to exploit periodicity cues may be an even more important factor when attempting to understand speech embedded in noise than the ability to benefit from masker fluctuations.

Chapter 3 investigated EEG signals in response to acoustically degraded speech with more or less periodicity. Here, unambiguously interpreting the results is complicated by the fact that speech signal manipulations affect acoustic and intelligibility alike. In the current study, the acoustic properties of the stimuli were thus altered and the trials were sorted according to the correctness of the listeners' spoken responses to separate out these two factors. Firstly, more periodicity rendered the event-related potentials (ERPs) more negative during the first second after sentence onset, indicating a greater cortical sensitivity to auditory input with a pitch. Secondly, a larger contingent negative variation (CNV) was observed in the ERP during sentence presentation when the subjects could subsequently repeat more words correctly. Additionally, slow alpha power (7–10 Hz) before sentences with the least correctly repeated words was increased, which may indicate that subjects have not been focussed on the upcoming task. These results suggest that acoustic periodicity is a factor that should not be overlooked when investigating the neural correlates of speech perception and the cognitive processes involved. In particular, it appears that aperiodic noise-vocoded speech leads to diminished evoked cortical responses, suggesting that speech materials preserving the natural mix periodic and aperiodic segments are the better choice.

In chapter 4, the same EEG data were analysed in the frequency domain. Although several studies have investigated neural oscillations in response to acoustically degraded speech, it is still a matter of debate which neural frequencies reflect speech intelligibility. It was found, firstly, that the total EEG power changes in response to completely aperiodic (noise-vocoded) speech and speech with a natural mix of periodicity and aperiodicity were almost identical, while an increase in theta power (5–6.3 Hz) and a trend for less beta power (11–18 Hz) were observed in response to completely periodic speech. These two effects are taken to indicate an information processing conflict caused by the unnatural acoustic properties of the stimuli, and that the subjects may have internally rehearsed the sentences as a result of this. Secondly, we separately investigated effects of intelligibility by sorting the trials in the periodic condition according to the listeners' spoken responses. The comparison of intelligible and largely unintelligible trials revealed that the total EEG power in the delta band (1.7–2.7 Hz) was markedly increased during the second half of the intelligible trials, which suggests that delta oscillations are an indicator of successful speech understanding. Although increased delta power in the frontal cortex has been associated with a state of concentration, no similar effect has so far been reported in the speech perception literature.

Chapter 5 investigated the role of periodicity in perceiving speech in noise after simulated cochlear implant (CI) signal processing. The materials used were similar to those introduced in chapter 2, but the mixture of speech and noise was in addition acoustically degraded to mimic the signals transmitted by a typical CI. As current CIs provide very restricted access to spectral information, this study tested whether temporal cues are sufficient to benefit from periodicity when listening to speech embedded in background noise. Furthermore, this experiment included maskers that promoted glimpsing by minimising the energetic overlap with the target speech, in order to test if it is at all possible to benefit from masker envelope fluctuations in CI-like listening conditions. Same as in chapter 2, it was found that listeners did not benefit from more periodicity in the target speech and that, compared to normal hearing, the benefits obtained from masker amplitude modulations and masker

periodicity were reduced to about half their size. However, although the ability to perceive periodicity information was severely restricted, the masker periodicity benefit was still twice as large as the benefit obtained from masker fluctuations, highlighting the importance and robustness of periodicity cues in the perception of speech in noise. Performance with the glimpsing-promoting maskers, on the other hand, was similar to those with steady maskers, suggesting that other factors, such as difficulties in perceiving segmental boundaries, need to be considered in addition to the susceptibility to energetic masking when explaining the poor performance with fluctuating interferers.

## 6.2 Psychophysical data

Although the finding is not new (see e.g. [de Cheveigné et al., 1995](#); [de Cheveigné et al., 1997b](#)), it is still a puzzling result that the amount of periodicity in the target speech had little effect on speech recognition performance in the presence of a masker. One plausible reason for this, also proposed by [de Cheveigné \(1993; de Cheveigné et al., 1997a\)](#), is that the harmonic structure of the target speech is difficult to extract for the auditory system at low signal-to-noise ratios (SNRs). The data in chapter 2 are very much in line with this idea, as there was a small benefit from target periodicity in experiment 2, where the SNRs were mostly positive, but no such effect at the negative SNRs in experiment 3. In contrast, this finding does not seem to be due to the fact that non-speech maskers were used that could easily be identified as the masker signal, since the studies by de Cheveigné and colleagues referred to above, were all based on concurrent artificial vowels. This rules out an explanation based on cognitive factors, such as attention.

