University of Southampton

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Auditory cortical responses as an objective predictor of speech-in-noise performance

by

Suwijak Deoisres

ORCID ID 0000-0003-1384-3632

Thesis for the degree of Doctor of Philosophy

March 2023

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Thesis for the degree of Doctor of Philosophy

Auditory cortical responses as an objective predictor of speech-in-noise performance

by

Suwijak Deoisres

The cortical entrainment of speech envelope is a phenomenon where the electrical activity in the brain fluctuates with the change in stimulus intensity, i.e., the stimulus envelope. While clearly speech stimuli are closer to everyday conversation, it is not clear whether the cortical entrainment of speech envelope closely reflects speech intelligibility in the brain more than cortical responses to other types of sound. The overall goal of this thesis is to assess the applicability of cortical responses to speech and non-speech sound for use as a predictor of behavioural speech intelligibility in normal hearing people. Initial works were carried out to explore how to best measure and detect cortical responses to continuous speech.

In the first study cortical responses to continuous speech were stronger when additional pauses were inserted into the stimulus, and this consequently led to a greater the number of detected responses. This also demonstrate that the amount of pauses could introduce a comparison bias as they have different envelopes, it was therefore decided that the continuous speech and non-speech stimuli used in the third study would comprise identical envelope shapes. In the second study, the backward linear modelling (reconstructing speech envelope from cortical responses) was able to detect greater number of significant responses to continuous speech than the forward linear modelling (predicting cortical responses from speech envelope), however, the number of detected responses was lower than the cortical auditory evoked potentials (CAEP) using Hotelling's T-squared. This suggests that evoked responses to continuous speech are harder to detect than more standard evoked responses. In the third study, behavioural speech recognition scores and objective measures showed significant correlation and it was possible to utilise cortical responses to continuous speech, modulated noise, and /da/ as objective measures to predict speech intelligibility in some subjects. However, the reliability of response measurement was poor, consequently the speech intelligibility prediction was inaccurate in many individuals.

With regards to the main goal of this thesis, all types of cortical responses showed expected monotonic relationship with the stimulus signal-to-noise ratio on a group level, cortical responses to continuous speech were worse in predicting the speech intelligibility in individuals than other types of responses. This gives empirical evidence to show that measurement of cortical responses to continuous speech through the linear modelling approach with scalp responses measurement are less reliable compared to cortical responses to less natural speech stimuli at an individual level. For normal hearing people, a measurement that simply confirms audibility (not necessarily speech intelligibility or comprehension) may be sufficient to predict speech-in-noise test performance.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name:	Suwijak Deoisres

Title of thesis:	Auditory cortical responses as an objective predictor of speech-in-noise performance

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature:		Date:	

Acknowledgements

This thesis was made possible with the support from several wonderful people. I, Suwijak, would like to dedicate this section to acknowledge their contribution and say thank you to:

David Simpson, as main my supervisor, for your mentorship since I was an MSc student and during my PhD. Thank you for your ideas and suggestions that form the foundation of my PhD research. Thank you for sharing your valuable knowledge and experience, being a great role model in research. Thank you for showing your care when my sponsor was being mean to me. I also thank you for the good conversations unrelated to work, which reminds me to make the most of my time during my stay in the UK.

Steve Bell, as my second supervisor, for your kind supervision and timely feedback on my work. Thank you for regularly sharing interesting papers and providing insightful ideas from an audiologist's point of view. I am also grateful for your help in getting the experiment for the third study ready.

Michael Chesnaye, for helping me with many things (coding, statistics, EEG simulations, participating in the pilot study, and etc.) since I was an MSc student.

Yuhan Lu and Fred Vanheusden, for the collecting dataset used in the first and second study. Many interesting findings in this thesis arise from the analysis of your dataset.

Sam Perry, for allowing me to join in your data collection showing me how to setup the tools. Although this need to be cancelled due to COVID-19, I am still grateful for your help.

Ghada Aljarboa, for performing the pure-tone testing in the third study and helping me with the data collection.

All participants in the third study, for joining my study and enduring several hours of experiment. Thank you for your good cooperation, giving me a great experience for my first time of data collection.

My PhD journey would not have been possible without the sponsorship from The Royal Thai Government (Ministry of Higher Education, Science, Research and Innovation).

Last but not least, I would like to thank you my family and friends for their love and support, encouraging me to forge ahead.

Definitions and Abbreviations

ABR	Auditory brainstem responses
AERs	Auditory evoked responses
ALR	Auditory late responses
AM	Amplitude modulated
AMLR	Auditory middle latency responses
APD	Auditory processing disorder
APD	Central auditory processing disorder
ASSR	Auditory steady-state responses
BACKWARD-CORR	Correlation coefficient obtain from the backward TRF
BKB-SIN	Bench-Kowal-Bamford Speech-in-noise
CA	Coherent average
CA-da	Coherent average of cortical responses to /da/
CAEP	Cortical auditory evoked potentials
CF	Carrier frequency
CT	Correlation threshold
DC	Direct current
EEG	Electroencephalogram
F0	Fundamental frequency
FFR	Frequency-following response
FIR	Finite impulse response
H0	Null-hypothesis
H2	Second harmonic
HINT	Hearing in noise test
HL	Hearing level
HT^2	Hotelling's T-squared

IEEE	Institute of Electrical and Electronic Engineering
ISI	Inter-stimulus interval
LOOCV	Leave-one-out cross validation
MF	Modulation frequency
MSE	Mean squared error
PTA	Pure-tone audiometry
SD	Standard deviation
SIN	Speech-in-noise
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRT	Speech reception threshold
SWN	Speech weighted noise
TRF	Temporal response function
TRF-model	Forward TRF model weight
TVM	Time-voltage mean
WRS	Word recognition score

Chapter 1 **Introduction**

Perception of speech is essential for learning, communication, and socialising. A major challenge affecting the quality of life for people with hearing impairments, and even for people with normal hearing, is the ability to understand speech particularly in a noisy environment. This problem can lead to, for example, difficulties in learning at school or social isolation especially for the elderly. Speech perception refers to the ability to perceive and understand speech (Heald and Nusbaum, 2014). It can be measured through behavioural tests, where the subjects are expected respond by repeating the word they hear or click/press one of a set of choices displayed on a screen. It is usually measured as the percentage of speech that can be recalled by the listener in a quiet or noisy condition. Since this type of measurement involves cognitive performance, such as attention, memory, executive function, and processing speed (Dryden *et al.*, 2017), some group of subjects (e.g., children and people with cognitive impairments) may not provide reliable behavioural responses.

Conventionally, hearing testing is most commonly carried out using pure-tone audiometry (PTA). Although the degree of hearing loss can be measured using PTA, the pure-tone stimuli used in the test does not represent the real-world issues in speech perception. Moreover, PTA is also a behavioural measure, and so has the same limitation as the behavioural measure of speech intelligibility. For example, participants are required to respond to the test whereas some participants, such as children or people with cognitive impairment, may not be able to respond reliably or at all (Hirsh, 1948). Behavioural tasks in speech-in-noise tests are even more challenging for the participants than in PTA because the participants are required to memorise, verbally repeat the speech material, or even understand the speech and answer questions. Auditory evoked responses (AERs), objective electrophysiological responses that measure brain response to sound stimuli, can be an alternative for estimating hearing threshold and are also used in other clinical applications (Stapells, Gravel and Martin, 1995; Ronnberg *et al.*, 2016). The stimulus used in AERs can be as simple as clicks and tone pips, or more complex such as syllables in a word, sentences, and narrative speech.

AERs to short repeating stimuli have been used to predict speech intelligibility. AERs to short repeating stimuli have been shown to be a potential predictor for speech intelligibility, as the amplitude of AERs is found to decrease as the stimuli is presented

with increasing noise levels and the strength of responses correlate with the behavioural speech intelligibility performance (Wong, Cheung and Wong, 2008; Papakonstantinou, Strelcyk and Dau, 2011; Zhang, Gong and Zhang, 2016). Evidence suggests that AERs may reflect two main processes relating to speech intelligibility, which are auditory perception and cognitive activity related to speech discrimination in noise. The latter has been found in the AERs at the cortex level generated by short speech stimuli (Billings *et al.*, 2013; Goossens *et al.*, 2018).

The response in AERs to continuous speech can be observed in correlations between the EEG and the envelope of the speech signals and is usually referred to as envelope entrainment. The envelope is however only a crude representation of the speech stimuli. It is also found that attention plays an important role in the envelope entrainment (Naatanen, 1990; Power *et al.*, 2012; O'Sullivan *et al.*, 2015; Reetzke, Gnanateja and Chandrasekaran, 2021), which makes the protocols requiring attention unsuitable for some groups of listeners (e.g., children and people with cognitive impairment). Moreover, measurement of AERs to continuous speech normally requires longer test times than AERs to brief stimuli because the latter takes advantage of powerful noise reduction from the coherent averaging technique, whereas AERs to continuous speech require more complex system identification methods. This leads to the question of whether there is any benefit of using AERs to continuous speech over AERs to brief stimuli in predicting speech intelligibility.

The overall goal of this thesis is to examine whether speech intelligibility is better predicted from the envelope entrainment rather than conventional-AERs to transient stimuli. As a first step, this study will explore how best to measure evoked responses to natural speech and better understand how those responses compare with those of simple repeating stimuli.

1.1 Aim of thesis

The overall goal of this thesis is to investigate whether cortical responses to continuous speech are better than cortical responses to less natural speech (continuous speech with additional pauses and repeating monosyllables) or non-speech (modulated broadband noise) stimuli as an objective measure for detecting cortical responses to speech sound and for the application of predicting individual speech-in-noise performance.

Objectives:

1. Assess how adding additional pauses between each word in the natural speech stimulus changes the cortical response.
2. Determine whether onset or non-onset segments in the continuous speech dominates the cortical responses measured.
3. Compare the characteristics of cortical responses to continuous speech and repeating short stimulus.
4. Compare the effectiveness of different AERs (TRF-model, backward TRF, correlation coefficient with continuous speech, and coherent averages of CAEP to repeating short stimulus) in detecting cortical responses to speech.
5. Compare the accuracy of behavioural speech reception threshold (SRT) prediction from cortical responses to continuous speech, and a non-speech stimulus, and CAEP to repeating short stimulus.

1.2 Publications

- Conferences
    - Gravenor H., Deoisres S., and Bell S.L. (2019). Exploring the effect of stimulus level on da responses. International Evoked Response Audiometry Study Group (XXVI). Sydney, Australia. *Poster presentation*.
    - Deoisres S., Lu Y., Simpson D.M., and Bell S.L. (2020). Cortical auditory evoked responses to continuous speech stimuli with extended pauses. European Medical and Biological Engineering Conference (8th). Portoroz, Slovenia. *Presentation abstract (conference cancelled due to COVID-19)*.
    - Deoisres S., Lu Y., Simpson D.M., and Bell S.L. (2021). Cortical auditory evoked responses to continuous speech stimuli with extended pauses. International Evoked Response Audiometry Study Group (XXVII). Online conference. *Oral Presentation.*
    - Deoisres S., Bell S.L., and Simpson D.M. (2022). Predicting speech reception threshold using cortical responses to sound. UK Hearing Audiology and Sciences Meeting. Southampton, United Kingdom. *Poster presentation*.
- Journals
    - Deoisres S., Lu Y., Vanheusden F.J., Bell S.L., and Simpson D.M. Continuous speech with pauses inserted between words enhances cortical envelope entrainment. *PLOS one*. *Submitted and in revision*.
    - Deoisres S., Bell S.L., and Simpson D.M. Predicting speech reception threshold using cortical responses to sound. *Preparing for submission.*

1.3 Outline of thesis

The organisation of the thesis chapters is briefly described in this section. The thesis begins with an introduction (Chapter 1) to auditory evoked response, including a description of the purpose of this thesis and the contribution to the field of research.

Chapter 2 is dedicated to a review of the literature. This chapter firstly describes speech intelligibility and the important factors which affect the ability to understand speech both in quiet and with background noise. Some of the commonly used speech-in-noise tests in clinic and research are described. Finally, the relation between speech intelligibility and the gold standard hearing test (pure-tone audiometry) is briefly explored.

Chapter 3 provides an explanation of auditory evoked responses. Then the use of cortical responses to various types of sound stimuli for assessing individual speech intelligibility is reviewed. This is followed by a description of objective methods to detect cortical responses to continuous sound stimuli and cortical auditory evoked potentials (CAEP), which are used throughout this thesis.

Chapter 4 describes the first study, exploring whether cortical responses to continuous speech stimuli can be enhanced by modifying the stimulus through the insertion of silent pauses between words to improve response detection sensitivity. Then there is more examination of the contribution of onset and non-onset segments in the speech envelope.

Chapter 5, the second study, further investigates the dataset presented in the previous chapter regarding the characteristics of cortical responses and comparing its morphology to the CAEP waveform. This study also includes a comparison between two methods for detecting cortical responses to continuous speech to selecting the most sensitive method to be used in detecting responses in condition with background noise in the next study.

Chapter 6, the third study, explores cortical responses to natural speech, noise modulated by the same envelope as the natural speech, and repeating /da/ stimuli as an objective measure to predict speech in noise performance.

The thesis concludes in Chapter 7 with the general discussion from the main findings and suggestions for future work.

Chapter 2 A review of speech intelligibility

This chapter provides a review of the literature relevant to the research project, which will be divided into four sections. The first section (2.1) introduces the importance of speech intelligibility. The second section (2.2) provides an overview of the behavioural measurements of speech intelligibility that are commonly used in clinic and research. The third section (2.3) reviews how pure-tone audiometry has been related to the behavioural speech intelligibility test. The fourth section (2.4) is a summary of the findings on how subcortical and cortical auditory evoked responses (AERs) have been used to predict speech intelligibility.

2.1 Speech intelligibility

Speech perception is a process in which the listener perceives and use their knowledge to combine the linguistic components to form their understanding of the speech (Heald and Nusbaum, 2014). The process of speech perception is more complex than the perception of non-speech sound. In addition to sound perception, it involves cognitive capacities to discriminate between speech and sounds that are not relevant to speech perception (Curtin *et al.*, 2017). The process of speech perception can be divided into 4 stages: detection, discrimination, identification, and comprehension (Erber, 1976). The role of each stage is described below.

1. Detection: the listener perceives sound.
2. Discrimination: the listener perceives and is able to distinguish between different sounds or between speech and environmental noise.
3. Identification: the listener perceives and can identify the word they heard.
4. Comprehension: the listener understands the meaning of speech sound and links it to prior knowledge.

In this thesis, speech intelligibility is defined as the ability to hear and identify speech under different listening conditions. Specifically, for the current study, the speech intelligibility will only measure up to the identification stage of speech perception. The speech comprehension stage will not be investigated in this thesis.

The intelligibility of speech has been altered primarily in four different ways:

1. Adding noise to compete with the target speech (Plomp and Mimpen, 1979; Meyer, Dentel and Meunier, 2013)
2. Corrupting the spectral resolution of the speech via noise-vocoding; speech is filtered in to single or multiple frequency bands, then use the amplitude envelope of filtered speech to modulate a noise signal corresponding to the same frequency band (Roberts, Summers and Bailey, 2011).
3. Manipulating the temporal envelope of speech (Noordhoek, Houtgast and Festen, 2001; Jorgensen, Decorsiere and Dau, 2015)
4. Playing the speech stimulus forward or backward in time (Di Liberto, O'Sullivan and Lalor, 2015)

The current study will only focus on the effects of background noise on speech intelligibility, as it is the most common way used in the literature.

To understand speech, people with hearing impairment normally require higher levels of speech relative to background noise level than people with normal hearing. Signal-to-noise ratio (SNR) is a term used to describe the level of desired signal, e.g., target sound or speech, relative to the level of background noise, this is expressed in decibel (dB). In a virtual restaurant study by Culling (2016), a tolerable level for conversation in normal hearing people is at approximately -5 dB SNR, before half of the conversation cannot be understood. The ratio between an increase of approximately 3 to 6 dB SNR is required, depending on type of background noise, for people with hearing impairment to understand speech when compared to people with normal hearing (Plomp, 1994). The increase in SNR also depends on the degree of hearing loss, for example, people with cochlea implants may need up to 20 dB SNR (Zaltz *et al.*, 2020). Adults who have difficulty in understanding speech have difficulties in daily communication (Gordon-Salant, 2005). In children, the difficulty in understanding speech leads to problems such as impairment in psychosocial development and in the development of language skills, which could greatly impact their long-term quality of life. Those problems are associated with academic performance, behaviour, and emotional development (Bess, Dodd-Murphy and Parker, 1998). Children with hearing loss also show delayed development of language compared to children with normal hearing (Tomblin *et al.*, 2015).

2.2 Factors affecting speech intelligibility

There are several factors that may affect intelligibility of speech to consider when designing an experiment or clinical measurement. The effects of the factors on speech intelligibility will be explained in this section.

2.2.1 Types of background noise

The acoustic cues in speech are less perceivable by listeners as a consequence of having noise competing with the target speech. Listeners may be unable to recognise some words in the sentence and the severity increases with decreasing SNR. However, the intelligibility of speech is affected differently depending on the type of background noise. The types of noise, characterised by the temporal and spectral features that has been used in speech-in-noise studies are (Meyer, Dentel and Meunier, 2013; Lee *et al.*, 2015; Le Prell and Clavier, 2017):

1. Stationary noise: temporal and spectral characteristics remains fairly constant with time (e.g., environmental, ambient, stationary white or coloured noise, speech-shaped noise).
2. Non-stationary noise: spectral characteristics change with time (e.g., conversation both intelligible or not, and road traffic).

Most of the acoustic cues for speech can still be distinguished from stationary noise as the spectral and temporal content does not fully overlap with the temporal and spectral feature of speech (Le Prell and Clavier, 2017). The stationary noise could be modified to a speech-shaped noise, by using the long-term average of the target speech spectra (Byrne *et al.*, 1994), which overlaps fully with the target speech.

The temporal fluctuations in non-stationary masking sounds can sometimes be used by the listeners to improve speech intelligibility. Studies have found better speech intelligibility performance when normal hearing people listen to speech in fluctuating noise rather than stationary noise of the same SNR. This can probably be explained through the existence of short silence gaps in the noise which makes the speech clearly perceivable by the listener (Miller, 1947; Festen and Plomp, 1990; Bacon, Opie and Montoya, 1998) –sometimes referred to as ‘glimpsing’. In contrast, people with hearing impairment did not benefit from the short silent gaps (Nejime and Moore, 1998).

Speech-shaped noise can sometimes be referred as speech-weighted noise (SWN) because it mainly contains the spectral power in the frequency band of 100 Hz to 5 kHz analogous to the spectral power of speech (Renz, Leistner and Liebl, 2018). SWN can be presented as stationary (same as speech-shaped noise) or fluctuating (more similar to real speech waveform). The masking of SWN can be greatest when the modulation rate of SWN is similar to the syllabic rate of speech (Sek *et al.*, 2015).

Informational noise is masking of speech with competing speech, this simulates the "cocktail party problem" (Cherry, 1953), where the listener needs to pay more attention to the speaker in interest and ignore other speakers. The competing speech typically may include up to 16-talkers to distract the listener from the target speech. Information masking significantly lowers the speech-in-noise performance when there are lower number of speakers (from 1 to 2 or 4) compared greater number of talkers (8 to 16) (Rosen *et al.*, 2013), as in the former the masking speech remains somewhat intelligible. The speech intelligibility performance also was found to be affected by intelligibility of informational noise (e.g., language that the listener can vs cannot understand), listeners were able to focus on the target speech better if the informational noise is less intelligible (spoken in language that the listener does not understand) (Presacco, Simon and Anderson, 2016).

Informational noise is often used in preference to stationary noise for speech intelligibility measurements because they mostly simulate the everyday situation where it is difficult to understand speech.