It has been pointed out in several previous studies ([Bernstein and Grant, 2009](#); [Freyman et al., 2012](#); [Smits and Festen, 2013](#)) that the fluctuating-masker benefit (FMB) becomes larger when a test is carried out at a lower SNR, and that this is one important reason for the reduced benefit of hearing-impaired listeners and CI users. This finding was replicated in the current thesis (see Fig. 2.5) but, moreover, a similar pattern was also observed for the masker-periodicity benefit (MPB; see Fig. 2.6). Hence, it appears to be a general rule that the size of the benefit obtained from some

particular acoustic feature depends on the SNR, which is largely determined by the intelligibility of the target speech in quiet listening conditions.

The results of the CI simulation experiment (chapter 5) have shown that the MPB is surprisingly robust, even when periodicity information is transmitted primarily by temporal cues. Compared to the target speech conditions in chapter 2, that led to similar SRTs (i.e. the ones with the lowest intelligibility in quiet; Nx7, FxNx7, and Fx12), the MPB was in fact hardly reduced at all. However, the absolute decrease of the MPB compared to normal-hearing was also larger than that of the FMB (about 6 dB compared to about 2 dB). This once more suggests that the reduced ability to exploit periodicity information may be an even more important factor in explaining the poor performance of CI users in the perception of speech in noise than the inability to benefit from masker fluctuations.

### 6.3 Electrophysiological data

The time-domain analyses of evoked neural activity presented in chapter 3 and the changes in total EEG power in the frequency domain analysed in chapter 4 have in common that effects of intelligibility were more pronounced than effects of periodicity. Differences in intelligibility were found to lead to effects with a greater temporal extension and, at least when including the completely unintelligible spectrally-rotated condition, also to smaller *p*-values. Apart from that, however, there are surprisingly few similarities, given that both chapters are based on the same raw data and that the evoked EEG activity is necessarily also part of the total activity.

In particular, the only measure that parametrically varied with the amount of periodicity in the stimuli was the amplitude of the event-related potentials during the first second after sentence onset, which became more negative. Furthermore, although the intelligibility of the sentences was reflected in both evoked and total EEG activity, the respective time courses varied considerably. The effect was observed from early after sentence onset until shortly after offset in the former case, but first emerged during the second half of the sentences in the latter analysis. It also appears that the delta power increase in response to intelligible speech persists beyond the analysis window ending 2500 ms after sentence onset. This, in turn, raises the

question whether this effect might be driven by the ensuing spoken response, particularly as the time-frequency decomposition algorithm uses future sample points to compute present values. As described in section 4.2.5, the length of the analysis window linearly decreased from 1 to 0.5 s as the frequencies increased from 1 to 30 Hz. Thus, for the 1 Hz frequency bin, sample points of up to 0.5 s ahead of a given data point determined its value. Fig. 4.6, however, shows that the increase in delta power is visible from as early as about 1 s after sentence onset. Given that the mean duration of the stimulus sentences was just over 2 s (2.04 s,  $SD = 0.24$  s), followed by another 0.25 s of silence, it thus cannot be that the verbal response itself elicited the delta power increase. Similarly, the data in Fig. 4.6 show that this effect is not confounded by the beep that signalled the participants to respond after the silent gap, since there are no substantial deviations from baseline in the unintelligible rotated condition towards the end of the stimulus window. Nevertheless, it could be argued that the effect reflects the preparation to respond rather than the intelligibility of the sentences, or a mixture of the two processes. If the stimulus materials are more intelligible to the participants, this would necessarily also require a longer spoken response, which could also explain the greater increase in delta power. In order to tease apart these two explanations, future studies could use a longer interval between sentence and response to reduce the temporal overlap, or even collect no response at all.

Furthermore, although no significant differences between the fully intelligible trials in the aperiodic, mixed, and periodic conditions were observed apart from the increase in theta power in the latter condition (see Fig. 4.3), it should be emphasised that only this condition was considered in the analysis of intelligibility. As the periodic condition is also the one that sounds least natural, which led to a sufficient number of unintelligible trials in the first place, future research is needed to see if similar increases in delta power are also elicited by other sets of stimuli.