2.2.2 Listening effort and background noise type

Listening in a noisy condition usually involves increased listening effort, the mental effort for maintaining attention to and comprehending speech (Rennies *et al.*, 2018). Generally, listening effort (often measured by subjective ratings) increases for decreasing SNR (less intelligible speech) and the demand for listening effort is higher for people with hearing impairment (Rennies *et al.*, 2014). Listening effort can also vary for different types of masking noise. Krueger *et al.* (2017) showed that more listening effort is required when speech is presented in stationary noise when compared to fluctuating noise. However, a study by Desjardins and Doherty (2013) did not find any difference in listening effort across masking conditions, even in 6-talker babble noise. This could indicate that listening effort varies between listeners and it may depend on other factors as well (e.g., cognitive function). It might also be pointed out that in low SNR conditions, effort may again decrease, as the listener may disengage from the difficult task (Peelle, 2018).

2.2.3 Binaural release from masking

Binaural release from masking is the ability of the auditory system to isolate the target sound of interest from background noise using auditory localization cues, such as interaural time and level differences (i.e., the delay between the sound arriving at each ear and differences in sound level) (Avan, Giraudet and Buki, 2015). Binaural hearing provides advantages in locating the source of sound and perceiving louder sound through binaural summation (van Hoesel, 2004; Epstein and Florentine, 2012). It is also beneficial in discriminating speech in noise when the target speech is spatially separated from background noise (Avan, Giraudet and Buki, 2015). The relation between phases of presented sound in the left and right ears, referred to as the interaural phase, can also affect the masking release. Hirsh (1948) observed lower binaural reception threshold (better perception) compared to monaural threshold when signal is presented binaurally out of phase, but noise is presented binaurally in phase or vice versa (opposite interaural phase). The binaural reception threshold becomes higher than monaural threshold when the interaural phase between the signal and noise is the same. The monaural threshold decreases when noise presented binaurally are in phase rather than out of phase.

2.2.4 Speaker gender

Whether female or male voice is more intelligible to the listener has been debated for a long time. Many studies tend to support that female speakers produce more intelligible speech with the advantage of having higher pitched sound (fundamental frequency) when there is no background noise (Bradlow, Torretta and Pisoni, 1996; Ferguson, 2004; Johnson and Ferguson, 2016). Some showed that male voices are more favourable to the listener (Ellis *et al.*, 1996; Kilic and Ogut, 2004; Johnson and Ferguson, 2016). Considering the fact that gender-biases may be the main factor in speech intelligibility (Ellis *et al.*, 1996), it also important to focus on the key factors that contribute to the intelligibility of speech. Bond and Moore (1994) suggested that less intelligible speech contains these characteristics: words with short duration, shorten vowel duration, and high variation in the amplitude of stressed vowels.

2.2.5 Phonetic balance

Speech audiometry is an audiological test to determine the intensity level or SNR level of speech which the listener can hear, discriminate between words, recognise, and verbally

repeat the word, in quiet or with background noise. In word recognition assessment, phonetic balance is a property of the speech material in which various phonemes occur at relative rate comparable to everyday speech (Martin, Champlin and Perez, 2000). It may be impossible to construct a perfect phonetically balanced speech materials for speech audiometry, and aiming to do so might lead to excessive amount of test time. A phonetically balanced list of speech material may be created by including each initial consonant, vowel, and final consonant with the same rate of occurrence (Lehiste and Peterson, 1959). A study by Martin, Champlin and Perez (2000) demonstrated that there were no significant effects from using phonetically-balanced and non-phonetically-balanced word lists on speech intelligibility performance. However, the speech intelligibility performance may be affected when the word list is spoken by different speakers, even though the list is phonetically-balanced (Killion *et al.*, 2004). The use of phonetically-balance word list is preferable, as it is a more standard material than non-phonetically balanced word list.

2.2.6 Prosodic feature

Prosodic features (e.g. stress, word duration, pauses, and fundamental frequency) are important cues for the listeners to understand speech (Kalikow, Stevens and Elliott, 1977). More intelligible speech normally contains a longer duration and higher number of pauses (slower speaking rate) and longer syllable and vowels duration (stronger stress) (Mayo, Aubanel and Cooke, 2012). These characteristics makes the speech different to “plain” speech (i.e., equivalent to conversation in normal circumstances), and can be generated by the speaker following instructions such as speaking as if communicating with an older adult and putting more effort into speaking. Such speech is usually referred to as clear speech (Uchanski, 2005). It is still under debate on how exactly pauses in speech affects speech intelligibility. While young (18-29 years) normal hearing listeners did not show to benefit from different pauses duration in speech (Krause and Braida, 2002), older adults (>60 years) or people with hearing impairment showed better speech intelligibility when pauses are slightly longer (Tanaka, Sakamoto and Suzuki, 2011).

2.2.7 Contextual cues

Contextual cues refer to the acoustic and phonetic features (e.g., formants) which the listener perceives and uses to comprehend speech (Rogers, Jacoby and Sommers, 2012). Formants in speech are frequencies with high acoustic energy as a result of vocal tract

acoustic resonance (Story and Bunton, 2015). Older adults rely on the contextual cues to understand speech more than young adults, and they appear to benefit from the context cues differently from young adults where the intelligibility in background babble noise is different (Presacco, Simon and Anderson, 2016). In a listening condition with meaningful background noise (e.g., native language conversation), the context cues of the target speech will be similar to those in the background noise, making the target speech more challenging for comprehension. In contrast, when the background noise is meaningless (e.g. foreign language conversation), the context cues of the target speech could be distinguished from the background noise with less effort and more success (Tun, O'Kane and Wingfield, 2002; Rogers, Jacoby and Sommers, 2012; Lash *et al.*, 2013).

In the context of speech recognition tests, a speech test based on listening to sentences would allow more use of contextual cues than a speech test based on phoneme or word. This is due to the possibility to use down processing, e.g., language experience, to aid the bottom-up processing, e.g., sensory input (Dingemanse and Goedegebure, 2019).

2.3 Behavioural measures of speech intelligibility

Speech intelligibility can be measured through behavioural tests, where subjects may be asked to verbally repeat the sentences they hear in both quiet and with background noise. In some cases, the subject may need to respond by choosing the correct speech material shown on a screen. Many test paradigms are being used in speech intelligibility studies, the methods may differ in terms of the speech material and noise masking. According to the practice guidance by the British Society of Audiology, responsible for setting up audiological protocols and guidance in the UK, speech-in-noise (SIN) tests can be conducted to obtain either a speech reception threshold (SRT) or word recognition scores (WRS). SRT is typically defined as the signal-to-noise ratio (SNR) where the subject can correctly repeat 50% of the speech material they heard (Plomp and Mimpen, 1979). SRT in quiet refers to the condition where words or sentences in SIN tests are presented without background noise. The SNR for SRT in quiet will instead be the intensity level of the presented speech where 50% of the presented speech material is correctly repeated. WRS, may also be referred to as “Isolated Word Recognition” and is the ability to identify monosyllabic words correctly; the score is typically reported as percentage of correctly repeated words from the total test words.

To obtain the SRT, two commonly used methods in SIN-tests are 1.) fixed SNR, and 2.) adaptive procedure (staircase). In the fixed SNR method, the speech materials can be presented at equally spaced SNR levels (e.g., 0, 2, 4, 6, and 8 dB) to obtain SIN scores at each SNR, then fit a psychometric function to estimate the 50% recognition score and its corresponding SNR level as SRT (Kollmeier *et al.*, 2015). In the adaptive procedure, a starting SNR level is selected for the first trial, the SNR level will then be adjusted for the following trial depending on the recognition score. If the score of the previous trial is greater than threshold (e.g., 50 % correct), the SNR for the following trial decreases, and the SNR increases if the score of the previous test is below threshold (Smits *et al.*, 2022). The adaptive procedure offers a more time efficient SRT estimation with a smaller number of trials compared to the fixed SNR method.

2.3.1 Types of speech-in-noise tests

This section provides an overview of different types of speech-in-noise tests commonly used in clinic and research. The SRT can be estimated form either word scoring or sentence scoring.

Primarily, there are two ways to employ the speech in noise test:

1. Maintain the noise at a particular level and vary the intensity of speech.
2. Maintain the intensity of speech and vary the noise level.

2.3.1.1 Quick Speech in Noise Test (QuickSIN)

The QuickSIN (Killion *et al.*, 2004) was developed by Etymotic Research Inc.. The development for this test was aimed at reducing the duration of previous SIN tests and representing well the real-world problem of listening to speech. As given by the name, QuickSIN test could be done within two minutes with easy administering and scoring.

QuickSIN consist of 18 lists containing six sentences each. Each sentence is adjusted to different SNR level: 25, 20, 15, 10, 5, and 0 dB. Sentences are recorded from a female speaker, and four-talker noise is used as the background noise. Sentences are presented at 70 dB hearing level (HL). The unit dB HL is an audiometric measurement referenced to the sound pressure level in the acoustic coupler (Svec and Granqvist, 2018).

The speech can be presented to the subject in two modes:

1. The standard mode, typically used for aided condition assessment, presents both target speech and multiple-talker noise through the same loudspeaker.
2. The split track mode: the target speech and multiple-talker noise are presented through different speakers which are spatially separated.

Each sentence in the list contains five keywords, one point will be given when the subject correctly repeated the word, and half a point will be given when the repeated word is not fully correct but close to the actual word (e.g., when actual word is "bats" but repeated as "bat"). Equation (2.1) shows how the SNR loss is calculated. (Killion *et al.*, 2004) defined SNR loss as the difference between the testing subject SRT and the SRT averaged across normal hearing population. SNR loss is an approximation of the increase in SNR relative to normal hearing people needed to understand about 50% of speech. The maximum point that the listener can score in each list is 30 points (each of the six sentences contains five keywords).

$$SNR\ loss\ (dB) = 25.5 - total\ points\ of\ correctly\ repeated\ words\ (\max 30\ points) \quad (2.1)$$

According to the QuickSIN user manual, the SNR loss can interpret as shown in Table 2.1, to classify the test subject's hearing condition:

Table 2.1. Hearing condition classified according to the SNR loss in dB

SNR loss	Hearing condition
0-3 dB	Normal
3-7 dB	Mild hearing loss
7-15 dB	Moderate hearing loss
>15 dB	Severe hearing loss

2.3.1.2 Hearing in Noise Test (HINT)

The HINT (Nilsson, Soli and Sullivan, 1994) was designed to provide more natural speech material in a SIN test, where much of the tests use spondees (words with two long (or stressed) syllables). Moreover, it was developed to overcome the floor and ceiling effect in speech intelligibility measurement. The floor and ceiling effect is found when most of the listener in the test scored close to the minimum (0-5% intelligibility) or the maximum (95-100% intelligibility) score of the test, respectively (Snik *et al.*, 1997). Though HINT may not be widely available for routine clinical use, it is still often used in research.

The speech material comprises 25 list with 10 sentences in each list, (250 sentences in total) obtained from the Bamford-Kowal-Bench (BKB) British sentence speech material (Bench, Kowal and Bamford, 1979). All sentences were recorded from a male speaker. The sentences are equivalent in terms of length, intelligibility, difficulty, and phonemic distribution. The test measures the threshold level where speech is correctly repeated and is thus comparable to the SNR loss measure in QuickSIN. Dissimilar to QuickSIN, the subject must correctly identify every word in the sentence to obtain a score.

There are three test conditions in a standard HINT protocol, each condition differs in the spatial direction from which speech and constant background noise are introduced to the subject. The three directions are to the front, left, and right of the subject and may differ for speech and noise. In this respect, HINT overcomes one of the limitations of QuickSIN, as it explores binaural cues, as well as the monaural ones used by the latter. HINT is thus able to ascertain the advantage of binaural hearing in speech in noise reception.

2.3.1.3 Matrix sentence tests

The matrix sentence test was designed by Hagerman (1982), and was first made available in Swedish. The test consists of an original list of 10 sentences, each sentence contains 5 words with the following structure, Name-Verb-Numeral-Adjective-Noun (e.g., Peter got three large desks). Ten alternative words are available to be selected for each position (this is the matrix), each word will appear only once in a list of 10 sentences. The words are similar in terms of difficulty and are selected to attain phonetic balance. A new list of sentences can be created by randomly selecting words in each position and constructing new sentences which are different from sentences in the original list. Sentences in the original list were recorded from a female speaker with an attempt to avoid transitions between words with a discrete production of each word.

The noise used in this test was generated using the recorded words in the original list to simulate multiple-talker noise. Periodic noises with different modulation rate between 10 to 30 Hz were produced, these noises were then combined with an independent noise with unperceivable periodicity which has identical spectral content to the speech material.

The matrix sentence test has been utilised in many research areas, such as speech audiometry (Nuesse *et al.*, 2019), hearing rehabilitation (Ronnberg *et al.*, 2016), hearing device testing (Kaandorp *et al.*, 2015), and assessing automatic speech recognition system

(Schadler *et al.*, 2015). Particularly for speech audiometry research, the test has been developed in more than 14 languages (Kollmeier *et al.*, 2015).

The matrix sentence tests provide both open- and closed-set response formats to be administered. An open-set response format is normally used clinically in speech audiometry, where the subject must verbally repeat the speech material correctly to be scored by the test administer. In a closed-set response format, the subject can select a word shown on a screen from a short list. The latter can be self-administered by the subject without the need for the test administer to understand the meaning of the speech material. The closed-set format response also reduces the need of the listener's short-term memory to repeat the words. However, when there is no requirement to repeat the speech material from the closed-set format, the listener still has access to the closed-set words and has the chance of choosing the correct word they hear, even though they did not clearly hear the word. The difficulty in perceiving specific speech feature or phonemes might not be noticed by the test administrator with this paradigm (Clopper, Pisoni and Tierney, 2006).

The behavioural SIN tests for measuring speech intelligibility in noise has been explained in this section, the next section will be a review of literature on how the ability in hearing (measured through pure-tone audiometry) is related to speech intelligibility.

2.4 Relationship between the pure-tone audiogram and speech reception threshold

PTA is the most used clinical measure for diagnosing hearing loss; unfortunately, it does not assess how well a person can understand speech in a noisy environment. Some normal hearing people with a nearly normal audiogram and speech comprehension in quiet, report to audiologist that they have difficulties to comprehend speech in noise. Several studies have found that the audiogram does not correlate well with the behavioural measure of speech intelligibility for normal hearing people (Middelweerd, Festen and Plomp, 1990; Smoorenburg, 1992; Ferman, Verschuure and Van Zanten, 1993). Possible causes for this discrepancy are related to central auditory processing disorders (APD), a condition which sound is processed abnormally in the brain (Middelweerd, Festen and Plomp, 1990; Bamiou, Musiek and Luxon, 2001). In contrast, the correlation between the audiogram and SRT becomes stronger for subjects with hearing impairment. However, this relationship is complicated, the relation between the audiogram and SRT is best described when referenced to a specific type of hearing loss (Smoorenburg, 1992). For example, for people

with sensorineural hearing loss, hearing loss caused by a damage of hearing organs in the inner ear and the audiogram indicates hearing loss from around 0.5 to 6 kHz, the audiogram at 2 and 4 kHz is found to be strongly correlated with SRT ($r = 0.72$).

In studies on relating the audiogram to SRT, the audiogram in the 0.5, 1, 2, 3, and up to 4 kHz frequency bands are most correlated with the SRT (Curry, 1949; Fletcher, 1950; Harris, Haines and Myers, 1960). When relating the audiogram to SRT, it is widely accepted to use the average value of the audiogram from three frequency bands to find any correlation with the SRT (i.e., average value of the audiogram at 0.5, 1, and 2 kHz), this is called the "three-frequency average" (Smoorenburg, 1992). SRT in quiet has been found to be most related to audiogram in the 0.25, 0.5, 1, and 2 kHz frequency bands, while SRT in noise is best correlated to 2, 3, and 4 kHz frequency bands (Hagerman, 1984; Smoorenburg, 1992). It is suggested that the relation between audiogram below 2 kHz and SRT in quiet could be explained by the ability to perceive sound (Noordhoek, Houtgast and Festen, 2001). The SRT in noise is associated with the lessened ability to distinguish words and phonemes due to hearing loss at 2-4 Hz (Hagerman, 1984). These results showed that hearing in high frequency (from 500 Hz up to 4 kHz) plays an important part in speech in noise comprehension.

Both PTA and the speech intelligibility test are done behaviourally. The downside of the behavioural measurement is that the subject must be able to respond to the test, thus this test could not be done in some groups of patients (e.g. infants, young children, people with language disorders, and people with cognitive impairments) (Purcell *et al.*, 2004). In the case of infants and young children, it is well known that early diagnosis and treatment of problems related to hearing and speech is essential for improving their later quality of life. However, it is impossible to retrieve reliable behavioural measurements from children due to their limited cognitive and linguistic skill. People with language disorders not only have difficulties in understanding speech but they may also have problems in producing the correct speech sound and their speech might not be correctly understood (Bishop and Snowling, 2004). Through the behavioural measure of speech intelligibility, it may not be possible to tell whether the person has difficulty with language processing or word production, or both. For people with cognitive impairments, tasks involving memory, decision making, and attention will be unfavourable and their attentiveness would decline throughout the test (Yanhong, Chandra and Venkatesh, 2013).

Due to the limitation of behavioural measurements described in the previous paragraph, the objective measurement of AERs (an alternative method for hearing test) is proposed by researchers as a possible predictor of behavioural measurement of speech intelligibility. The studies using AERs as a predictor of speech intelligibility will be reviewed in the next chapter.

Chapter 3 Cortical responses to sound

In the previous section, the relationship between PTA and SRT was described, but due to the limitation that behavioural measures cannot be carried out reliably in some groups of patients, objective electrophysiological measures may be required to try and predict SRT. Firstly, this section will explain the auditory evoked responses (AERs). This will be followed by descriptions of the method for measuring AERs to repeating short and continuous speech stimuli. Finally, this chapter summarises of the findings on how subcortical and cortical auditory evoked responses have been used to predict speech intelligibility.

AERs include auditory brainstem responses (ABR), auditory middle latency responses (AMLR), auditory steady-state responses (ASSR), and cortical auditory evoked potentials (CAEP), and all of these except AMLR have been used for predicting the behavioural measure of speech intelligibility. AERs to both speech and non-speech stimuli are described in this chapter.

3.1 Auditory evoked responses

AERs are usually elicited by inserting a transducer that transmits sound into the ear. The stimuli could be presented either through a headphone, insert phone, bone conductor, or a loudspeaker in a free field (when there is no or minimal sound reflection). As the stimulus linked activity progress through the auditory pathway, they elicit electrical activity from each stage of auditory processing; starting from the cochlea, auditory brainstem, midbrain and auditory cortex. These responses can be measured non-invasively using electrodes placed on the ear canal or most commonly, the scalp, i.e., via the electroencephalogram (EEG).

AERs are naturally low in SNR because these responses are a fraction of numerous electrical activities in the human brain which form background noise. The signal quality of AERs also suffers from muscle artefacts from the participant and electrical noise from equipment. These noises recorded in the EEG (other brain electrical activities and artefacts other than the response from the auditory system) may have relatively larger amplitude than the AERs themselves (Rompelman and Ros, 1986). AERs may not be present or data analysis can mislead, if the SNR of EEG is too low. In the pre-processing step, a number of

methods for reducing noise includes differential amplification, filtering, artefact rejection and finally signal averaging.

3.1.1 Differential amplification

Differential amplification in EEG recording consist of two input electrodes, a non-inverting and an inverting electrode placed on the head (vertex) and neck (ipsilateral neck) respectively (Beattie, 1988). The differences between signals from the two electrodes are amplified, while the common signals (same voltage change in both inputs relative to ground) are attenuated. The use of differential amplification for recording AERs is advantageous when noise is presented in both electrodes and the AERs signal in interest is mainly detected by one electrode. More detailed technical guidelines for the use of differential amplification in electrophysiological recordings can be found in Brigell *et al.* (2003).

3.1.2 Filtering

The purpose for filtering the EEG is mainly to suppress noise by attenuating components that are out of the frequency range of interest, thus increasing SNR. Slow drifts and power line noise are often seen in EEG signals and can easily be removed by filtering. Slow drifting occurs due to the change in skin potential or poor contact between electrode and skin. The slow drift can be removed using a high pass filter with a cut off frequency higher than the frequency of the drifting wave. Power line noise can also be suppressed by simply applying a notch filter, however, the 50 or 60 Hz frequency might be important for some AERs at the cortical level (de Cheveigne and Nelken, 2019). The use of filters should be carefully considered whether the AERs components of interest will be affected or not.

Filters can also be used to further emphasise specific EEG frequency bands, e.g., delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), gamma (30-80 Hz) (Tsipouras, 2019). Each EEG frequency band is considered suggested to be associated with different cognitive processes, e.g., perception, memory, and information processing (Klimesch, 1999). In AERs research, sub-band EEG analysis can be performed to explore whether certain aspects of auditory perception, peripheral hearing or comprehension, are more related to some particular EEG frequency bands (Ding and Simon, 2014; Etard and Reichenbach, 2019).

3.1.3 Artefact rejection

Artefacts in EEG can be caused by biological, instrumentation, or environmental factors. Biological factors, such as eye blinking, swallowing, and heart beating, appears to be the most challenging as they are unavoidable, and the severity of artefacts varies depending on the subject's behaviour. Moreover, the spectral bandwidth of this type of artefact, approximately 20-300 Hz, tends to overlap with the frequency of cortical auditory activities (Muthukumaraswamy, 2013). For recorded EEG, prior to the signal averaging, segments of the EEG signal containing auditory response (epochs) holding samples with amplitude greater than the threshold level can be simply removed (Golding *et al.*, 2009). This strategy can also be applied during online recordings to obtain desired number of artefacts free epoch. One might consider using computational methods, such as independent component analysis, to remove the artefacts which are presented across the raw EEG (Comon, 1994).

3.1.4 Signal averaging

A vital process for AERs measurement is signal averaging, as the AERs cannot be observed directly from the recorded EEG. Tens, hundreds, or thousands of stimuli are presented repeatedly into the ear and the EEG from the repetition of stimulus is stored to average out the noise, as shown in Figure 3.1. Following each stimulus presentation, each response is expected to contain dominant peaks that is hindered in noise occurring at the same time delay (latency) and with a similar amplitude; noise is considered to arise randomly, regardless of the timing of the stimulus. Greater number of responses from the subject being included in the averaging process will result in a waveform of AERs with greatly reduced noise. This averaging process is called coherent averaging, where the epochs synchronised with the trigger point of the stimuli are averaged down the columns (Tagare, 1987).

Equation (3.1) shows an epoch array (ensemble) containing multiple epochs in a 2-dimensional matrix X with each epoch in a separate row.

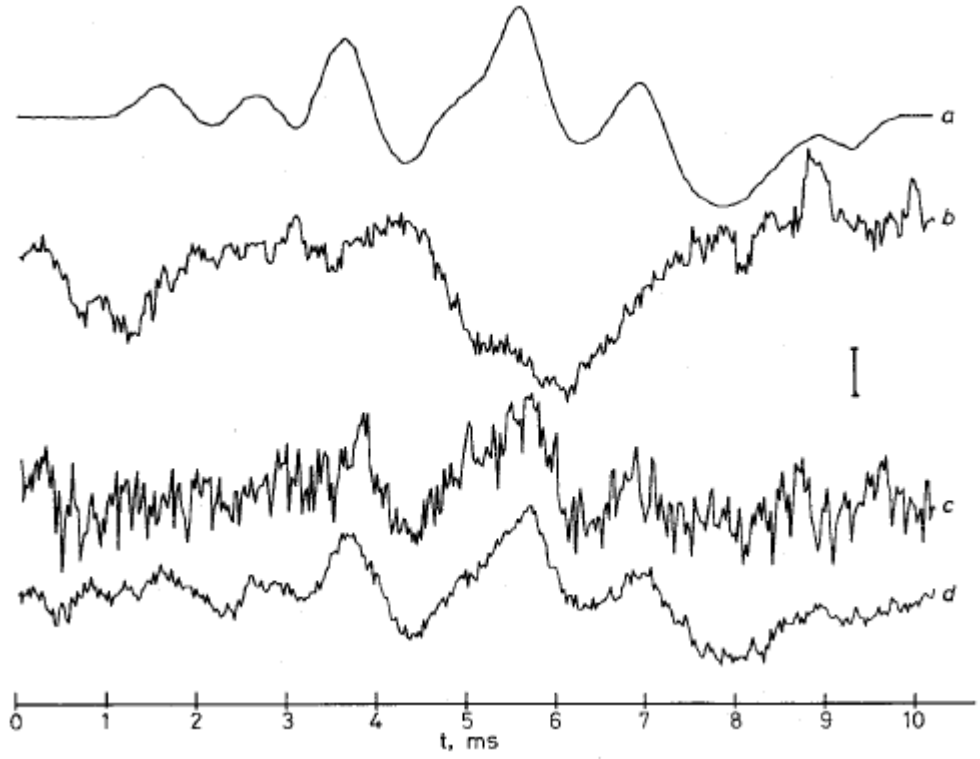

Figure 3.1. (a) A noise free auditory brainstem response at 80 dB Hearing level. (b) Auditory brainstem response from 1 epoch, (c) average of 100 epochs, and (d) average of 2,048 epochs. *Reprinted from Kaplan, R., Özdamar, Ö. Microprocessor-based auditory brainstem response (ABR) simulator. Med. Biol. Eng. Comput. 25, 560–566 (1987). With permission from Springer.*

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NK} \end{bmatrix} \tag{3.1}$$

Where K is the number of samples per epoch and N is the total number of epochs. From the matrix X, x_{11} represents the first sample from the first epoch and x_{NK} is the last sample from the last epoch. The first column is where the trigger points (typically the time of a stimulus) occur. The coherent averaging is obtained by coherently averaging down each column separately, resulting in a coherent average waveform with K samples.

3.1.5 Auditory brainstem responses

The auditory brainstem response (ABR) occurs within about 10 ms after stimulation (Sharma, Bist and Kumar, 2016). Major components in ABR are generally labelled with Roman numerals (I to V) as shown in Figure 3.2. Wave I is the response generated from distal part of the eighth cranial nerve, as this wave is a response leaving the cochlea. Wave II is also a response from the eighth cranial nerve but from the proximal portion, which is

the part that leads to the brainstem (Starr *et al.*, 2010). Wave IV is generated from the brainstem but this wave is less important to the clinicians compared to other waves, as it is only observed as a contribution to wave V or as the wave IV/V complex (Hall, 2007). Wave V is the most important feature in ABR that is analysed by clinicians. The latency of the peak is about 5-7 ms from the stimulus. The origin of wave V is from the inferior colliculus, with contributions from the termination in the lateral lemniscus (Porter and Tharpe, 2010).

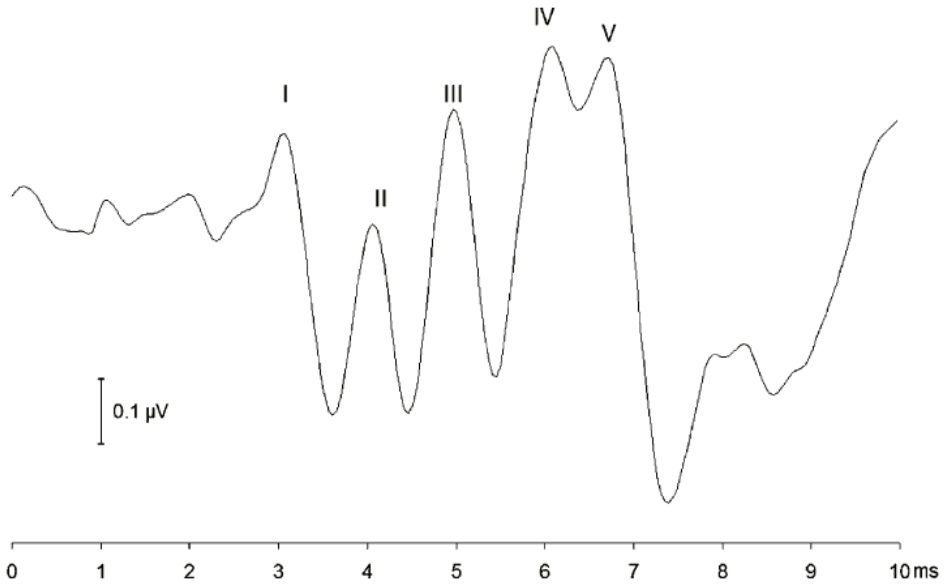

Figure 3.2. Auditory brainstem response waveform. *Reprinted from Kraus, N., Nicol, T. (2008). Auditory Evoked Potentials. In: Binder, M.D., Hirokawa, N., Windhorst, U. (eds) Encyclopedia of Neuroscience. Springer, Berlin, Heidelberg. With permission from Springer.*

ABR can be evoked by numerous types of stimuli, which includes clicks, chirps, tones, monosyllable speech and continuous speech. One of the differences between ABR to speech and non-speech stimuli is its application. ABR to non-speech stimuli are generally used in hearing screening and hearing threshold measurement. Speech-ABR (Figure 3.3) consist of onset response and transition components (frequency-following response (FFR) and offset response) (Anderson and Kraus, 2010; Sinha and Basavaraj, 2010). The response additionally reflects some cortical activity which is the FFR, making it potentially useful in diagnosing phonological disorders (an inability to discriminate distinct speech sound which can result in an error in speech production (Edwards, Fox and Rogers, 2002)) and cognitive impairments. The FFR region of speech-ABR is the part where the ABR is phase-locked to the temporal fluctuation of the speech stimuli (Bidelman, 2015). An example of ABR in response to a monosyllable speech stimulus (/da/) is shown in Figure 3.3.

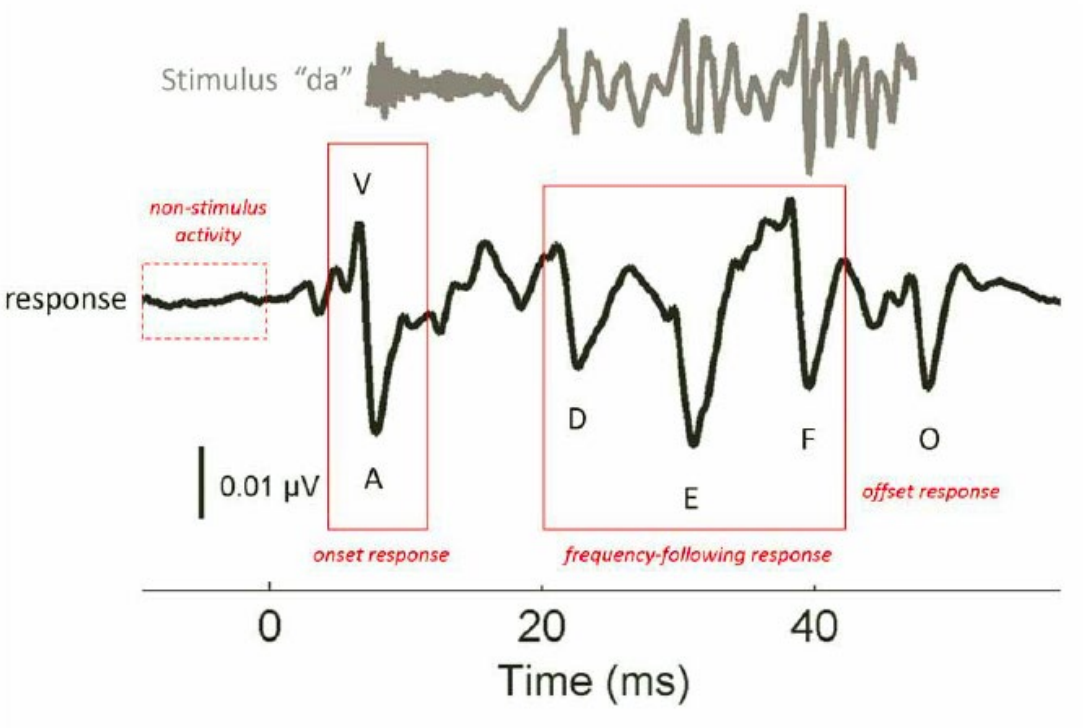

Figure 3.3. ABR in response to /da/ stimulus. The /da/ stimulus waveform is shifted by about 6.8 ms to the right, to help visualise the coherence between the stimulus and the frequency-following response. *Reprinted from Skoe E and Kraus N (2013) Musical training heightens auditory brainstem function during sensitive periods in development. Front. Psychol. 4:622. Used under Creative Commons CC-BY license.*

3.1.6 Auditory middle latency responses

Auditory middle latency responses (AMLR) waveform consist of Na, Pa, and Pb peaks as main components, occurring within 20 to 70 ms post stimulus onset (Bell *et al.*, 2004) (see Figure 3.4). AMLR compose activities from both subcortical and cortical generators (Musiek and Nagle, 2018). A few studies relating AMLR and behavioural performance observed no significant correlation between the two measures (Paludetti *et al.*, 1991; Makhdoum *et al.*, 1998) in normal hearing people and cochlear implant users. A more recent study by Alemei and Lehmann (2019) also report the same finding from the same group of subjects.

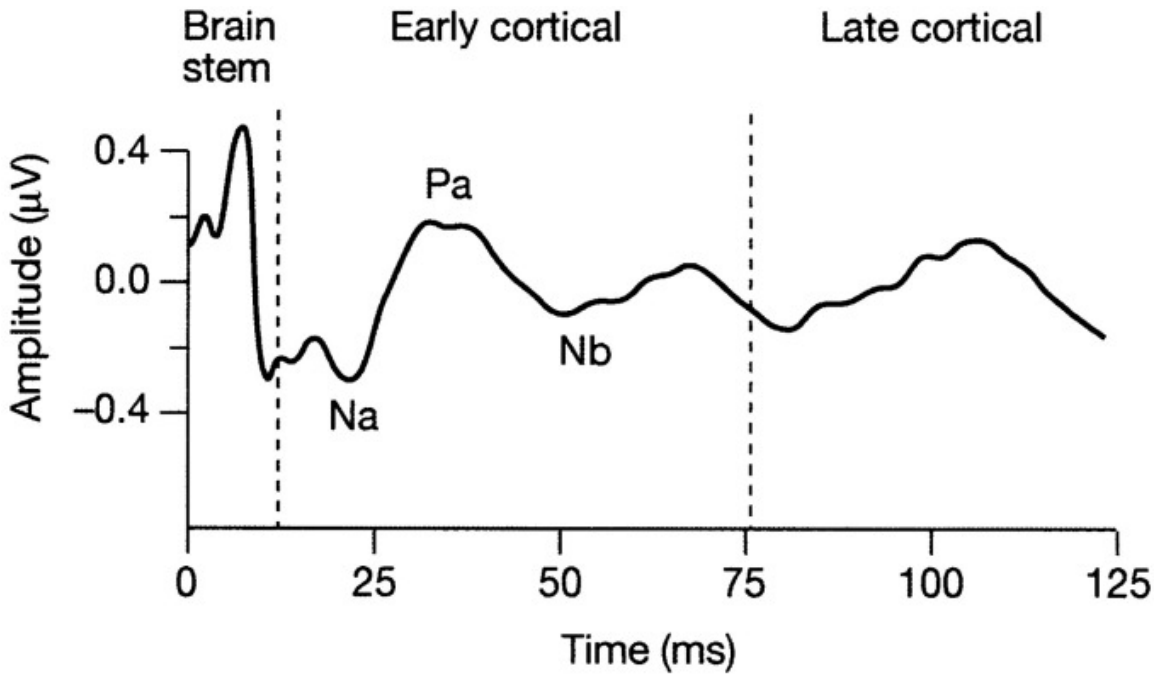

Figure 3.4. Auditory middle latency responses waveform shown as early cortical responses. *Bell, S.L. et al. (2004) 'Recording the middle latency response of the auditory evoked potential as a measure of depth of anaesthesia. A technical note', British Journal of Anaesthesia, 92(3), pp. 442-445. With permission from British Journal of Anaesthesia.*

3.1.7 Auditory steady-state responses

Auditory steady-state responses (ASSR) are typically generated by amplitude modulated (AM) tone stimuli. The AM tone stimuli comprise of carrier frequency (CF) tonal stimulus, normally high frequencies (e.g. 0.5 or 1 kHz) and a modulation frequency (MF) that is much lower (e.g. 40 or 80 Hz). ASSR can be interpreted in both time and frequency domain. In the time domain, ASSR will show a temporal fluctuation corresponding to the MF. For example, in Figure 3.5 (A, top), the inter-peak latency in the time domain is approximately 25 ms for ASSR to a 40 Hz MF. In the frequency domain, a strong power spectrum is expected to show at 40 Hz and its harmonics, 80 and 120 Hz (Bidelman and Bhagat, 2016). Different stimulus modulation rates generate ASSR from distinct sources along the auditory pathway (Luke, De Vos and Wouters, 2017). For example, ASSR to 40 Hz AM tone stimuli mostly arises from thalamic source, while ASSR to 80 Hz AM tone stimuli has its origin in the auditory brainstem.

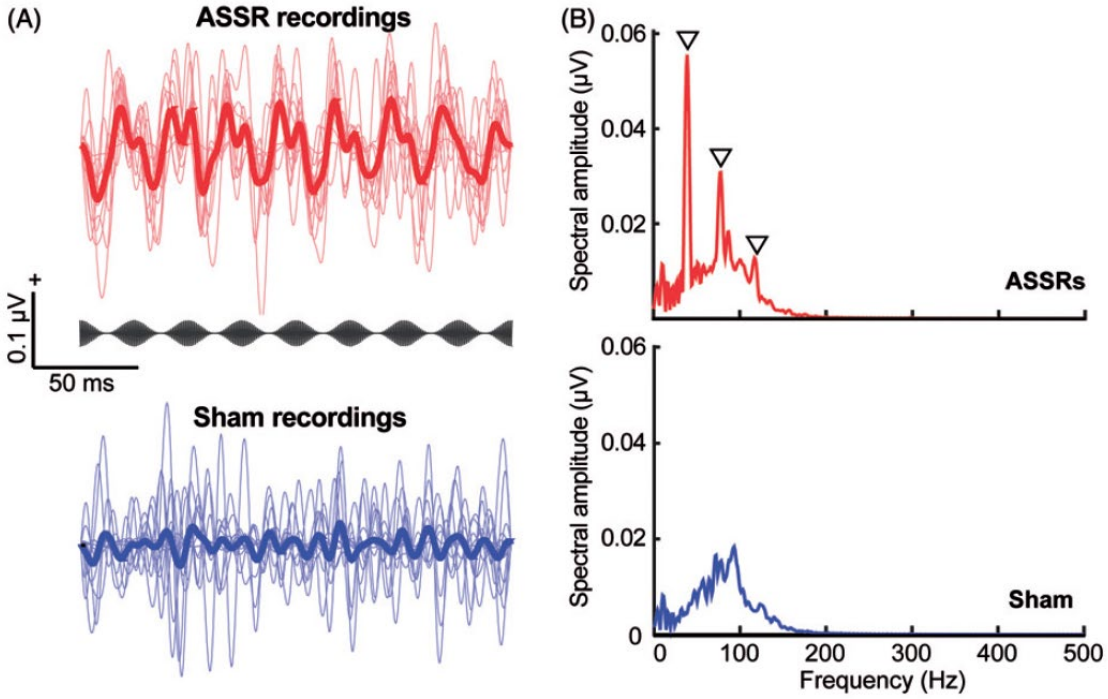

Figure 3.5. (A, top) Auditory steady-state responses (ASSR) to a 40 Hz sine wave modulating a 1 kHz tone. (A, bottom) Response with no auditory stimulation. (B, top) Averaged response spectra for ASSR. (B, bottom) Averaged response spectra with no auditory stimulation. Bidelman, G.M. and Bhagat, S.P. (2016) 'Objective detection of auditory steady-state evoked potentials based on mutual information', *International Journal of Audiology*, 55(5), pp. 313-319. W*ith permission from Taylor & Francis.*

3.1.8 Cortical auditory evoked responses

The slowest AERs are the cortical evoked potentials (CAEPs) or auditory late response (ALR), responses generated from the auditory cortex, with the latency of about 50-300 ms (see Figure 3.6). Major components of CAEPs are labelled as P1, N1, P2, and N2 (Figure 3.6). The main components, N1 and P2, occurs within about 75 to 150 ms and 150 to 200 ms respectively after the stimulus (Kerr, Rennie and Robinson, 2008; Zhang, Gong and Zhang, 2016). The amplitude of CAEPs, negative peak N1 to positive peak P2, found in the literatures is typically about 3 to 4 µV. Components of CAEPs are sorted into two groups, exogenous and endogenous (Leite *et al.*, 2018): exogenous components are the components that are elicited by the attributes from the stimuli which are P1, N1 and P2. Endogenous responses are determined by the interaction of the subject with the stimulus involving some cognitive processing the subject may be required to perform a task on the stimuli (e.g. attending or ignoring odd stimuli). This type of responses may be reflected in peaks occurring 300 ms, post stimulus onset, e.g., P3 peak (de Melo *et al.*, 2016).