## 6.4 Perspectives

One of the central findings of this thesis was the large benefit obtained from periodicity in the masker. As discussed in chapters 2 and 5, this effect is attributed

to a combination of harmonic cancellation, spectral glimpsing, the absence of random envelope modulations, and the lack of low-frequency envelope modulations crucial for speech intelligibility. Future studies should attempt to disentangle these factors and quantify their respective contributions. Chapter 5 already constitutes a step in that direction, as the CI simulation processing eliminated the components of the masker-periodicity benefit that are based on the accurate perception of spectral information, i.e. harmonic cancellation and spectral glimpsing. However, the noise-excited CI simulation used in this experiment also introduced random envelope fluctuations, which considerably changed the modulation spectrum of the periodic maskers (see Fig. 5.7), which makes it difficult to draw conclusions that also apply to normal hearing. Still, the results have shown that the temporal regularity introduced by the F0-related modulations of these maskers by itself substantially aids their segregation from the target speech. The individual contribution of harmonic cancellation, for instance, could be examined by shifting the individual harmonics of the maskers up or down. This would render the periodic maskers inharmonic and hence also *aperiodic*, but would neither affect the ability to glimpse in between the lower masker harmonics nor would it significantly change the modulation spectrum. Such a study is currently under way.

Furthermore, the results presented in chapter 5 should obviously be complemented with data obtained from genuine CI users, as CI simulations only provide an approximation of the acoustic information transmitted via a typical CI. In particular, CI simulations tend to somewhat overestimate performance (e.g. [Cullington and Zeng, 2008](#)), as they ignore individual differences and assume that the devices work ideally in all participants.

In order to gain a better understanding of the complex behavioural data described in chapter 2, comparing the speech intelligibility predictions of different auditory models appears to be a promising approach. Such models vary considerably regarding their assumptions and components, which could prove insightful in determining which acoustic features were driving the observed effects. For example, the recent multi-resolution speech-based envelope power spectrum model

(mr-sEPSM; [Jørgensen et al., 2013](#), see also Figs. 4.8 and 5.7), which contains a modulation filter bank with a wide range of centre frequencies (1–256 Hz), seems well-suited to predict the masker-periodicity benefit, as this effect is in part thought to be caused by the different modulation spectra of periodic and aperiodic maskers. The mr-sEPSM should thus outperform models such as the audibility-based extended speech intelligibility index (ESII; [Rhebergen and Versfeld, 2005](#)), that does not include a modulation filter bank, or the speech-based speech transmission index (sSTI; [Goldsworthy and Greenberg, 2004](#)), which only considers slower envelope modulations ( $<12.5$  Hz). Secondly, none of the aforementioned models (ESII, sSTI, and mr-sEPSM) takes pitch information into account. As harmonic cancellation and stream segregation both depend on the ability to perceive pitch information, these models should in theory not predict the masker-periodicity benefit accurately. Preliminary results indeed support both of these assumptions.

Consequently, these data could also prove useful for improving existing auditory models, which are usually evaluated with maskers that vary widely regarding their spectro-temporal properties (e.g. interfering talkers, car noise, pub noise; see [Jørgensen et al., 2013](#); [Taal et al., 2011](#)), rather than individual acoustic features such as periodicity. Additionally, relatively little is known about how acoustic degradations of the target speech, for example vocoding, affect the predictions of current auditory models. The vast majority of modelling studies has been concerned with the acoustic properties of the maskers, leaving the target speech untouched (for exceptions see e.g. [Christiansen et al., 2010](#)). Spectral resolution and intelligibility of the target speech materials used in chapter 2 vary over a wide range, making them an ideal data set for investigating this issue. Specifically, it would be interesting to see whether the models predict a similarly small effect of periodicity in the target speech.



## 6.5 Conclusions

The main findings of this thesis, grouped into behavioural and electrophysiological results, may be summarised as follows:

1. Acoustic periodicity was shown to be an important factor in the perception of speech in quiet and in noise. Participants substantially benefitted from periodicity in the masker, even more so than from masker amplitude modulations. Surprisingly, however, the presence or absence of periodicity in the target speech did hardly affect speech recognition performance. Even in simulations of cochlear implants, the masker-periodicity benefit was to some extent maintained, demonstrating the robustness of the effect. On the other hand, the absolute reduction compared to normal hearing was also larger than the reduction of the fluctuating-masker benefit, which appears to be an important factor when explaining the poor performance of CI users.
2. EEG signals recorded at the cortical level reflect the amount of acoustic periodicity in the speech signals, their intelligibility, and even the attentional state of the listener before speech onset. Firstly, the event-related potentials during the first second after sentence onset had greater negative amplitudes with more periodicity in the speech stimuli. Secondly, after controlling for acoustic differences, it was found that the power of slow neural delta oscillations was increased when the stimulus sentences were more intelligible to the listeners. A similar effect was observed in the time domain, where intelligible sentences elicited a slow negativity in the ERP. Finally, as reflected by increased neural alpha power, listeners appeared to be less attentive before sentences that turned out to be unintelligible to them.

## References

- Apoux, F., Youngdahl, C. L., Yoho, S. E., and Healy, E. W. (2015). Dual-carrier processing to convey temporal fine structure cues: Implications for cochlear implants. *J. Acoust. Soc. Am.* 138, 1469–1480.
- Ardoint, M., Green, T., and Rosen, S. (2014). The intelligibility of interrupted speech depends upon its uninterrupted intelligibility. *J. Acoust. Soc. Am.* 136, EL275–EL280.
- Bacon, S. P., Opie, J. M., and Montoya, D. Y. (1998). The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. *J. Speech Lang. Hear. Res.* 41, 549–563.
- Becker, R., Pefkou, M., Michel, C. M., and Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Front. Syst. Neurosci.* 7, Article 121.
- Bernstein, J. G. W. and Brungart, D. (2011). Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio. *J. Acoust. Soc. Am.* 130, 473–488.
- Bernstein, J. G. W. and Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 125, 3358–3372.
- Birbaumer, N., Elbert, T., Canavan, A., and Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. *Physiol. Rev.* 70, 1–41.
- Bird, J. and Darwin, C. (1998). Effects of a difference in fundamental frequency in separating two sentences. In *Psychophysical and Physiological Advances in*

- Hearing*, edited by I. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (London: Whurr), pp. 263–269.
- Blessner, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *J. Speech Hear. Res.* 15, 5–41.
- Boersma, P. and Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program]. Version 5.3.49. <http://www.praat.org> (Last viewed May 13, 2015).
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652.
- Botvinick, M. M., Cohen, J. D., and Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* 8, 539–546.
- Bregman, A. S. (1994). *Auditory scene analysis*. Cambridge, MA: MIT press.
- Brokx, J. and Nötteboom, S. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phonetics* 10, 23–36.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* 109, 1101–1109.
- Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929.
- Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., and Lui, C. (1994). An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am.* 96, 2108–2120.
- Calandruccio, L. and Smiljanic, R. (2012). New sentence recognition materials developed using a basic non-native English lexicon. *J. Speech. Lang. Hear. Res.* 55, 1342–1355.
- Chait, M., Poeppel, D., and Simon, J. Z. (2006). Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cereb. Cortex* 16, 835–848.
- Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., and Moreno, A. (1995). EUROM – A

- spoken language resource for the EU. In *Proceedings of Eurospeech*, pp. 867–880.
- Christiansen, C., Pedersen, M. S., and Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Commun.* 52, 678–692.
- Cohen, M. X. and Donner, T. H. (2013). Midfrontal conflict-related theta-band power reflects neural oscillations that predict behavior. *J. Neurophysiol.* 110, 2752–2763.
- Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58, 306–324.
- Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215.
- Culling, J. F. and Darwin, C. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *J. Acoust. Soc. Am.* 93, 3454–3467.
- Cullington, H. E. and Zeng, F.-G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *J. Acoust. Soc. Am.* 123, 450–461.
- Darwin, C. (2008). Listening to speech in the presence of other sounds. *Philos. Trans. R. Soc. London B* 363, 1011–1021.
- Davis, M. H. and Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *J. Neurosci.* 23, 3423–3431.
- de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acoust. Soc. Am.* 93, 3271–3290.
- de Cheveigné, A. (1998). Cancellation model of pitch perception. *J. Acoust. Soc. Am.* 103, 1261–1271.
- de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* 111, 1917–1930.