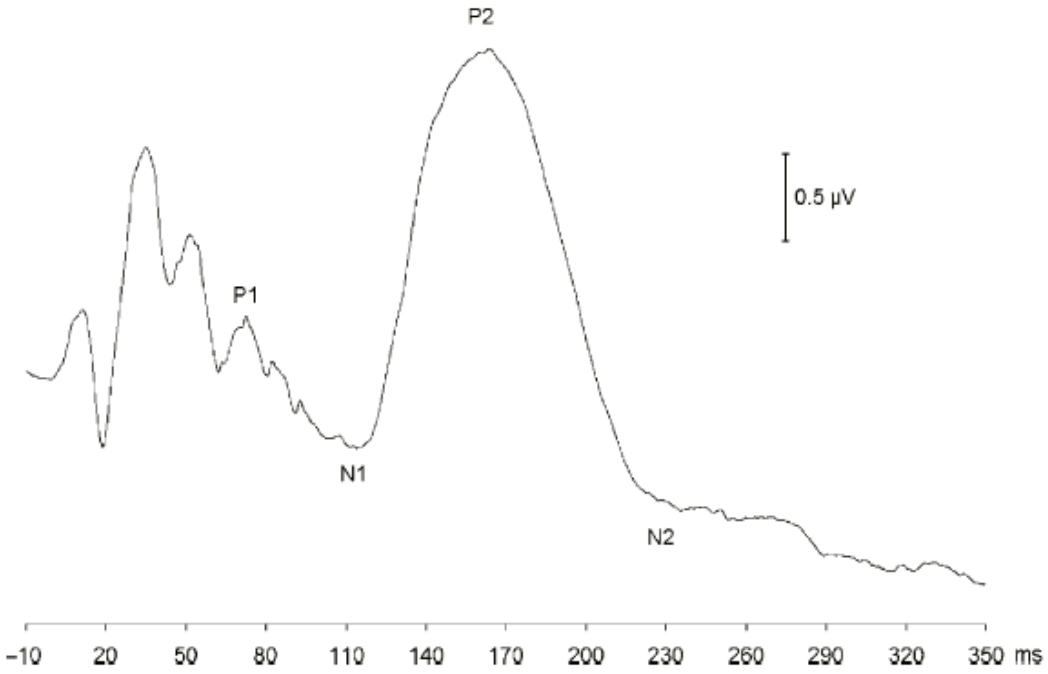

Figure 3.6. Cortical auditory evoked potentials waveform. *Reprinted from Kraus, N., Nicol, T. (2008). Auditory Evoked Potentials. In: Binder, M.D., Hirokawa, N., Windhorst, U. (eds) Encyclopedia of Neuroscience. Springer, Berlin, Heidelberg, with permission from Springer.*

3.2 Detection of conventional auditory evoked responses: Hotelling's T-squared

Inspecting AERs visually can be dependent on the audiologist's experience. To deem whether AERs is present or not can be most challenging when responses are generated by stimulus with intensity near hearing threshold, the conclusion can be variable across different examiners (Lv, Simpson and Bell, 2007). Objective methods based on statistical tests can be used to assist the audiologist to determine the presence or absence of AERs, providing more consistent result across examiners. The false positive or negative detection of AERs, i.e., incorrectly determine a response is present or absent, can also be controlled with the use of objective methods (Golding *et al.*, 2009; Chesnaye *et al.*, 2018).

The one-sample Hotelling's T-squared (HT^2) is an objective AERs detection method that is widely used and it is the main tool for detecting CAEPs throughout this thesis. The method is used to test whether the mean vector from the epoch array is significantly different from zero. In the absence of a stimulus response (i.e. under the null hypothesis) this is expected, because the direct current (DC) offset (mean amplitude displaced from zero) is removed from AERs by high-pass filtering (Golding *et al.*, 2009). The HT^2 was suggested to be the

best among several detection methods in terms of sensitivity and detection time for ABR (Chesnaye *et al.*, 2018). In the time domain analysis, the multiple variables are obtained by calculating the 'time-voltage' means (TVMs), an average of amplitude change over a short interval of the CAEPs. The TVMs are then tested for statistically significant difference form zero.

It has been suggested to be cautious in choosing the TVM windows, bins that are too long might cover both positive and negative peaks, and as a result the mean value in the bin will be close to zero, making the test insensitive in detecting responses. Making the TVM window too short will lower the sensitivity of the statistical test (Golding *et al.*, 2009; Chesnaye *et al.*, 2018), as too many variables are included in the test. A suggested TVM window length from the previous study by Golding *et al.*, 2009 was 50 ms. The method was focused on detecting the onset response (detecting N1 and P2), which is within the first 500 ms of the data, which is the analysis window, the interval where the TVMs will be calculated.

The HT^2 test statistic is obtained from the equation (3.2).

$$T^2 = N(\bar{x} - \mu_\circ)S^{-1}(\bar{x} - \mu_\circ)^{\mathrm{H}} \tag{3.2}$$

Where:

- N is the number of epochs for the test (rows in the epoch array as shown in equation (3.1) in section 3.1.4)
- $\bar{x}$ is the mean vector of the epoch array
- $\mu_\circ$ is the hypothesised mean vector (zero for evoked potentials)
- S^{-1} is the inverse covariance matrix of the variables (TVMs)
- The superscript $^{\mathrm{H}}$ indicates the Hermitian matrix

Note that the number of rows in the epoch array for the test must be greater than the numbers of columns to be able to calculate the covariance matrix.

The T^2 is then converted to F-statistic with the degrees of freedom $v1$ and $v2$ by using the equation (3.3).

$$F_{v1,v2} = \frac{N-K}{K(N-1)}\,T^2 \tag{3.3}$$

Where:

- $v1 = K$
- $v2 = N - K$

- K is the number of variables for the test (columns of TVMs)

The F value with the degrees of freedom of $v1$ and $v2$ is then tested for its significance from the F-distribution table. If the p-value is ≤ 0.05, the H0 will be rejected, i.e., assumption that the mean value of TVMs is zero and thus that there are no responses present, is rejected and a response is deemed to have been detected.

3.3 Cortical responses to continuous speech

During the last decade, many studies on CAEP to speech have focused on using continuous speech stimuli. This type of stimulus is more natural and intelligible compared to the conventional transient stimuli, as it more closely represents the real-world problem. This allows researchers to study the relationship between cortical auditory processing and cognitive processing (e.g., attention to target speech).

Continuous speech stimuli are more complex than conventional transient stimuli. Unlike the repeating transient stimuli, the amplitude and latency of each syllable or word in the continuous speech waveform changes over time. Therefore, would be suboptimal to extract the AERs using the coherent average process, as there are no synchronisation of syllable or word onset at a specific occurrence rate. Studies commonly use a mathematical model based on a system identification method to find a linear relationship of the AERs to some acoustic features of speech. The speech envelope is the most selected feature of the speech to relate with the AERs.

One of the best-established tools for measuring human response to speech stimuli is the Multivariate Temporal Response Function (mTRF) (Crosse *et al.*, 2016). As it is known that the EEG can entrain to the envelope of speech (Aiken and Picton, 2008b), the mTRF was design to measure the strength of envelope entrainment. Two approaches can be employed with the mTRF (Figure 3.7), either encoding (predicting the EEG using the speech envelope, forward TRF) or decoding (reconstructing the speech envelope from the EEG, backward TRF). While the former follows the causal psychophysiological process of speech driving EEG entrainment, the latter has practical advantages in allowing multiple EEG channels to be analysed simultaneously and thus potentially permits more powerful analysis of the association between speech envelope and a set of EEG signals than repeated single channel analyses. An outline of the backward TRF is given below in section 3.3.1 and details of the implementation used will be shown in the methods section of each study (chapter 4, 5, and 6). The backward TRF not only reflects auditory responses at the cortical

level, but has also been shown to be able to predict speech intelligibility (Vanthornhout *et al.*, 2018).

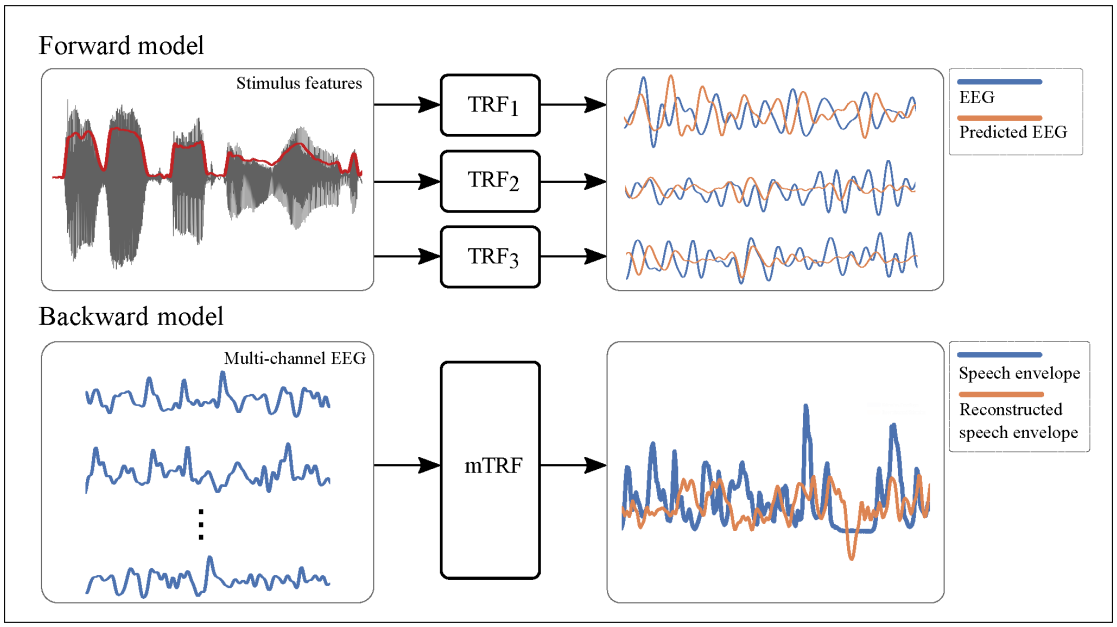

Figure 3.7. The temporal response function estimation in the forward and backward modelling approach. The forward modelling approach or encoding model (forward TRF) uses a stimulus feature (i.e., speech envelope plotted in red) to predict the EEG response to a stimulus (here three TRFs and their associated EEG channels are shown). The backward modelling approach or decoding model (backward TRF) uses EEG response to reconstruct the stimulus feature. *Adapted from Crosse MJ, Di Liberto GM, Bednar A and Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Front. Hum. Neurosci. 10:604, used under Creative Commons CC-BY license.*

3.3.1 Backward modelling approach

The backward TRF, also known as the stimulus reconstruction or decoding approach, is a model-based approach which utilize a linear model ($g(\tau, n)$ in equation (3.4)), to reconstruct the speech envelope from the EEG response signal (Crosse *et al.*, 2016), using multichannel convolution:

$$\hat{S}(t) = \sum_{n} \sum_{\tau} r(t+\tau, n) g(\tau, n) \tag{3.4}$$

where $\hat{S}(t)$ refers to the reconstructed speech envelope and $r(t+\tau, n)$ to the EEG signal, and t is the time index, τ is the range of time lags in the convolution (corresponding to model order), and n represents the channel of the EEG. Equation (3.4) represents the so-called 'inverse model', since in reality the speech signal causes changes in the EEG, but in

this model, the speech envelope is estimated from the EEG using a non-causal filter (hence the + sign in $r(t+\tau, n)$). The inverse model is calculated using the regularised least squares method,

$$\min e = \sum_{t} \left[S(t) - \hat{S}(t) \right] \tag{3.5}$$

$$g = (\boldsymbol{R}^T \boldsymbol{R} + \lambda \boldsymbol{M})^{-1} \boldsymbol{R}^T S \tag{3.6}$$

where $\boldsymbol{R}$ is the EEG signal in lagged time series (in matrix form), and S the stimulus envelope (vector). $R^T R$ and $R^T S$ are the auto-correlation of the EEG, and the cross-correlation of the EEG and a parameter of the stimulus (usually the envelope of the speech stimulus), respectively, across all EEG channels and time lags. Conventionally, the model is evaluated through the Pearson's correlation coefficient and mean-squared error (MSE) between the estimated and actual parameter of the stimulus.

The Tikhonov regularisation, also called ridge regression, is denoted by $\lambda \boldsymbol{M}$. The regularization matrix $\boldsymbol{M}$ is chosen to reduce the off-sample error (i.e. the error in predicting $\hat{S}(t)$ on previously unseen data) (Lalor *et al.* (2006) , as follows (missing terms are zero):

$$\boldsymbol{M} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix} \tag{3.6}$$

Moreover, the regularisation is used to prevent overfitting. An issue with a reverse correlation analysis on biological data, explaining the statistical relationship between the preceding stimuli and sensory response using a linear model (Ringach and Shapley, 2004), is that the estimated model can present properties that are not likely to be seen in biological data (e.g. high frequency fluctuations). The decoder is then estimated and unintentionally overfitted to those data, resulting in a suboptimal estimation of unseen data (data that were not used for building the decoder model). The regularisation reduces overfitting by diminishing the large differences between adjacent values of the decoder model, making the model more generalisable and more optimal to estimate unseen data (Crosse *et al.*, 2016).

The identity matrix can be replaced by the regularisation matrix M to reduce the out-of-sample error (estimation error when estimating on a new data set), used to measure the performance of the model in predicting unseen data, resulting in a smaller estimation error.

The lagged time series of the AERs is denoted by R, where the AERs of the channel n at sample point t = 1, 2, …, $T-1$, T is denoted by $r(t,n)$. The matrix R below shows the lagged time series from a single-channel EEG. In a multiple EEG channel case, each column of R, representing an individual time lag in samples (τ_{max} and τ_{min} as the maximum and minimum time lag respectively), will be replaced by N columns of AERs from N EEG channels at each time lag. In this multivariate case, the R matrix will result in the dimension of $T \times (\mathrm{N} \times \tau_{window})$. Where τ_{window} is the mTRF estimation window, calculated from the difference between τ_{max} and τ_{min}.

$$\mathrm{R} = \begin{bmatrix} r(1-\tau_{min},1) & r(\tau_{min},1) & \cdots & r(1,1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & r(1,1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & r(1,1) \\ r(T,1) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & r(T,1) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r(T,1) & r(T-1,1) & \cdots & r(T-\tau_{max},1) \end{bmatrix} \quad (3.8)$$

S is defined as the column-wise vector of the stimulus representation.

$$\mathrm{S} = \begin{bmatrix} s(1) \\ s(2) \\ s(3) \\ \vdots \\ s(T) \end{bmatrix} \quad (3.9)$$

3.3.2 Leave-one-out cross-validation

Leave-one-out cross-validation is an approach to validate the inverse model. As an example, the method will begin by segmenting the speech stimuli and EEG responses into k segments, such that the speech stimuli and EEG response in each segment correspond to each other. Suppose that the speech stimuli and the EEG response were divided into four segments: in the 1st iteration, data in segment 1-3 will be used to estimate a decoder model. This decoder model will then be tested with the remaining data in segment 4, which will

result in two validation metrics, Pearson's correlation coefficient and MSE. In the 2nd iteration, data in segment 1, 2, and 4 will be used to estimate the inverse model and the averaged inverse model will be tested with data in segment 3. This process will be repeated until the 4th iteration, such that all data segments are used in testing, resulting in four Pearson's correlation coefficient and four MSEs, one from each iteration. The validation metrics will then be averaged across the four trails. The inverse model fitting can be optimised in terms of generalisability on testing data, by adjusting the regularisation values (λ), normally by estimating the inverse model over a range of λ and selecting the model which gives the lowest overall estimation error (highest correlation coefficient).

3.4 Auditory evoked responses as an objective measure to predict speech-in-noise performance

The previous section (section 3.3) provided an overview of the auditory evoked responses and the methods for detecting responses to both repeating and continuous stimuli.

This section will provide an outline of studies relating auditory evoked responses to behavioural speech-in-noise performance. Section 3.4.1 to 3.4.3 are dedicated to the conventional AERs, which includes ABR, ASSR, and CAEP. Section 3.4.4 is dedicated to the more novel cortical response to continuous speech.

3.4.1 Auditory brainstem responses

3.4.1.1 Click-ABR

The Wave I ABR component has been observed to be associated with the degree of neural loss within the cochlea for animals and humans (Kujawa and Liberman, 2009; Makary *et al.*, 2011). It is hypothesised in the study by Bramhall *et al.* (2015) that the ABR wave-I would correlate with the behavioural measurement of speech intelligibility in noise. Fifty-seven adults (22 females) age 19-90 years old (mean 48.6 years) were included in the study. Participants showed normal hearing to mild hearing loss within the frequency range of up to 1 kHz (PTA 0-40 dB HL) and normal hearing to severe hearing loss (PTA 0-80 dB HL) within the frequency range of 2-8 kHz. The QuickSIN test was utilised to obtain the speech perception score. Results from the study indicate that the decrease in ABR wave-I amplitude is correlated with the increasing age and poor performance in the QuickSIN test in noisy conditions but not in quiet. A later study further found that the longer latency between ABR wave I and V peaks is correlated to poorer speech intelligibility (Valderrama

et al., 2018). However, the click-ABR lacks the frequency following responses (FFR) which appears only in ABR to a more speech-like /da/ stimulus, so it is unclear whether speech intelligibility in noise can simply be related to click-ABR as it may only reflect hearing ability and cannot be used to verify encoding of speech sound (consonants and vowels) in the auditory system.

3.4.1.2 ABR to short speech stimulus

FFR in speech-ABR are generated by vowels, such as /a/ and /i/, this response is a distinct component which is absent in ABR to clicks or tone burst. Tone bursts are sinusoidal waveform which has a rise-time before reaching a certain amplitude level, a plateau region where the amplitude is constant for a specific time interval, and a fall-time where the amplitude decreases (Lewis and Henry, 1989). FFR can either be represented in the time or frequency domain which can be referred to as envelope-FFR and spectral-FFR respectively (Aiken and Picton, 2008a). In the time domain, specifically in the FFR region of speech-ABR, the SNR of a stimulus tends to affect the latency of temporal features rather than the amplitude. This phenomenon is known as 'phase-locking' where the neural activity is synchronised with the oscillation part, envelope or fine structure, of the stimuli (Plyler and Ananthanarayan, 2001; Akhoun *et al.*, 2008). In the frequency domain, encoding of speech fundamental frequency (F0), speech harmonics or vowel frequency is indicated by strong spectral components at the corresponding frequency bands.