- de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997a). Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. *J. Acoust. Soc. Am.* 101, 2839–2847.
- de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *J. Acoust. Soc. Am.* 97, 3736–3748.
- de Cheveigné, A., McAdams, S., and Marin, C. M. (1997b). Concurrent vowel identification. II. Effects of phase, harmonicity, and task. *J. Acoust. Soc. Am.* 101, 2848–2856.
- Dellwo, V., Fourcin, A., and Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1129–1132.
- Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21.
- Deroche, M. L. and Culling, J. F. (2011). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *J. Acoust. Soc. Am.* 130, 2855–2865.
- Deroche, M. L., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014a). Roles of the target and masker fundamental frequencies in voice segregation. *J. Acoust. Soc. Am.* 136, 1225–1236.
- Deroche, M. L., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014b). Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity. *J. Acoust. Soc. Am.* 135, 2873–2884.
- Ding, N., Chatterjee, M., and Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88, 41–46.

- Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.*, *19*, 158–164.
- Dudley, H. (1939). Remaking speech. *J. Acoust. Soc. Am.* *11*, 169–177.
- Elliott, T. M. and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* *5*, e1000302.
- Evans, S., Kyong, J., Rosen, S., Golestani, N., Warren, J., McGettigan, C., Mourão-Miranda, J., Wise, R. and Scott, S. (2014). The pathways for intelligible speech: multivariate and univariate perspectives. *Cereb. Cortex* *24*, 2350–2361.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.
- Fant, G., Liljencrants, J., and Lin, Q.-G. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory – Quarterly Progress and Status Report* *4*, 1–13.
- Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Berlin: Springer.
- Faulkner, A., Rosen, S., and Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *J. Acoust. Soc. Am.* *108*, 1877–1887.
- Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.* *88*, 1725–1736.
- Fourcin, A. (2010). A note on voice timing and the evolution of connected speech. *Logoped. Phoniatr. Vocology* *35*, 74–80.
- Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (2012). Intelligibility of whispered speech in stationary and modulated noise maskers. *J. Acoust. Soc. Am.* *132*, 2514–2523.

- Fu, Q.-J. and Nogaki, G. (2005). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *J. Assoc. Res. Otolaryngol.* 6, 19–27.
- Fu, Q.-J., Shannon, R. V., and Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *J. Acoust. Soc. Am.* 104, 3586–3596.
- Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517.
- Gnansia, D., Pean, V., Meyer, B., and Lorenzi, C. (2009). Effects of spectral smearing and temporal fine structure degradation on speech masking release. *J. Acoust. Soc. Am.* 125, 4023–4033.
- Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.* 116, 3679–3689.
- Green, T., Faulkner, A., and Rosen, S. (2002). Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. *J. Acoust. Soc. Am.* 112, 2155–2164.
- Green, T., Faulkner, A., and Rosen, S. (2004). Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *J. Acoust. Soc. Am.* 116, 2298–2310.
- Green, T., Faulkner, A., Rosen, S., and Macherey, O. (2005). Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification. *J. Acoust. Soc. Am.* 118, 375–385.
- Green, T. and Rosen, S. (2013). Phase effects on the masking of speech by harmonic complexes: Variations with level. *J. Acoust. Soc. Am.* 134, 2876–2883.
- Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J. Acoust. Soc. Am.* 87, 2592–2605.

- Griffiths, T. D., Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., Brugge, J. F., and Howard, M. A. (2010). Direct recordings of pitch responses from human auditory cortex. *Curr. Biol.* 20, 1128–1132.
- Guimond, S., Vachon, F., Nolden, S., Lefebvre, C., Grimault, S., and Jolicœur, P. (2011). Electrophysiological correlates of the maintenance of the representation of pitch objects in acoustic short-term memory. *Psychophysiology* 48, 1500–1509.
- Gutschalk, A., Patterson, R. D., Rupp, A., Uppenkamp, S., and Scherg, M. (2002). Sustained magnetic fields reveal separate sites for sound level and temporal regularity in human auditory cortex. *Neuroimage* 15, 207–216.
- Gutschalk, A., Patterson, R. D., Scherg, M., Uppenkamp, S., and Rupp, A. (2004). Temporal dynamics of pitch in human auditory cortex. *Neuroimage* 22, 755–766.
- Hanslmayr, S., Pastötter, B., Bäuml, K.-H., Gruber, S., Wimber, M., and Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *J. Cognit. Neurosci.* 20, 215–225.
- Harmony, T. (2013) The functional significance of delta oscillations in cognitive processing. *Front. Integr. Neurosci.* 7, Article 83.
- Hazan, V. and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Am.* 130, 2139–2152.
- He, B. J. and Raichle, M. E. (2009). The fMRI signal, slow cortical potential and consciousness. *Trends Cogn. Sci.* 13, 302–309.
- Herrmann, B., Henry, M. J., Haegens, S., and Obleser, J. (2016). Temporal expectations and neural amplitude fluctuations in auditory cortex interactively influence perception. *Neuroimage* 124, 487–497.
- Hopkins, K. and Moore, B. C. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *J. Acoust. Soc. Am.*, 125, 442–446.