Speech intelligibility is suggested to be correlated with either envelope- or spectral FFR (Anderson and Kraus, 2010). Anderson and Kraus (2010) observed a significant delay in the latency of the time domain peaks in the envelope-FFR of speech-ABR evoked by a monosyllable /da/ stimulus, for adults with poor SIN perception (with normal audiogram). The same study also found that the magnitude of spectral-FFR at F0 and second harmonic (H2) significantly correlates with the behavioural measures of speech intelligibility for adults. Later they found the same result for children (Anderson *et al.*, 2010b). FFR has also been used for detecting deficits in speech encoding in children with learning problems (Cunningham *et al.*, 2001).

Studies have shown that speech-ABR reflects the encoding of simple sound to more complex speech-like sound by the auditory system, trough onset/offset responses, modulation responses, and phase locking at low frequency (F0). An advantage of using ABR over later responses, such as CAEP, is that the hearing function and encoding of speech sound can be assess early in young children (5 years old) (Milaine Dominici *et al.*,

2019), due to earlier maturity of auditory brainstem relative to auditory cortex. The subcortical hearing function is indeed important for SIN perception, however, the variability of behavioural speech intelligibility can be associated with the aspect of cognitive ability (e.g., attention and working memory) to process and comprehend speech, which the cortical activity related to SIN perception may not be captured at the subcortical level.

3.4.2 Auditory steady-state responses

Temporal envelopes, one of the critical temporal features of sound, represent the low frequency fluctuation of the stimulus amplitude which correspond to the change of intensity and length of syllables, words, or phrases (Rosen, 1992). In ASSR measurements, the response to the temporal envelope can be referred to as neural envelope encoding (Goossens *et al.*, 2018). The strength of neural envelope encoding can be observed in the frequency domain at the frequency corresponding to the modulation rate of the stimulus. A study by Shannon *et al.* (1995) demonstrated that consonants, vowel, and words in sentences can still be identified with high performance even when limited spectral information (formants at 16, 50, 160, and 500 Hz) are available in the speech. Important units of speech for speech recognition such as syllables and phonemes are typically represented in the temporal envelope corresponding to the modulation rate of approximately 4 and 40 Hz respectively (Chait *et al.*, 2015).

It is demonstrated that cortical activities contribute more to the ASSR when using stimulus with modulation rate <30 Hz, ASSR at approximately 40 Hz comprise both cortical and brainstem activities, and ASSR to >80 Hz modulation rate are primarily generated from the brainstem (Herdman *et al.*, 2002; Goossens *et al.*, 2018).

ASSR at the brainstem level may only represent the audibility aspect of speech intelligibility, similar to the relationship between click-ABR and speech intelligibility, due to the use of modulated tone stimuli. A positive correlation between the brainstem level ASSR and SIN performance is mostly found in adults with normal hearing, however, the correlation is negative for adults with hearing impairments (Ahissar *et al.*, 2001; Dimitrijevic, John and Picton, 2004; Doelling *et al.*, 2014; Manju, Gopika and Arivudai Nambi, 2014). Although the correlation between brainstem ASSR amplitude and SIN performance is significant ($r = -0.49$, $p = 0.001$), the correlation only accounts for limited range of SIN performance, 17% of the variability in Leigh-Paffenroth and Murnane (2011). This was suggested to be due to the variability in inter-subjects ASSR.

A few studies reported that the correlation between ASSR at the cortical level and speech intelligibility is negative, this is opposite to findings from ASSR at the brainstem level. The increase in cortical level ASSR is found to be correlated with poorer of speech intelligibility in modulated background noise for normal hearing people in adults (Millman *et al.*, 2017; Goossens *et al.*, 2018; Guest *et al.*, 2018). However, these studies were not replicated in normal hearing people. This negative correlation between cortical ASSR and SIN performance is more evident in people with hearing impairment (Millman *et al.*, 2017; Goossens *et al.*, 2019; Heeringa and van Dijk, 2019). Millman *et al.* (2017) suggested that this may be linked to increase in top-down processing compensating the lost in cochlea sensorineural inputs (Auerbach, Rodrigues and Salvi, 2014).

Overall, studies shows that ASSR is significantly correlation to SIN performance and encoding of natural speech representations (envelope modulation at syllabic or phonemic rate) can be assessed. However, this correlation is considered only at the group level, so it is not known whether ASSR will be a robust indicator of individual SIN performance. Findings from these studies also demonstrated that ASSR and behavioural SIN performance in adults with normal hearing and with hearing impairments correlates in opposite direction. This suggest that the combined use of top-down and bottom-up auditory processing may contribute to the neural envelope encoding, which confound and lead to an unclear conclusion of whether ASSR is an assessment of hearing function or auditory cognition.

3.4.3 Cortical auditory evoked responses

3.4.3.1 CAEP to short speech stimulus in young adults

Billings *et al.* (2013) conducted a study to determine whether the behavioural SIN performance in normal hearing young adult participants (23-34 years old) can be predicted from the waveform components in the CAEP (amplitude and latency of peaks) while participants were not attending to the stimuli. The study also examined the effects of both SNR and signal level on the CAEP components and the behavioural measure. SNR ranged from -10 to 35 dB, signal level ranged from 50 to 80 dB SPL. The stimuli for measuring CAEP was a monosyllable /ba/, the sentences for the behavioural test were retrieved from the Institute of Electrical and Electronics Engineers (IEEE) recommended practice for speech quality measurement. Background speech-spectrum noise contains approximately the same spectrum as the speech material for the behavioural measurement.

The main findings from the study by Billings *et al.* (2013) indicated that the amplitude or latency of the CAEP waveform components (N1, P2, and N2) are positively correlated with behavioural SIN performance in all four signal level conditions, 50, 60, 70, and 80 dB SPL. The best CAEP component for predicting SRT is the N1 negative-peak, the correlation coefficient between the amplitude (and latency) of N1 and SRT ranged is 0.770 (and 0.627). The test condition using signal at 70 dB SPL with 5 dB SNR showed highest correlation between N1 component and SRT. Varying SNR significantly affects the amplitude and latency of CAEP components: larger amplitude and shorter latency positive/negative peaks occur with higher SNR. The increases of signal level did not significantly increase the amplitude or shorten the latency of CAEP components, suggesting that the CAEP components are more influenced by the change in the listening environment rather than the intensity of the target speech.

3.4.3.2 CAEP to short speech stimulus in children and older adults

Generally, CAEP from adults would show a complex of P1, N1, P2, and N2 in the waveform. For infants and children under 9 years old, the N1 and P2 tends to be merged with P1 and N2 respectively, resulting in a waveform with only P1 and N2 components (Ceponiene, Rinne and Naatanen, 2002). The N1 component in children and adults appears to be contributed from distinct neural sources which may also reflect different auditory processes (Ceponiene, Rinne and Naatanen, 2002). The N1 is also strongly associated with the attentiveness to the stimulus (Naatanen, 1990), where attentional task can be more challenging for children compared to adults. These findings implies that the N1 component may not be a reliable predictor of the behavioural measure of speech intelligibility in infants and children.

The N2 component has been found to be the most dominant negative peak for the CAEP waveform in children and the neural source contributing to this component is similar to what is found in adults (Ceponiene, Cheour and Naatanen, 1998). A study by Anderson *et al.* (2010a) found that better performance in the behavioural speech-in-noise test is inversely correlated with the amplitude of the N2 component. The authors also suggested that, in children, the N2 component is related to the need for greater listening effort.

The CAEP in older adults (age above 60 years old) with normal hearing is found to be correlated with behavioural speech-in-noise performance, similar to young adults (Billings *et al.*, 2015). Prediction of SRT in older adults, however, generally leads to greater prediction errors than in young adults (Billings *et al.*, 2015; Koerner and Zhang, 2018).

This may be due to the greater variability of hearing threshold in older adults compared to young adults, particularly at 4 kHz and above, which can also cause variability in both behavioural and electrophysiological measurements. Although not often taken into account in CAEP studies, the cognitive abilities (e.g., attention, memory, and processing speed) in older adults may also vary considerably between individuals.

Previous studies shown that amplitude and latency of peaks in the CAEP waveform are correlated to SIN performance in children and adults. In contrast to subcortical measurements, CAEP to speech stimuli may be capable of capturing both acoustic and cognitive processing related to speech intelligibility and has been studied in many age groups. This shows that CAEP to speech stimuli could be a promising objective measure to predict speech intelligibility. However, the complications of using CAEP to predict SRT may be caused by the influence of other cognitive processes, such as attention. This is an important issue especially when testing in children and hence the method may not be suitable to predict speech intelligibility.

3.4.4 Cortical response to continuous speech

Studies have shown the ability of the human brain to entrain to the temporal envelope of sound; human brain activity measured by EEG synchronises temporally with the stimulus envelope (Aiken and Picton, 2008b; Ding and Simon, 2014). This phenomenon was found when using either AM tones or continuous speech stimuli. Researchers have been trying to explain the relationship between the neural entrainment to the speech envelope and speech intelligibility. An increase in temporal envelope entrainment correlates with better performance in speech comprehension tasks (Ahissar *et al.*, 2001). This also links with the suggestion that better cortical tracking of speech at syllable rate (~5 Hz) improves intelligibility (Doelling *et al.*, 2014).

Another important finding in the research area is that attention plays an important role in the envelope entrainment phenomenon. Several electrophysiological studies stimulated the cocktail party scenario where multiple talkers are competing with the target speech. The studies found that the neural response strongly encodes the speech envelope from the attended talker, and weakly correlates to the unattended talkers (O'Sullivan *et al.*, 2015; Biesmans *et al.*, 2017; Das, Bertrand and Francart, 2018). This is also a top-down processing phenomenon, where the attention enhances the intelligibility of the target speech by suppressing the unattended speech stream (Ding and Simon, 2012a). This selective attention is also found for auditory processing at the brainstem level (Etard *et al.*,

2019), where the listener shows significantly stronger cortical tracking to the fundamental frequency of the targeted speaker compared to the competing speakers.

Another way to relate neural entrainment to the speech envelope and speech intelligibility is to manipulate the intelligibility of the continuous speech stimuli, using the four different methods mentioned in section 2.1. The manipulation of the stimulus intelligibility normally results in a change in the acoustic property of the stimuli. As a result, it becomes unclear whether the neural entrainment becomes different in consequence of intelligibility, or the acoustic property of the stimuli. For example, several studies found correlation between the neural entrainment and continuous speech envelope, even when the speech is played backwards (unintelligible) (Howard and Poeppel, 2010; Pena and Melloni, 2012). If the cortical envelope entrainment is mainly driven by the envelope of the stimulus, then it may not represent whether the stimulus is intelligible or not. To directly use the cortical envelope entrainment to predict SRT could be a conceptual limitation of the method, if it is not clearly understood of what aspects of speech perception the cortical envelope entrainment is representing.

An example of reducing stimulus intelligibility by adding stationary noise is from a study by Ding and Simon (2013). The authors demonstrated that only the neural entrainment in the theta-band (4-8 Hz) correlates with the stimulus intelligibility, while the neural entrainment in the delta-band (1-4 Hz) is unaffected by the reduced stimulus intelligibility. It is suggested that the speech intelligibility affects only the neural entrainment in the theta-band but not in the delta band (Ding and Simon, 2014). The cognitive function related to speech comprehension may be reflected more in the neural envelope entrainment in the delta band when the perceived speech is degraded by background noise (Etard and Reichenbach, 2019).

Some researchers have been trying to predict SRT using the neural response to continuous speech (Vanthornhout *et al.*, 2018; Iotzov and Parra, 2019). The studies by Vanthornhout *et al.* (2018) and Iotzov and Parra (2019) used speech shaped noise as background noise prevent the confound between intelligibility and acoustic properties. Both studies showed strong correlation between the envelope entrainment and the behavioural measure of speech intelligibility. A study by Lesenfants *et al.* (2019) showed that the accuracy of SRT prediction at the individual level is within 2 dB difference in 88% of the participants (17 out of 19). However, these studies were only done in healthy normal hearing young adult subjects, this finding still needs to be validated in other age groups and hearing conditions.

Another factor that may be considered as a disadvantage is that the measurement of cortical responses to continuous speech requires relatively lone test time (~ 10 minutes per SNR condition) compared to ABR and CAEP (~3 minutes per SNR condition).

There has been a suggestion that the envelope entrainment may not be a major predictor of speech intelligibility (Ding and Simon, 2014). The evidence is from a study by Steinschneider, Nourski and Fishman (2013) who found the encoding of speech and non-speech sound in both human and monkey. Not only was it found that the envelope entrainment is non-speech specific, it is also not specific to humans. Another concern is the difference in nature between the objective and behavioural test. Since many behavioural tests involve some short-term memory to repeat the speech material, the objective measure may never flawlessly predict the behavioural performance (O'Sullivan *et al.*, 2015).

Apart from the envelope entrainment, it is hypothesised that auditory cortex may possibly track the syllable onset and collective features of the speech (Howard and Poeppel, 2010; Ding and Simon, 2012b). The 'syllable onset tracking' hypothesis was proposed by Howard and Poeppel (2010): the hypothesis was that the envelope entrainment is mostly driven by the brain response to the sharpness of the onset (rising) and offset (falling) of each syllable. The hypothesis is supported by the finding from Doelling *et al.* (2014), who found a decrease in the envelope entrainment when the temporal fluctuations at the syllabic rate (2-4 Hz) was removed. The 'tracking of collective feature' hypothesis supposes that the stimulus envelope contains the combination of pitch, source location, and timbre of the speech (Shamma, 2001). These collective features will then be disentangled and captured in successive processing steps, as the speech travels along the auditory pathway.

The analysis stage is a part of the acoustic processing is where the sensory organs, starting from the cochlea, disintegrate the incoming speech or non-speech sound into acoustic features. It is considered that both syllable onset tracking and collective feature tracking are processed at the analysis stage (Ding and Simon, 2014). The language and cognitive top-down process is not yet done at the analysis stage. However, high-level acoustic representations (e.g., phonological representation at the acoustic level) related to ability to discriminate speech from noise could be extract by combining acoustic features across frequencies (Nahum, Nelken and Ahissar, 2008; Ding and Simon, 2014).

3.5 Summary

Chapter 2 and 3 has outlined how behavioural measures and AERs, as an objective measure, have been used in predicting individual's speech speech-in-noise performance. It has extended the availability of the speech intelligibility tests to participants who are unable to undergo gold standard tests, such as PTA and behavioural speech intelligibility tests. AERs to several type of sounds (clicks, AM tones, monosyllable speech, continuous speech) has been demonstrated to be significantly correlated with behavioural SIN performance in many groups of subjects, however, it is unclear whether the complexity of stimulus will affect the accuracy or SRT prediction. With regards to the cortical entrainment to speech envelope, it remains debatable whether it is only a measure of access to speech sound (detection stage) or it can be related to more complex processing of speech sound (discrimination, identification, and comprehension). Finally, to make the framework of using AERs to predict SRT more applicable in clinics, it is also important to consider the required time to measure the responses and accuracy of SRT prediction especially at the individual level.

Chapter 4 Experiment 1: Decoding cortical responses to continuous speech with additional pauses inserted between words

4.1 Introduction

The first study explores whether the detection of cortical responses to continuous speech can be improved by modifying the stimulus without affecting the speech meaning. To the best of the authors knowledge, no study involving auditory evoked responses utilises the method of inserting fixed duration pauses between phrases into speech, a method so far only used in behavioural such as studies by Tanaka, Sakamoto and Suzuki (2011) and Ghitza and Greenberg (2009). Therefore, in this study we investigate how continuous speech with additional pauses inserted between words affects human cortical auditory responses using the backward TRF. It was hypothesised that additional pauses in speech will generate stronger cortical responses because the effect of having longer inter-stimulus interval (ISI) in repeating sound stimulus is known to result in a stronger CAEP (Davis *et al.*, 1966). The cortical responses were then further explored to see how strongly they were influenced by the onset and non-onset segments of the stimulus envelope, to explore whether the increase in cortical envelope entrainment strength is primarily an effect of onsets similar to CAEP. Finally, this study examines whether cortical responses to continuous speech with additional pauses inserted between words could improve (reduce) detection time of backward TRF, over the use of natural speech in detecting cortical responses to speech.

Speech rate is one of the elements which influences individual's speech perception (Picheny, Durlach and Braida, 1989; Nejime and Moore, 1998; Krause and Braida, 2002). In audiological research, this element is often manipulated by compressing or expanding the temporal waveform of the speech stimulus (Schmitt, 1983; Picheny, Durlach and Braida, 1989; Nejime and Moore, 1998; Krause and Braida, 2002). The effect of compressed speech (faster speech rate) and expanded speech (slower speech rate) on intelligibility is highly variable. Compressed speech and expanded speech have both been found to increase (Schmitt, 1983; Krause and Braida, 2002) and decrease (Picheny, Durlach and Braida, 1989; Nejime and Moore, 1998; Kemper and Harden, 1999) individual's speech intelligibility and in some cases to cause no effect (Small, Kemper and

Lyons, 1997). Time-compressed and -expanded speech both cause distortion in the acoustic signal. It is clear that they have a strong effect on intelligibility when compressed or expanded more than 0.5 times relative to the original speech rate (Nejime and Moore, 1998; Nourski *et al.*, 2009). It is currently thought that that intelligibility of time-compressed and -expanded speech relates to the individual's cognitive processing capabilities (Nejime and Moore, 1998). It has been suggested that changes in brain function with age can influence speech intelligibility (Cerella, 1990; Wingfield, 1996; Janse, 2009). It has been reported that older adults, especially those who have reduced cognitive function, benefit from listening to speech presented at a slower rate, as there is more time to process linguistic information (Wingfield *et al.*, 1999). One of the cognitive functions that has been found to most correlate with speech perception in noise is working memory, while overall IQ has been found to be less significant (Akeroyd, 2008). It should also be noted that higher speech intelligibility can also be achieved by simply producing clear speech (i.e., speaking louder or with better articulation) without changing the rate of speech (Krause and Braida, 2002; Smiljanic and Bradlow, 2009). A challenge in interpreting responses to compressed and expanded speech is that it is difficult to disentangle the effects of changes in speech intelligibility from various acoustic properties of the stimuli, for example, intensity envelope and duration of pauses (Schmitt, 1983; Vaughan *et al.*, 2002; Ghitza and Greenberg, 2009).

Kayser *et al.* (2015) investigated the effect of an irregular speech rate on auditory responses in different brain locations and different frequency bands of the EEG. The modification of speech stimuli was carried out by extending or shrinking existing pauses between syllables and words randomly with limits to the modified pause of not more than three times the original duration. Auditory responses to speech with irregular rate generated weaker left frontal alpha power and cortical entrainment in the delta frequency band. The weaker responses were suggested to be related to reduce top-down control, e.g., less attention to stimuli, by the frontal and premotor cortices over the auditory cortex. Another study by Hambrook, Soni and Tata (2018) modified speech stimuli by inserting periodic silent pauses. They consistently found weaker cortical entrainment to speech. It is suggested that the interruption of silent pauses in speech degrades the acoustic properties and the rhythm. Some studies involving the backward TRF and other objective measure tools manipulated pauses in speech (typically pauses between phrases and sentences) by shortening them to 0.3 – 0.5 seconds. For example, in dual attention task studies by Power *et al.* (2012) and Kong, Mullangi and Ding (2014), long pauses were truncate to minimize

subject's attention to the unattended speech when there is silence in the target speech, and to make the speech stimulus more continuous. The two studies found that the amplitude of positive peaks in the response waveform are mostly affected by the attention to the story stimulus, while the negative peak was less affected.

EEG was analysed in the delta and the theta band to explore whether the responses in these two frequency bands would reflect the difference in auditory processing time scales and functional roles, which is frequently reported in other studies (Ding and Simon, 2014; Kayser *et al.*, 2015; Etard and Reichenbach, 2019). For example, the cortical entrainment of speech envelope in the delta band reflects the processing of words and phrases and tends to be strongly correlate with the intelligibility of speech (Vanthornhout *et al.*, 2018; Etard and Reichenbach, 2019), while the cortical entrainment in the theta band reflects the processing of syllables and is more correlated with acoustic features influencing speech segmentation (Kayser *et al.*, 2015; Etard and Reichenbach, 2019). The current study will analyse the backward TRF for different segments of the speech envelope, in particular comparing the backward TRF calculated from the entire speech envelope to that of just onset and offset regions.

The results are expected to provide new insights for comparing the backward TRF from different speech corpora which may have different speeds of delivery and a range of silent pauses between words, and the extent to which the observed cortical responses can be deemed to be dominated by the onset.

4.2 Methods

This section will describe the original speech stimuli and the modified versions and will describe the acquisition of the EEG data from participants. The EEG data used in the current study were collected by Yuhan Lu for his MSc project. Yuhan also recruited the participants and designed the experimental procedure. Therefore, the current study was a further work in terms of improving data analysis. The pre-processing of EEG data, the main analysis tool used (the backward TRF approach), and an outline the statistical analyses performed are also presented.

4.2.1 Participants

Sixteen native English speakers participated in this study (9 males; aged 18-41 years, mean 25 years old; 13 right-handed). All subjects self-reported as normal hearing. Hearing

thresholds were tested with pure-tone audiometry in a sound-proof room, using air conduction. Thresholds for all participants were below 25 dB HL in the frequency range from 250 Hz to 8 kHz. Ethics were approved by the University of Southampton Ethics Committee (ethics reference number is 20741). All participants provided written informed consent prior to the experiments.

4.2.2 Stimulus

The continuous speech stimulus used in this study was a segment of an auditory recording narrated by a female narrator in the free audiobook "The Children of Odin: Chapter 2 - The building of the wall", available online at https://librivox.org/the-children-of-odin-by-padraic-colum/. The speech stimulus was manually split into four segments. Segments varied slightly in length to fit in with natural breaks, but each was approximately 2 minutes and 30 seconds in duration. In addition to using the recorded speech directly, the recording was also modified by inserting either 250 ms (short), or 500 ms (long) pauses between words, resulting in three speech pause conditions (natural speech, short, and long pauses). The length of each segment with short and long pauses inserted was approximately 4 minutes and 20 seconds and 6 minutes respectively. A total of 12 segments of speech were used to test each participant, four segments per speech condition. The total duration for continuous speech stimulus for the no pause, short pause, and long pause conditions were therefore approximately 10 minutes and 20 seconds, 16 minutes, and 21 minutes and 42 seconds, respectively.

4.2.3 Experimental procedures

The experiment was carried out in a quiet sound-proof room with lights turned off. Participants sat on a comfortable chair in a relaxed position and were instructed to close their eyes during the experiment to reduce ocular movement and consequent artefacts. They could take a break during the test if needed.

Participants were presented with the 12 segments of speech containing natural speech, speech with short, and long pauses (henceforth referred to as speech pause conditions). The speech pause conditions were presented in a randomised order to reduce order effects, but the 4 blocks within each condition were presented in a chronological order to maintain the flow and progression of the story. Participants were asked to attend to the stimuli. Each participant was asked a multiple-choice question at the end of each speech segment, to assess their attention to the stimuli. All 12 segments of speech were presented at 70 dB

LeQ(A) through ER-2 insert phones (Etymotic, Elk Grove Village, IL) to both ears. Calibration was done only on the original stimulus (natural speech) via Type 4230, Bruel and KjAERs, pauses were added afterward to create speech with short and long pauses stimuli.

EEG data were recorded using a 32-channel BioSemi EEG system (ActiveTwo, BioSemi BV, Amsterdam, Netherlands). Electrodes were positioned according to the international 10-20 system. Additional external electrodes were placed bilaterally on the mastoid and on the chin as reference channels and to detect artefacts from swallowing. The sampling rate for the EEG data was 4,092 Hz. An online band-pass filter from 0.1-100 Hz and a notch filter at 50 Hz was applied during data collection.

Analysis of EEG and speech stimulus were performed with the MNE-Python software (Gramfort *et al.*, 2013). EEG data from every participant were re-referenced to common average. They were then band-pass filtered using a zero-phase (non-causal) FIR filter (filter length was 6.6 times the reciprocal of the shortest transition band) over the range 1-4 Hz (delta band) and 4-8 Hz (theta band), and then resampled to 128 Hz. The EEG recordings from each participant were normalised prior to backward TRF analysis, to give a mean of zero and a standard deviation of 1.

4.2.4 Extraction of speech envelope

The speech envelope was extracted using the Hilbert transform applied to the original speech stimuli (sampling frequency 44.1 kHz). The envelope was then band-pass filtered using the same filter settings as in the EEG analysis in the ranges 1-4 Hz and 4-8 Hz (matching the delta and theta frequency bands used in the EEG also – see below) and resampled at 128 Hz.

As the auditory cortex is sensitive to acoustic edges, we processed onset and non-onset segments separately (see section 4.2.5 below). To construct these segments, we detect the pauses in the speech envelope by setting a low threshold level around the zero value, then replaced the samples below the threshold level with Not a Number (NaN) to exclude them from further analysis. Pauses are excluded from both the onsets and non-onsets representation prior to the backward TRF analysis, to avoid confounding from having different lengths of segments without any sound stimulus. In order to selectively process onsets, only samples within the first 150ms following each pause were kept, other samples (the non-onsets) were replaced with NaNs. This 150ms window was selected based on the duration of syllables as suggested by other studies (Greenberg *et al.*, 2003; Giraud and

Poeppel, 2012). To select only the non-onsets, the process was reversed, the onset segments samples in the speech envelope were replaced with NaNs and the remaining samples left with their original value. The full recording refers to the data including onsets, non-onsets and pauses. Henceforth, the full envelope, onsets, and non-onsets will be referred to as the different speech features.

4.2.5 The temporal response function

To test the hypothesis that the cortical response to speech with pauses would be enhanced by the onsets following pauses, the decoders were trained on the full envelope as well as the onset and non-onset segments; thus, three decoders were produced for each speech feature (full envelope, onsets, and non-onsets) and each participant. For the samples where the speech envelope was set to NaN, the error in model fit cannot be calculated and they are thus excluded from the least-mean-square fitting in equation (2). Then each decoder was tested on all three speech features to assess its ability to generalise across different parts of the recording. If the onsets dominate the EEG response and hence the decoder, then the decoder derived from the onsets should be able to reconstruct the full envelope better (i.e., give a higher correlation coefficient) than the decoder derived from the non-onsets. Similarly, in this case one might also expect that the decoder derived from the full envelope would reconstruct the onset responses better than that for the non-onsets. By testing each of the three decoders on all speech features, nine correlation coefficients obtain from the backward TRF (BACKWARD-CORR) are obtained, which can provide insight into which aspect of speech may be dominating the EEG response.

The recordings with pauses have a longer duration, as described in section 4.2.2. Comparison of decoder response detection performance may be biased by this difference in length of recording. Therefore, an additional analysis was carried out to compare the detection rates for the different stimulation conditions using only 2 minute and 1 minute segments from each stimulus, with the same length of recording used in cross-validation.

4.2.6 Statistical analysis

Permutation tests were performed to assess the significance of the BACKWARD-CORR. A null distribution of BACKWARD-CORR for individuals in each speech condition was obtained by using the speech envelope and a mismatched (permutated) EEG segment to train the decoder and perform the cross-validation on (previously unseen) testing data with the speech envelope and EEG segment also mismatched. Randomization was repeated 500

times to construct the null distribution of BACKWARD-CORR. The BACKWARD-CORR from the correct matching of speech envelope and EEG segment was then tested against this null distribution, at a significance level of $\alpha=0.05$. In this way the significance of estimated BACKWARD-CORR was tested in each recording, and not just of the average performance across the cohort.

Friedman tests were used to explore differences in the BACKWARD-CORR within the same decoder training and testing combination across three speech pause conditions (multiple tests on natural speech, short, and long pauses conditions). Wilcoxon signed rank tests were used to explore differences in the BACKWARD-CORR between each decoder training and testing combination. Bonferroni corrections were applied in all multiple comparisons. The adjusted α-level after Bonferroni correction was 0.0167 (0.05/3) across three speech pause conditions for comparison within the same decoder training and testing combination. The adjusted α-level after Bonferroni correction for tests within each speech condition was 0.00185 (0.05/27) for pair wise comparison, resulting from the nine different combinations of training and test datasets used (27 Wilcoxon tests in total for each EEG frequency band). Results were reported as statistically significant only in accordance with this Bonferroni correction, when $p \geq \alpha/N$.

4.3 Results

Figure 4.1 shows the correlation coefficients between the true speech envelope and that reconstructed via the backward TRF, obtained from different decoder training and testing combinations, as a function of the pauses used (natural speech, short, and long pauses) in the delta and theta bands. Although nine BACKWARD-CORR values were obtained for each speech condition (three training conditions and three testing conditions), for the sake of clarity only the results for training on the full envelope are shown, with others provided in the Appendix A (Figure A1 and Figure A2). Three combinations of testing and training data were displayed in Figure 4.1 and Figure 4.2: training and testing on the full envelope, training with the full envelope and testing on onset regions only, training with the full envelope and testing non-onset regions only. The increasing duration of pauses generally raised the BACKWARD-CORR for all decoding features in both delta and theta bands. Friedman tests indicated statistically significant difference in BACKWARD-CORR for comparison within the same decoder training and testing combination across the three pauses conditions ($p<0.001$ for the Bonferroni corrected level for significance $p \geq 0.0167$).

The results of pairwise Wilcoxon signed rank test between pairs of speech features across all speech pauses conditions are shown in Table 4.1.

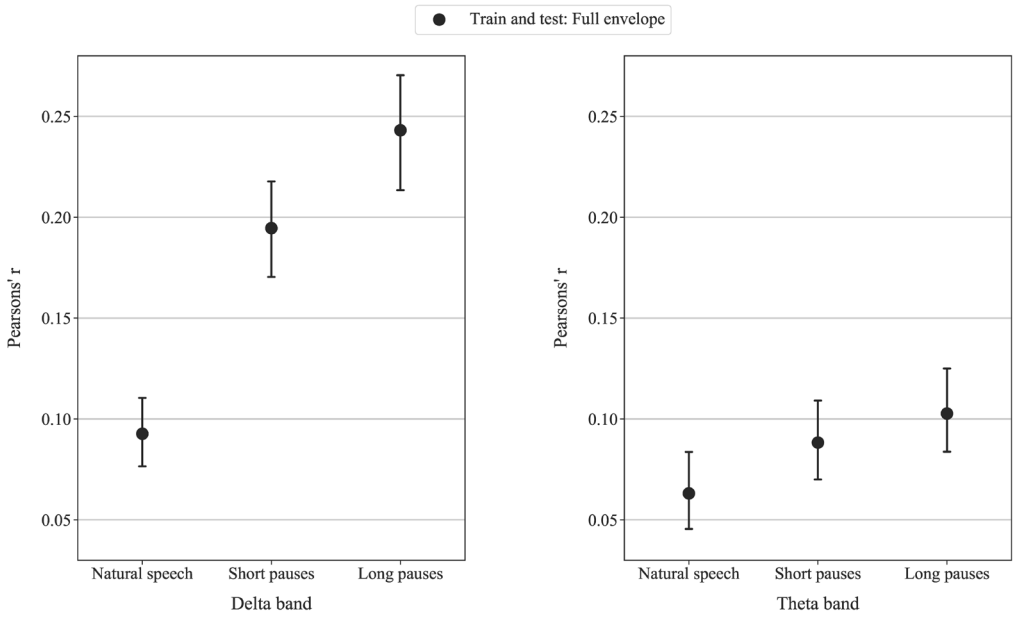

Figure 4.1. The mean correlation coefficient from backward TRFs trained and tested on the full envelope in (left) delta and (right) theta bands across three speech pause conditions. Error bars indicate the 95% confidence interval for the mean.

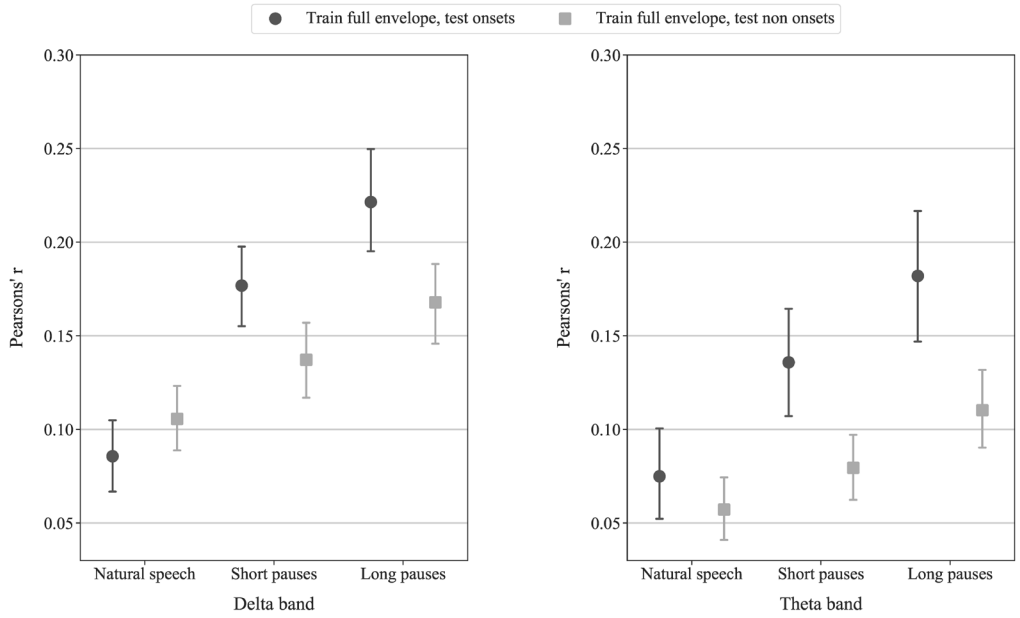

Figure 4.2. The correlation coefficient from backward TRFs trained on the full envelope and tested on onsets (circles) and non-onsets (squares) in (left) delta and (right) theta bands across three speech pause conditions. Each point shows the average Pearson's correlation coefficient across sixteen participants. Error bars indicate the 95% confidence interval for the mean.

Table 4.1. P-values for all possible pairwise tests (Wilcoxon Signed Rank Tests) across all speech pause conditions and speech features tested using model trained on the full envelope for both the delta and theta bands. Full envelope is abbreviated as Full. Significant p-values are shown in bold and italic (critical values from Bonferroni correction). P-values which are underlined indicate that the speech feature labelled at the top of the column with an underline has significantly greater correlation coefficients, or else the other speech feature is greater.

Speech pause conditions	**Speech feature comparison pair**		
Delta band	Full/Onsets	Full/Non-onsets	Onsets/Non-onsets
Natural speech	0.26	***0.001***	0.02
Short pauses	0.013	***<0.0001***	***0.001***
Long pauses	0.006	***<0.0001***	***0.001***
Theta band	Full/Onsets	Full/Non-onsets	Onsets/Non-onsets
Natural speech	0.004	0.039	0.004
Short pauses	***<0.0001***	0.07	***0.001***
Long pauses	***<0.0001***	0.011	***<0.0001***

In order to analyse the results in more detail, the decoder trained and tested on the full envelope were firstly considered, and then the decoder trained on the full envelope tested on the onsets, and finally the decoder trained full envelope tested on the non-onsets.

4.3.1 Effects of extended duration of pauses in continuous speech on the decoder trained and tested on the full envelope

From Figure 4.1, it was observed that for the decoders trained and tested on the full speech envelope, the BACKWARD-CORR gradually increase across the speech pause conditions in both the delta and theta bands ($p<0.001$, Friedman tests). The improved reconstruction of the full speech envelope indicates that responses to speech with extended pauses generate a stronger EEG response (i.e., enhanced linear relationship to the speech envelope), as originally hypothesised.

A similar trend of increasing BACKWARD-CORR for the full envelope decoders is also shown in the theta band (Figure 4.1.right) ($p<0.001$, Friedman). This further reinforces that auditory responses to speech with extended pauses are stronger than responses to natural speech. It may also be noted that the BACKWARD-CORR from the delta band is higher than for that the theta band. This agrees with previous studies using the stimulus

reconstruction approach (Etard and Reichenbach, 2019; Verschueren, Vanthornhout and Francart, 2020).

4.3.2 Effects of extended duration of pauses in continuous speech on the decoder trained on the full envelope tested on the onsets and non-onsets

From Figure 4.2, when training the decoder using the full envelope, the BACKWARD-CORR for testing on onsets or non-onsets were not significantly different in the natural speech condition. However, when short or long pauses were included in the speech, the reconstruction of the onsets was significantly better ($p<0.001$) than for the non-onsets, indicating that the model is better adjusted to the onset than the non-onset speech envelope segments. This suggests that with the longer pauses, the backward TRF model becomes dominated by the onsets. Similar impacts of pauses in the speech on the reconstruction of onset and non-onset segments were observed in both delta and theta bands, though more dramatically so in the delta band (Figure 4.2.left). The higher BACKWARD-CORR achieved in the onset segments compared to the non-onset segments are very clear, with statistical significance, in accordance with our original hypothesis.

4.3.3 Comparing detection of cortical entrainment to speech with extended pauses using different amount of testing and training data

Figure 4.3 shows the Box and Whiskers plot of BACKWARD-CORR across subjects using EEG data with different durations per segment across the three speech pause conditions in both the delta and theta band. With the same amount of training and testing data, the cortical entrainment to continuous speech with additional pauses inserted between words generally show significantly stronger BACKWARD-CORR (see Table 4.2) compared to cortical entrainment to natural speech ($p<0.001$), except when comparing BACKWARD-CORR between response to natural speech and short pauses in the theta band. This implies that the increase in BACKWARD-CORR in speech pause conditions is not simply a result of having longer EEG recordings.