- Hopkins, K., Moore, B. C., and Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *J. Acoust. Soc. Am.* 123, 1140–1153.
- Jensen, O. and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4, Article 186.
- Jørgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *J. Acoust. Soc. Am.* 134, 436–446.
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936.
- Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.* 30, 620–628.
- Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195.
- Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends Cogn. Sci.* 16, 606–617.
- Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., and Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neurosci. Lett.* 244, 73–76.
- Kononowicz, T. W. and Penney, T. B. (2016). The contingent negative variation (CNV): timing isn't everything. *Curr. Opin. Behav. Sci.* 8, 231–237.
- Kwon, B. J., Perry, T. T., Wilhelm, C. L., and Healy, E. W. (2012). Sentence recognition in noise promoting or suppressing masking release by normal-hearing and cochlear-implant listeners. *J. Acoust. Soc. Am.* 131, 3111–3119.
- Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., and Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* 94, 1904–1911.

- Lakatos, P., Barczak, A., Neymotin, S. A., McGinnis, T., Ross, D., Javitt, D. C., and O'Connell, M. N. (2016). Global dynamics of selective attention and its lapses in primary auditory cortex. *Nat. Neurosci.* 19, 1707–1717.
- Laufs, H., Holt, J. L., Elfont, R., Krams, M., Paul, J. S., Krakow, K., and Kleinschmidt, A. (2006). Where the BOLD signal goes when alpha EEG leaves. *Neuroimage* 31, 1408–1418.
- Lefebvre, C., Vachon, F., Grimault, S., Thibault, J., Guimond, S., Peretz, I., Zatorre, R. J., and Jolicœur, P. (2013). Distinct electrophysiological indices of maintenance in auditory and visual short-term memory. *Neuropsychologia* 51, 2939–2952.
- Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press.
- Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc. Natl. Acad. Sci.* 103, 18866–18869.
- Makeig, S. (1993). Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. *Electroencephalogr. Clin. Neurophysiol.* 86, 283–293.
- Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190.
- Mazaheri, A. and Jensen, O. (2008). Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. *J. Neurosci.* 28, 7781–7787.
- Mazaheri, A. and Jensen, O. (2010). Rhythmic pulsing: linking ongoing brain activity with evoked responses. *Front. Hum. Neurosci.* 4, Article 177.
- McCallum, W. and Walter, W. G. (1968). The effects of attention and distraction on the contingent negative variation in normal and neurotic subjects. *Electroencephalogr. Clin. Neurophysiol.* 25, 319–329.
- Miller, G. A. and Licklider, J. (1950). The intelligibility of interrupted speech. *J. Acoust. Soc. Am.* 22, 167–173.

- Millman, R. E., Johnson, S. R., and Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J. Cognit. Neurosci.* 27, 533–545.
- Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J. Assoc. Res. Otolaryngol.* 9, 399–406.
- Moore, B. C. J. (2012). The importance of temporal fine structure for the intelligibility of speech in complex backgrounds. In *Speech Perception and Auditory Disorders*, edited by T. Dau, M. L. Jepsen, T. Poulsen, and J. C. Dalsgaard (Ballerup, Denmark: The Danavox Jubilee Foundation), pp. 21–32.
- Müller, N. and Weisz, N. (2012). Lateralized auditory cortical alpha band activity and interregional connectivity pattern reflect anticipation of target sounds. *Cereb. Cortex* 22, 1604–1613.
- Nelson, D. A. and Donaldson, G. S. (2001). Psychophysical recovery from single-pulse forward masking in electric hearing. *J. Acoust. Soc. Am.* 109, 2921–2933.
- Nelson, P. B. and Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *J. Acoust. Soc. Am.* 115, 2286–2294.
- Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *J. Acoust. Soc. Am.* 113, 961–968.
- Norman-Haignere, S., Kanwisher, N., and McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.* 33, 19451–19469.
- Obleser, J. and Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage* 55, 713–723.
- Obleser, J. and Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cereb. Cortex* 22, 2466–2477.

- Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., and Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *J. Neurosci.* 32, 12376–12383.
- Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants. *Trends Amplif.* 12, 316–331.
- Oxenham, A. J. and Simonson, A. M. (2009). Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference. *J. Acoust. Soc. Am.* 125, 457–468.
- Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., and Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36, 767–776.
- Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* 3, Article 320.
- Peelle, J. E., Gross, J., and Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* 23, 1378–1387.
- Peters, R. W., Moore, B. C. J., and Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *J. Acoust. Soc. Am.* 103, 577–587.
- Pfurtscheller, G., Stancak, A., and Edlinger, G. (1997). On the existence of different types of central beta rhythms below 30 Hz. *Electroencephalogr. Clin. Neurophysiol.* 102, 316–325.
- Picton, T. W., Hillyard, S. A., Krausz, H. I., and Galambos, R. (1974). Human auditory evoked potentials. I: Evaluation of components. *Electroencephalogr. Clin. Neurophysiol.* 36, 179–190.
- Plomp, R. and Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Int. J. Audiol.* 18, 43–52.
- Pratt, H. (2011). Sensory ERP Components. In *The Oxford Handbook of Event-Related Potential Components*, edited by S. J. Luck and E. S. Kappenman (New York: Oxford University Press USA), pp. 89–114.

- Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J. Acoust. Soc. Am.* *114*, 446–454.
- Rasch, R. and Plomp, R. (1999). The perception of musical tones. In *The Psychology of Music*, edited by D. Deutsch (San Diego: Academic Press), pp. 89–112.
- Rhebergen, K. S. and Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J. Acoust. Soc. Am.* *117*, 2181–2192.
- Romei, V., Gross, J., and Thut, G. (2010). On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: correlation or causation? *J. Neurosci.* *30*, 8692–8697.
- Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. London B* *336*, 367–373.
- Rosen, S. and Iverson, P. (2007). Constructing adequate non-speech analogues: what is special about speech anyway? *Developmental Science* *10*, 165–168.
- Rosen, S., Souza, P., Ekelund, C., and Majeed, A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise-vocoding. *J. Acoust. Soc. Am.* *133*, 2431–2443.
- Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* *17*, 225–246.
- Roux, F. and Uhlhaas, P. J. (2014). Working memory and neural oscillations: alphagamma versus thetagamma codes for distinct WM information? *Trends Cogn. Sci.* *18*, 16–25.
- Schoof, T., Green, T., Faulkner, A., and Rosen, S. (2013). Advantages from bilateral hearing in speech perception in noise with simulated cochlear implants and residual acoustic hearing. *J. Acoust. Soc. Am.* *133*, 1017–1030.

- Schubert, R., Haufe, S., Blankenburg, F., Villringer, A., and Curio, G. (2009). Now you'll feel it, now you won't: EEG rhythms predict the effectiveness of perceptual masking. *J. Cognit. Neurosci.* 21, 2407–2419.
- Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123, 2400–2406.
- Sedley, W., Teki, S., Kumar, S., Overath, T., Barnes, G. R., and Griffiths, T. D. (2012). Gamma band pitch responses in human auditory cortex measured with magnetoencephalography. *Neuroimage* 59, 1904–1911.
- Shahin, A. J., Picton, T. W., and Miller, L. M. (2009). Brain oscillations during semantic evaluation of speech. *Brain Cogn.* 70, 259–266.
- Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., and Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *Neuroimage* 41, 113–122.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304.
- Shulman, G. L., Astafiev, S. V., McAvoy, M. P., d'Avossa, G., and Corbetta, M. (2007). Right TPJ deactivation during visual search: functional significance and support for a filter hypothesis. *Cereb. Cortex* 17, 2625–2633.
- Smits, C. and Festen, J. M. (2013). The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: II. Fluctuating noise. *J. Acoust. Soc. Am.* 133, 3004–3015.
- Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* 32, 8443–8453.
- Steinmetzger, K. and Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *J. Acoust. Soc. Am.* 138, 3586–3599.
- Steinmetzger, K. and Rosen, S. (2017). Effects of acoustic periodicity, intelligibility, and pre-stimulus alpha power on the event-related potentials in response to speech. *Brain Lang.* 164, 1–8.