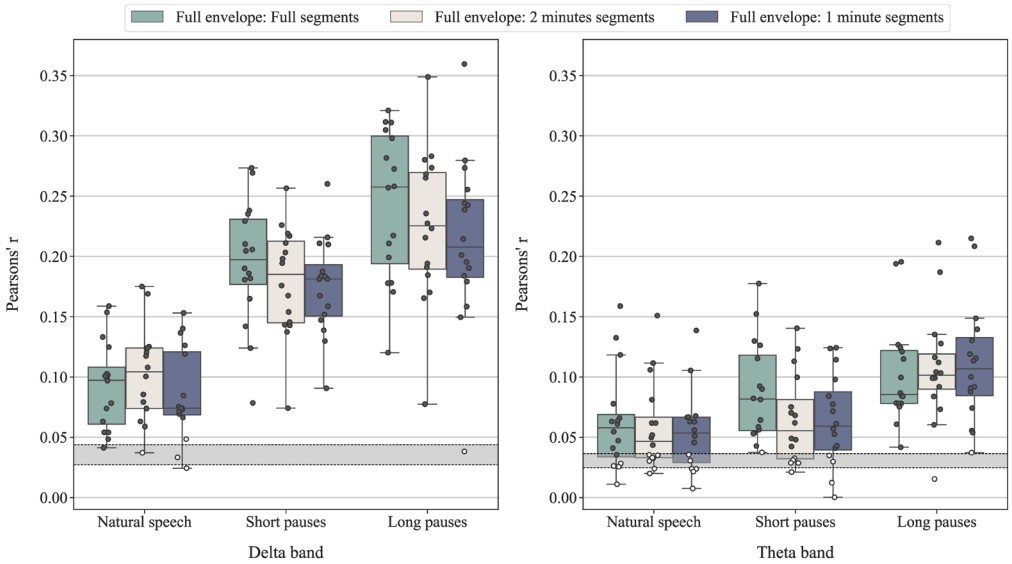

Figure 4.3. Box and Whiskers plot of correlation coefficients from each participant's backward TRF using different amount of training and testing data in the delta (left) and theta band (right). Light grey, grey, and dark grey boxes contain correlation coefficients from the full envelope using different stimulation durations (full length [2 minutes 30 seconds], 2 minutes, and 1 minute segments) recording from each data segment (in total 4 segments), respectively. Full segment stimulation refers to the full duration of the recording of each segment including natural pauses, whereas the 2 and 1 minute segments stimulation refers to recording segments with added pauses whose duration is truncated to 2 or 1 minute, respectively. The grey horizontal band indicates the range of critical values obtained from individuals in the sample, based on the null distribution of the correlation coefficients from the permutation test only from backward TRFs trained and tested on the full envelope with segments in full length (i.e. all estimates above this band are deemed significant, the band is higher for shorter duration recordings). Dots overlaid on each box are the backward TRF correlation coefficients from each participant. White dots indicate individual correlation coefficients that are not statistically significant based on subject's null distribution in each speech pause condition.

Table 4.2. P-values of differences in correlation coefficients between different stimulation durations (Wilcoxon signed rank test). Full envelope is abbreviated as Full. Bold and italic p-values indicate statistically significant difference (Bonferroni corrected) in correlation coefficients between data reduction conditions. P-values which are underlined indicate that the speech feature labelled at the top of the column with an underline has significantly greater correlation coefficients, or else the other speech feature is greater.

Segment lengths	Natural speech /short pauses	Natural speech/long pauses	Short pauses/long pauses
Full (delta)	***<0.0001***	***<0.0001***	***0.001***
2 minutes (delta)	***<0.0001***	***<0.0001***	***0.002***
1 minute (delta)	***<0.0001***	***<0.0001***	0.049
Full (theta)	0.034	***0.001***	0.015
2 minutes (theta)	0.679	***0.001***	***0.001***
1 minute (theta)	0.179	***<0.0001***	***0.001***

4.4 Discussion

In this study, it was found that continuous speech with additional pauses inserted between words generate stronger neural entrainment to speech envelope, as measured in correlation coefficient from between the actual and the reconstructed speech envelope (BACKWARD-CORR), in both delta and theta frequency bands. Analysis of the way in which different components of the speech envelope are reconstructed from the EEG signal demonstrated that there are two distinct responses: onset and non-onset responses. In the natural speech condition, responses in the delta band appear stronger to non-onset segments, while theta band responses appear stronger to onset segments. When pauses were introduced into the speech, onset responses become dominant in both the delta and theta bands. This might be a concern that the stronger cortical response generated dominated by the onset portions of speech with additional pauses might primarily be an exogenous response, mainly influence by the acoustic property, rather than a response that is influenced by the speech intelligibility.

Modifications in the speed of presented speech have typically been carried out in previous studies by either modifying silent pauses alone, or by altering (compressing or expanding) the temporal waveform of speech. An advantage of only modifying the duration of pauses, as used in this study, is that its effect on speech intelligibility can be investigated independently from effects of changes in acoustical properties, such as intensity and frequency and the speech envelope of individual words. The main disadvantage of manipulating pauses in speech is that the flow of speech can be severely altered when inserting pauses between words or phrases and in the current case the speech does indeed sound quite unnatural with the inserted pauses. The advantage of using time-compressed or -expanded speech is that the overall rhythm of speech does not change greatly compared to natural speech, however there are changes in acoustic properties affecting the articulation of phonemes (Vaughan *et al.*, 2002). Analysis of EEG responses to compressed or expanded speech may lead to confounding between the effects of changes in pauses and changes in phonemes. Our experimental protocol only affected the pauses and clearly demonstrated their powerful effect on EEG responses.

This study only shows the results from the decoder trained on the full envelope and tested on the full envelope, onset, and non-onset segments. It was an initial concern that onset and non-onset segments may have a different amplitude range and since larger amplitude ranges tend to lead to increased BACKWARD-CORR, such differences might bias results.

However, further analysis showed that onset and non-onset segments had similar speech envelope ranges, reducing this concern. It may also be noted that the onset and non-onset segments were of similar length in these recordings, with a slightly greater number of samples in the non-onsets compared to the onsets. This implies that the increase in BACKWARD-CORR was neither a result from greater variance in the sample nor bias towards a model which was trained on greater amount of data because all the models were trained on the full speech envelope.

Regarding the Bonferroni correction for multiple comparisons, the method is highly conservative compared to other correction methods. While the false positive rate (incorrectly identify the differences in the BACKWARD-CORR as significant) is controlled (5% for α=0.05), the false negative rate (incorrectly identify the differences in the BACKWARD-CORR as not significant) is inflated (Armstrong, 2014). Some results that are deemed non-significant after applying the Bonferroni correction may be significant if a different correction method, such as the false discovery rate, is used instead. However, it was imperative to employ a relatively conservative method to control the false positive rate, as the number of multiple tests is relatively high (27) for the decoder framework and the obtained p-values in post hoc Wilcoxon signed rank tests are generally lower than 0.05.

4.4.1 Effect of pauses in speech to cortical auditory responses

As shown in Figure 4.3, the pauses in speech not only increased the average BACKWARD-CORR in the group but could achieve this with shorter recordings. The BACKWARD-CORR was also statistically significant in each subject when using short and long pauses, but only in 9 out of 16 subjects when using natural speech and training data of 3 minutes (segments of 1 minute in the leave-one-out cross validation) or 6 minutes (segments of 2 minutes). The use of shorter segment length not only lower the BACKWARD-CORR value but also caused the critical values of the BACKWARD-CORR obtained from the permutation test to be greater compared to the use of full recording, which leads to greater number of non-significant BACKWARD-CORR.

There have been two previous studies that have explored the effects of stimulus manipulation on neural responses to speech, though the protocol was different to that used in the current study and the pattern of responses found was also somewhat different. Kayser *et al.* (2015) made a comparable study, investigating the effect of irregular speech rate on the neural and behavioural responses to speech. The irregular speech was achieved by increasing or decreasing existing pauses in the natural speech. This is considerably

different from what is done in the current study, as the pauses can either expand or shrink and there are no additional pauses inserted. They found a reduction in the cortical entrainment compared to natural speech only in the delta band, with no difference in other frequency bands. Behavioural speech intelligibility also remained approximately the same for both natural and irregular speech. It was suggested that the top-down processes of speech perception, using prior knowledge in language to comprehend speech (Zekveld *et al.*, 2006), reduced cortical entrainment in the delta band. Top-down processing has been found to be sensitive to the regularity of sound (Schroeder and Lakatos, 2009; Hickok, Farahbod and Saberi, 2015) which might affect the cortical entrainment. The reason that the cortical entrainment to an irregular speech envelope in other frequency bands remain similar to that of cortical entrainment to natural speech may be that the modification of pauses in the study by Kayser *et al.* (2015) was primarily controlled to preserve the overall mean duration of pauses. The modified duration of pauses only changed compared to their original duration (and was limited to a maximum of 300%), rather than consistently increasing, as was the case in the current study. The duration of pauses in Kayser's work was probably not consistent or long enough to enhance auditory onset responses.

Another study that can be compared with ours is by Hambrook, Soni and Tata (2018), who examined the effect of periodic introduction of pauses and noises into the continuous speech, on both behavioural responses and cortical entrainment. The aim of their study was to explore neural function during the phonemic restoration phenomena, which refers to the observation that speech intelligibility degrades when the speech stream is interrupted by silent pauses and partly restored when noises fill the interrupting pauses. The introduction of pauses with the duration of 166 ms every 333 ms into the continuous speech (50% of speech removed) was found to significantly reduce speech intelligibility and the cortical entrainment to speech envelope, however, speech intelligibility and cortical entrainment to speech envelope improved when pauses were filled with noise. The disruption of acoustic tracking in the auditory cortex was suggested to be the reason for the reduction in cortical entrainment. Although they suggested that it is possible that the onset and offset segments of the speech envelope can be removed and replaced by silence, causing disturbance in the cortical tracking on those segments, they found no evidence to support this. This reinforces the idea that speech comprehension is not a process which is driven by the stimulus alone. In their study, the actual speech was affected by interruption of pauses and noises, while in our study speech were not replaced by pauses and noises. We presume that their manipulation method might disrupt the cortical processing, as speech was removed,

whereas our speech manipulation presumably did not and indeed increase the cortical entrainment to speech envelope.

4.4.2 Differences in the methods to define onsets

The selection of onset segments in this study differed from previous studies using EEG response. In our work, the first 150 ms portion following word onset is defined as the onset segment. Previous studies commonly calculate onset envelopes from the first derivative of the speech envelope (gradient of the speech envelope e.g., (Hertrich *et al.*, 2012; Drennan and Lalor, 2019). The gradient of the speech envelope only contains the rate of amplitude change, which is greatest for onset and offset segments. The reconstruction accuracy of the gradient of the speech envelope often results in weaker BACKWARD-CORR compared to the reconstruction of the standard speech envelope (Hertrich *et al.*, 2012; Drennan and Lalor, 2019). One problem with the gradient envelope is that it not only removes pauses, but also relatively constant amplitude sections of for example voiced speech. Our method of specific analysis of onsets using backward TRF does not appear to have been used previously in this area and seems better able to focus on these signal segments than previously used alternatives.

A study by Hamilton, Edwards and Chang (2018) observed the effect of onsets in a similar manner. In their study, they found that invasive EEG responses following pauses longer than 200ms generates strong onset responses that can occur both within a sentence or before the sentence starts. Our study extended their findings, by demonstrating that effect of strong onsets response persists even when the pauses are 500ms in duration and it could be detected by a non-invasive EEG measurement. It may be possible to investigate certain regions of the brain where they are specifically sensitive to acoustic edges and onsets non-invasively. However, the current study has not specified whether the strong onset response from the EEG was generated from the same region, the Superior temporal gyrus or STG part of the temporal lobe, as shown in the invasive EEG studies.

Some studies may refer the cortical entrainment to as the phase-locked responses to amplitude modulation, specifically phase-locking to change in acoustic cues (Bieser and Muller-Preuss, 1996; Peelle, Gross and Davis, 2013). It was suggested that the strength of phase-locked responses to amplitude modulation may be associated with strong onset response (Bieser and Muller-Preuss, 1996). Other study also found that the phase-locked responses are enhanced when stimulated with intelligible speech compared to less-intelligible speech and was not due to the effect of onset response (Peelle, Gross and

Davis, 2013). These studies showed that there is a possible confounding onset and intelligibility effect on the cortical responses to continuous speech. Thus, measurement of auditory responses to speech tokens or sounds, such as phonemes or consonants, may not be sufficient when aiming to probe auditory responses to higher level information in continuous speech (Di Liberto, O'Sullivan and Lalor, 2015).

The different behaviours for onset and non-onset responses adds to the discussion on analysing responses to speech: Is it primarily an observation on the response to acoustic features or to higher level processing of speech? If the former, one might ask if speech is the most efficient stimulus to use and to what extent it provides additional information to that obtained by repeated transient synthetic stimuli or by speech tokens.

4.5 Conclusion

The use of continuous speech with additional pauses has the potential to reduce the time required to detect cortical responses to speech. However, the stronger cortical responses do not necessarily link to improved speech intelligibility, since the speech was less intelligible due to the unnatural speech regularity with added pauses. The relation between cortical responses to continuous speech with additional pauses and behavioural speech intelligibility was not further studied, as experiments involving human subjects were not permitted during the COVID-19 pandemic lockdown.

Chapter 5 Exploring the characteristics of cortical responses to speech with additional pauses between words and a comparison to cortical auditory evoked potentials

5.1 Introduction

The second study was a further investigation on the dataset used in Chapter 4, mainly to compare the forward and backward TRF in detecting responses to continuous speech and further evaluate the inference in the previous chapter that the cortical envelope entrainment to speech with pauses may mostly be an exogenous response. The primary aim of this study was to compare the use of the forward TRF model weight (TRF-model) and the correlation coefficient obtain from the backward TRF (BACKWARD-CORR) for response detection to find the most sensitive detection method for cortical responses to continuous speech, which will be used in the next study (second experiment). The comparison of response detection also incorporates CAEP to /da/ stimulus to see whether cortical responses to speech can be more easily detected by simply using repeating /da/ (with shorter EEG measurement time) than using continuous speech. A sensitive response detection method is desirable especially when detecting responses to stimuli with low signal-to- noise ratio (SNR) or aiming to determine certain thresholds, such as hearing or speech reception thresholds (as in the study in Chapter 6). Ideally, responses are expected to be detected down to a certain stimulus level or signal-to-noise ratio (if stimulus is presented with background noise), where the boundary of response present and absent shall be considered as the threshold. In the current study, all measurements were carried out with no background noise, so it was not possible to measure thresholds in terms of SNR.

The secondary aim was to apply the forward TRF approach to determine how the characteristics of cortical responses to continuous speech progress when additional pauses are inserted and compare the morphology or the TRF-model to cortical auditory evoked potentials (CAEP) waveform. The aim of analysis was to investigate whether the characteristics of the TRF-model of cortical responses to continuous speech with additional pauses will be similar to responses to more intelligible stimulus, cortical response to natural speech, or become more similar to response to unintelligible stimulus, CAEP to

/da/. Result from Chapter 4 showed that the BACKWARD-CORR of responses to speech with pauses is greater compared to the BACKWARD-CORR of responses to natural speech. However, the BACKWARD-CORR alone cannot indicate how sound is encoded in the brain (characteristics), it can only tell how strong it was encoded and mainly used for response detection. Interpretation of the peaks amplitude and latency of the TRF model and CAEP waveform may indicate whether the auditory system processes intelligible and unintelligible sound similarly or differently from one another.

Relationship between the CAEP and TRF-model

CAEP waveform consist of several positive and negative peaks manually or automatically labelled with, for example, P1, N1, P2, and N2. Where P and N denotes positive and negative peaks, and numbers indicate the order of appearance (Picton and Hillyard, 1974; Munro *et al.*, 2020). The TRF-model describes the relationship between the response and stimulus as model weights over a range of time lag (Crosse *et al.*, 2016), where greater model weights appears at the time lags which the cortical responses covary most with the stimulus feature. In addition, the TRF-modelling can also be applied to measurement of CAEP instead of using the coherent averaging method (CA), as the epochs (sequence of responses) of CAEP are highly correlated to the repeating stimulus. The CAEP to repeating sound and the TRF-model of cortical responses to continuous sound both reflect the magnitude of the brain response to sound at each time delay following the onset of stimulus, thus the waveform morphology of the two measurements can be similar but differ in the scale of measurement. However, the firm connection between the CAEP and TRF-model has not yet been established. In some case, TRF-model of cortical responses to continuous speech may be expected to exhibit waveform in which the major components appear at a distinct latency, for example CAEP positive peak at ~80ms and TRF model positive peak at ~150ms (Reetzke, Gnanateja and Chandrasekaran, 2021). The difference in the latency of the major components is suggested to arise from the difference in auditory processing stages between continuous and simple short sound (Lalor *et al.*, 2009).

Applications of TRF-model

The forward TRF modelling approach is frequently used as a complementary tool rather than the main tool for measuring the cortical response to continuous speech. Conventionally, one can interpret the TRF-model by observing statistically significant amplitude and latency changes of the peaks in the model, similar to the interpretation of CAEP dominant peaks. The goal for interpreting the TRF-model is to establish a distinct

characteristics of cortical entrainment to continuous speech envelope changes relevant to the experimental conditions and/or group of subjects, for example effect of stimulus intensity (Verschueren, Vanthornhout and Francart, 2021) and effect of language and attention between native and non-native English speakers (Reetzke, Gnanateja and Chandrasekaran, 2021). It is not clear whether the TRF-model or BACKWARD-CORR are more sensitive to detect significant cortical responses to continuous speech. If the detection of responses through the TRF-model will be as sensitive as the BACKWARD-CORR, this will be an advantage due to the reduced computational cost because only one EEG channel is needed. Moreover, the characteristic of the response can be interpreted through the TRF-model, providing more insights on how the auditory pathway encodes the stimulus, as the direction of causality is correct.

Aim of study

1. To investigate if the TRF-model might be used more effectively than the BACKWARD-CORR to detect cortical responses to continuous speech and compare the efficiency of response detection with CAEP to /da/.
2. To assess if the TRF-model can provide additional insight into how natural speech and speech with additional pauses inserted between words is processed and compare it to the coherent average waveform of CAEP to /da/

5.2 Methods

The current study is mostly based on further analyses of the dataset used in Chapter 4 with an additional CAEP to /da/ measurements, also collected by Yuhan Lu. Information related to continuous speech will be described briefly in this chapter. Please refer to Methods section in Chapter 4 for more detailed information on the continuous speech stimulus and the experimental procedure.

5.2.1 Participants

The 16 participants were native English speakers (9 males; aged 18-41 years, mean 25 years old). All subjects verified to have normal hearing, threshold <25 dB HL over the frequency range of 250 to 8 kHz. Ethics were approved by the University of Southampton Ethics Committee (ethics reference number is 20741). All participants provided written informed consent prior to the experiments.

5.2.2 Stimulus

Continuous speech stimulus

The continuous speech stimuli used in this study were the same as in the previous study (Chapter 4). The stimuli were narrative story and its modified versions with additional pauses, 250 ms (short pauses) and 500 ms (long pauses), inserted between words, these are referred to as speech pause conditions. The continuous speech stimuli were presented at 70 dB LeQ(A) (calibrated via Type 4230, Bruel and KjAERs).

Repeating /da/ stimulus

A monosyllable /da/ stimulus with a duration of 40 ms was used to present each participant 200 times, alternating between rarefaction and condensation polarisation (100 per polarity), with an inter-stimulus interval of 1.11 s. This stimulus was presented at 70 dB SPL (calibrated via Type 4230, Bruel and KjAERs).

5.2.3 Experimental procedure

The experimental procedure in the current study was the same as in the previous chapter for the conditions using natural speech and continuous speech with pauses added between words. The only difference is that the current study includes a condition using repeating /da/ stimuli. For every participant, 200 repeating /da/ stimuli were presented before the continuous speech stimuli. The continuous speech stimuli with different speech pauses conditions were presented in a random order but the progression of the story remains in a chronological order. Participants were instructed to pay attention to the stimuli and were required to answer multiple choice questions after listening to each of the continuous speech stimuli segment to verify their attentiveness.

EEG data were recorded using a 32-channel BioSemi EEG system (ActiveTwo, BioSemi BV, Amsterdam, Netherlands). Electrodes were positioned according to the international 10-20 system. The sampling rate for the EEG data was 4,092 Hz. An online band-pass filter from 0.1-100 Hz and a notch filter at 50 Hz was applied during data collection.