- Steinmetzger, K. and Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia* 95, 173–181.
- Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F.-G. (2007). Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences. *J. Acoust. Soc. Am.* 122, 1069–1078.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *J. Acoust. Soc. Am.* 116, 1081–1091.
- Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. (2011). The importance for speech intelligibility of random fluctuations in ‘steady’ background noise. *J. Acoust. Soc. Am.* 130, 2874–2881.
- Stone, M. A., Füllgrabe, C., and Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *J. Acoust. Soc. Am.* 132, 317–326.
- Strauß, A., Kotz, S. A., Scharinger, M., and Obleser, J. (2014a). Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage* 97, 387–395.
- Strauß, A., Wöstmann, M., and Obleser, J. (2014b). Cortical alpha oscillations as a tool for auditory selective inhibition. *Front. Hum. Neurosci.* 8, Article 350.
- Strauß, A., Henry, M. J., Scharinger, M., and Obleser, J. (2015). Alpha phase determines successful lexical decision in noise. *J. Neurosci.* 35, 3256–3262.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Language Process.* 9, 2125–2136.
- Tallon-Baudry, C. and Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn. Sci.* 3, 151–162.
- Tecce, J. J. and Scheff, N. M. (1969). Attention reduction and suppressed direct-current potentials in the human brain. *Science* 164, 331–333.

- Teder-Sälejärvi, W. A., Münte, T. F., Sperlich, F.-J., and Hillyard, S. A. (1999). Intra-modal and cross-modal spatial attention to auditory and visual stimuli. An event-related brain potential study. *Cognitive Brain Res.* 8, 327–343.
- Turner, C. W., Reiss, L. A., and Gantz, B. J. (2008). Combined acoustic and electric hearing: preserving residual acoustic hearing. *Hearing Res.* 242, 164–171.
- Van Dijk, H., Schoffelen, J.-M., Oostenveld, R., and Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *J. Neurosci.* 28, 1816–1823.
- Vestergaard, M. D. and Patterson, R. D. (2009). Effects of voicing in the recognition of concurrent syllables. *J. Acoust. Soc. Am.* 126, 2860–2863.
- Weisz, N. and Obleser, J. (2014). Synchronisation signatures in the listening brain: a perspective from non-invasive neuroelectrophysiology. *Hearing Res.* 307, 16–28.
- Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., and Obleser, J. (2015). Alpha oscillatory dynamics index temporal expectation benefits in working memory. *Cereb. Cortex* 25, 1938–1946.
- Wilson, B. S. and Dorman, M. F. (2008). Cochlear implants: a remarkable past and a brilliant future. *Hearing Res.* 242, 3–21.
- Whitmal, N. A. III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *J. Acoust. Soc. Am.* 122, 2376–2388.
- Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept. Psychophys.* 63, 1293–1313.
- Wisniewski, M. G., Thompson, E. R., Iyer, N., Estepp, J. R., Goder-Reiser, M. N., and Sullivan, S. C. (2015). Frontal midline  $\theta$  power as an index of listening effort. *Neuroreport* 26, 94–99.
- Woods, D. L., Alho, K., and Algazi, A. (1994). Stages of auditory feature conjunction: an event-related brain potential study. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 81–94.



- Wöstmann, M., Herrmann, B., Wilsch, A., and Obleser, J. (2015a). Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits. *J. Neurosci.* 35, 1458–1467.
- Wöstmann, M., Schröger, E., and Obleser, J. (2015b). Acoustic detail guides attention allocation in a selective listening task. *J. Cognit. Neurosci.* 27, 988–1000.
- Wöstmann, M., Herrmann, B., Maess, B., and Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proc. Natl. Acad. Sci.* 113, 3873–3878.
- Xu, Y. (2013). ProsodyPro – A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, pp. 7–10.
- Yrttiaho, S., Tiitinen, H., May, P. J., Leino, S., and Alku, P. (2008). Cortical sensitivity to periodicity of speech sounds. *J. Acoust. Soc. Am.* 123, 2191–2199.
- Yrttiaho, S., Tiitinen, H., Alku, P., Miettinen, I., and May, P. J. (2010). Temporal integration of vowel periodicity in the auditory cortex. *J. Acoust. Soc. Am.* 128, 224–234.
- Yrttiaho, S., May, P. J., Tiitinen, H., and Alku, P. (2011). Cortical encoding of aperiodic and periodic speech sounds: Evidence for distinct neural populations. *Neuroimage* 55, 1252–1259.