5.2.4 Data analysis

Analysis of EEG and speech stimulus were performed with the MNE-Python software (Gramfort *et al.*, 2013). EEG response to continuous speech and /da/ from every participant were re-referenced to channel T7 and T8. They were then band-pass filtered using a zero-

phase (non-causal) FIR filter (filter length was 6.6 times the reciprocal of the shortest transition band) over the range 1-4 Hz (delta band) and 4-8 Hz (theta band), and then resampled to 128 Hz. The EEG recordings from each participant were normalized prior to data analysis, to give a mean of zero and a standard deviation of 1. The Cz channel (vertex or midline central) was chosen for the data analysis as the channel site can be locate precisely by measuring distance of nasion (bridge of the nose between the eyes) to inion (occipital bone at base of the skull) (Schestatsky, Morales-Quezada and Fregni, 2013) and it is frequently used in CAEP studies (Billings *et al.*, 2011; Small *et al.*, 2018).
The speech envelope was extracted using the Hilbert transform applied to the original speech stimuli (sampling frequency 44.1 kHz). The envelope was then band-pass filtered using the same filter settings as in the EEG analysis in the ranges 1-4 Hz and 4-8 Hz (matching the delta and theta frequency bands used in the EEG also – see below) and resampled at 128 Hz.

5.2.5 CAEP to /da/

The CAEP to /da/ stimulus was obtained by coherent averaging epochs of responses to /da/, as described in section 3.1, responses were aligned by the onset where the stimulus starts then calculating the average at each sample (column wise). The coherent average resulted in a CAEP in a range of 0-1110 ms.

CAEP to /da/ was automatically detected using the one sample Hotelling's T-squared (HT2) statistics, section 3.3 of this thesis provides a more detailed description of the method. In the time domain, nine time-voltage means (TVMs) were calculated for each epoch over the range of 50-500 ms latency (50 ms for each bin) (Golding *et al.*, 2009). The HT2 was then applied to determine whether any of the TVMs are statistically greater than the null hypothesised mean of zero, a hypothesis of no response detected. If the p-value from the HT2 test is less than or equal to 0.05, a response is deemed present.

5.2.6 Forward TRF

The forward TRF modelling, also known as the encoding model, utilise a single-channel EEG forward TRF model $w(\tau, n)$ to predict the EEG response using the stimulus envelope (Crosse *et al.*, 2016).

$$r(t,n) = \sum_{\tau} s(t-\tau)w(\tau,n) \tag{5.1}$$

where $r(t,n)$ refers to the predicted EEG response at channel n with the time index t, $s(t-\tau)$ refers to the speech envelope with the time index t and the range of time lag τ in the convolution (corresponding to model order). Equation (5.1) represents the so-called 'forward model' (TRF-model), unlike the backward TRF approach, this forward modelling approach imitate the causal stimulus and response physiology (hence the – sign in $s(t-\tau)$). The forward TRF model is calculated using the regularised least squares method,

$$w = (S^T S + \lambda I)^{-1} S^T r \tag{5.2}$$

where S is the stimulus envelope in lagged time series and r is the EEG signal. The identity matrix (I) was chosen for the regularisation instead of the M matrix used in the backward TRF approach due to the incorrect baseline in the TRF-model when using the M matrix for regularisation which caused difficulty in identifying significant peaks in the model, this is demonstrated in Figure 5.1. The maximum absolute peak within the 0-350 ms time lag in the TRF-model was the detection parameter for the forward TRF approach. This absolute peak detection method was also applied for cortical responses to /da/, in addition to the HT2, to assess whether the forward TRF modelling and coherent average (CA) is different in terms of detection sensitivity.

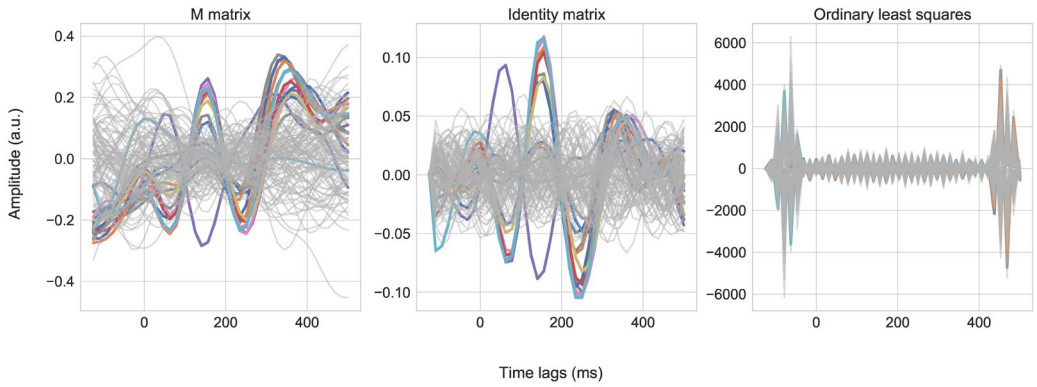

Figure 5.1. Effect of regularisation matrix on the true (coloured) and bootstrapped (grey) TRF-model. (Left) Using the M matrix for regularisation caused incorrect model baseline at the early and late samples. (Middle) Using the identity matrix for regularisation resulted in true and bootstrapped TRF-model with correct baseline at zero. (Right) The TRF-modelling without regularisation resulted in waveforms where significant peaks in the true TRF-model are not clearly seen within the expected time lag.

5.2.7 Backward TRF

The correlation coefficient between the actual and the reconstructed envelope stimulus or BACKWARD-CORR was the detection parameter for the backward TRF modelling of cortical responses to continuous speech. Detailed description of the backward TRF approach has been described in section 3.2. The reconstruction of stimulus envelope was calculated over the range of 0-300 ms lag, EEG lagging the stimulus. The identity matrix was applied for the regularisation. A leave-one-out cross validation (LOOCV) method was applied to find the optimal λ value over the range of 50 logarithmically spaced points between 0.01 to 10^5. Each speech pause condition contains the four segments of EEG and stimulus envelope, the optimal λ value gives the highest correlation coefficient after averaging from four iterations of LOOCV (all segments were allocated as testing set).

5.2.8 Statistical analysis

The bootstrapping technique was performed to determine if the detection parameters of cortical responses to speech (TRF-model and BACKWARD-CORR) and CAEP to /da/, is statistically significant compared to the parameters obtain from the misaligned response and stimulus null distribution. The null distribution was estimated by resampling random EEG segments (with replacement) while the stimulus feature remains the same, then calculate the mismatch version of the detection parameters following their described methods. The bootstrapping test was repeated 500 times for each detection parameters in each individual. The true TRF-model and CAEP to /da/, when responses and stimulus are correctly aligned, is considered statistically significant when the peak value exceeds the upper or lower $\alpha = 0.025$ of the null distribution. The true BACKWARD-CORR is considered statistically significant when the correlation value exceeds the upper $\alpha = 0.05$ of the null distribution.

A cluster-based permutation test was implemented on the TRF-model between speech pauses condition in the delta and theta frequency band to investigate the effect of additional pauses in the continuous speech on the shape of the TRF-model at the group level. This method was chosen to handle with the problem of increased chance of reaching statistical significance with multiple comparisons, even though there is no actual significant result present, when large number of statistical comparisons must be performed on multiple time points in the TRF-model from multiple subjects (Maris and Oostenveld, 2007; Weissbart, Kandylaki and Reichenbach, 2020; Reetzke, Gnanateja and

Chandrasekaran, 2021). The cluster-based permutation statistical analysis performs multiple comparison with corrections. With the use of this method, the critical alpha-level of 0.05 is controlled.

5.3 Results

5.3.1 Grand average and individual TRF-model

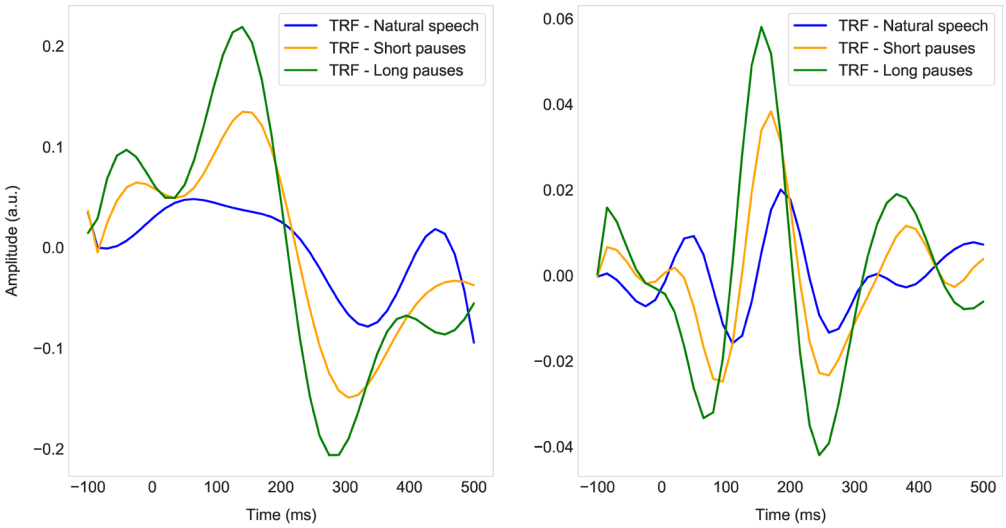

Figure 5.2. The grand averaged TRF-model across 16 subjects in all speech pause conditions in the delta (left) and theta band (right).

Figure 5.2 shows the grand averaged TRF-model across all subjects from each speech pause conditions within the time lag range of -100 to 500 ms in the delta and theta frequency band. Consistent with the findings in Chapter 4, the strength of cortical envelope entrainment progressively increases as the duration of pauses in speech increases. Visually inspecting, the grand averaged TRF-model in the theta band exhibit a negative-positive-negative peaks or 'N1-P2-N2' complex similar to CAEP to /da/.

The cluster-based permutation analysis indicated significant difference (t-cluster p-value < 0.05) in the positive and negative peak amplitude between TRF-model from the natural speech and speech with long pauses condition in both the delta and theta band. In the delta band (Figure 5.3), significant differences between the TRF-model from natural speech and TRF-model from speech with long pauses occurs within the time lag ranges of 55-155 ms, 230-290 ms, and 410 ms to 440 ms. In the theta band (Figure 5.4), significant differences between the TRF-model from natural speech and TRF-model from speech with long pauses occurs within the time lag ranges of 20-65 ms, 125-155 ms, 215-260 ms, and 350-

380 ms. There were no significant differences in amplitude change between TRF-model from speech with short pauses compared to the TRF-model from other two speech pause conditions in both delta and theta band.

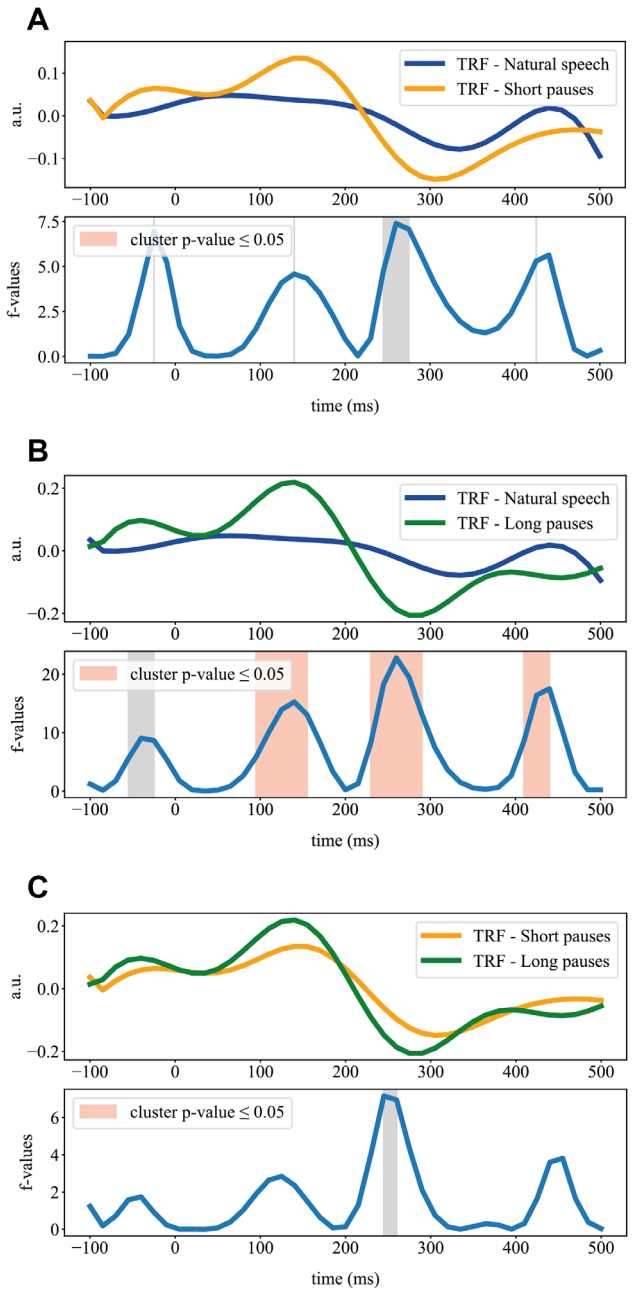

Figure 5.3. Pair-wise comparison of amplitude change between the TRF-models from each speech pause condition in the delta band. (A) Natural speech and speech with short pauses TRF-models. (B) Natural speech and speech with long pauses TRF-models. (C) Short and long pauses TRF-models. The time intervals with significant differences in amplitude after correction between two TRF-models are indicated by the orange colour band, while the grey colour band indicates cluster p-values not significant after correction.

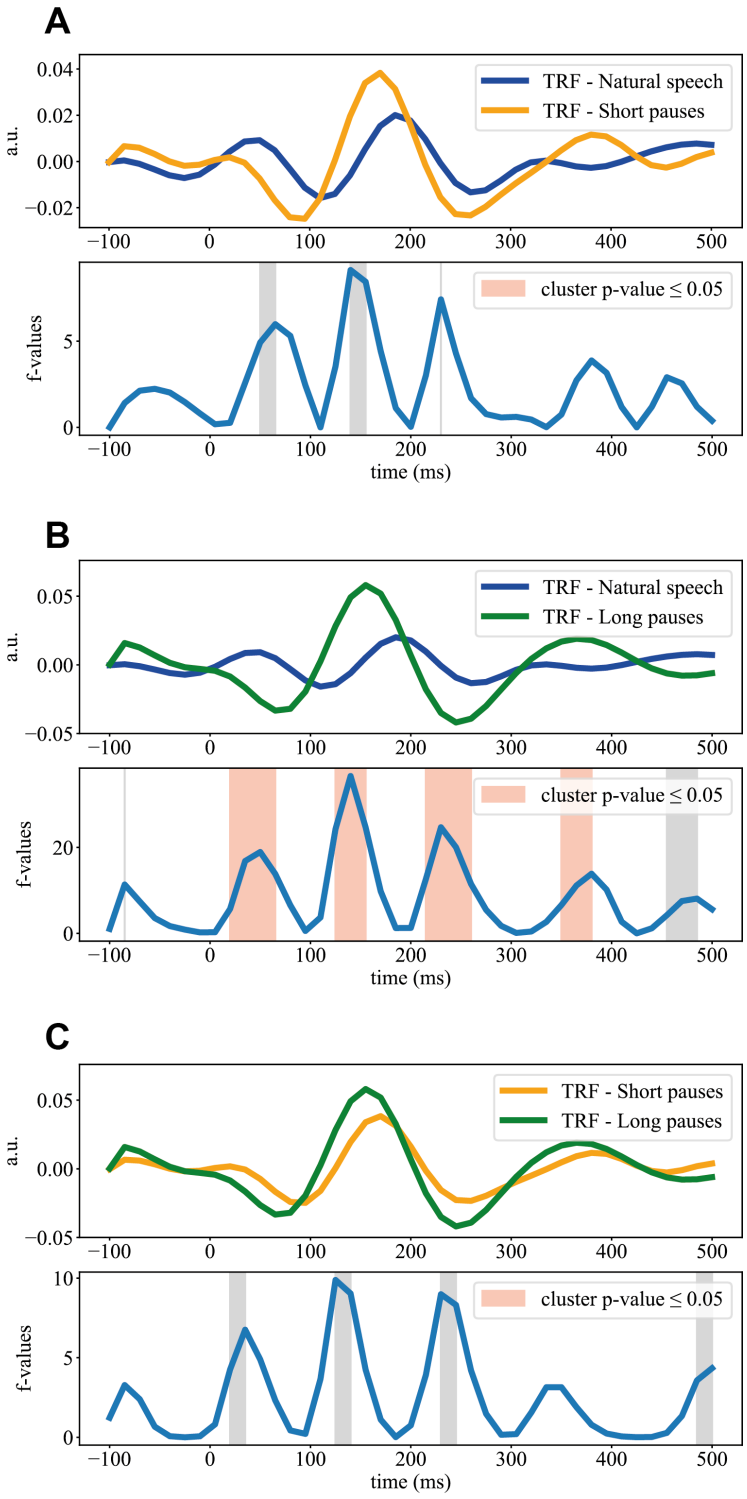

Figure 5.4. Pair-wise comparison of amplitude change between the TRF-models from each speech pause condition in the theta band. (A) Natural speech and speech with short pauses TRF-models. (B) Natural speech and speech with long pauses TRF-models. (C) Short and long pauses TRF-models. The time intervals with significant differences in amplitude after correction between two TRF-models are indicated by the orange colour band, while the grey colour band indicates cluster p-values not significant after correction.

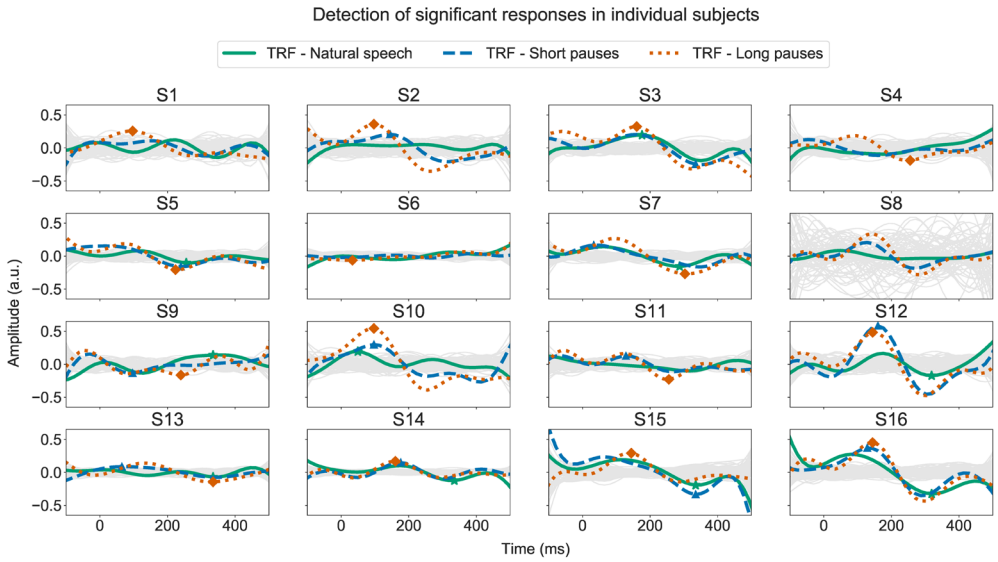

Figure 5.5. Detection of maximum absolute peak with statistical significance in the TRF-model in all speech pauses conditions in the delta band for each individual. Light grey plots show the bootstrapped TRF-models.

Figure 5.6. Detection of maximum absolute peak with statistical significance in the TRF-model in all speech pauses conditions in the theta band for each individual. Light grey plots show the bootstrapped TRF-models.

Figure 5.5 and Figure 5.6 show the subject TRF from 16 subjects from all speech pauses condition within the time lag range of -100 to 500 ms overlaid on the one hundred permutated TRF of each individual (labelled by S and number) in the delta and theta band. The significant absolute peak in the TRF of each subject is marked with the star, triangle, and diamond symbol for TRF within the natural speech, speech with short pauses and speech with long pauses respectively. The number of significant TRF (in the following order: natural speech, speech with short pauses and speech with long pause) was 10, 12, and 15 in the delta band (Figure 5.5) and was 10, 9, and 12 in the theta band (Figure 5.6).

5.3.2 Grand average of CAEP to repeating /da/

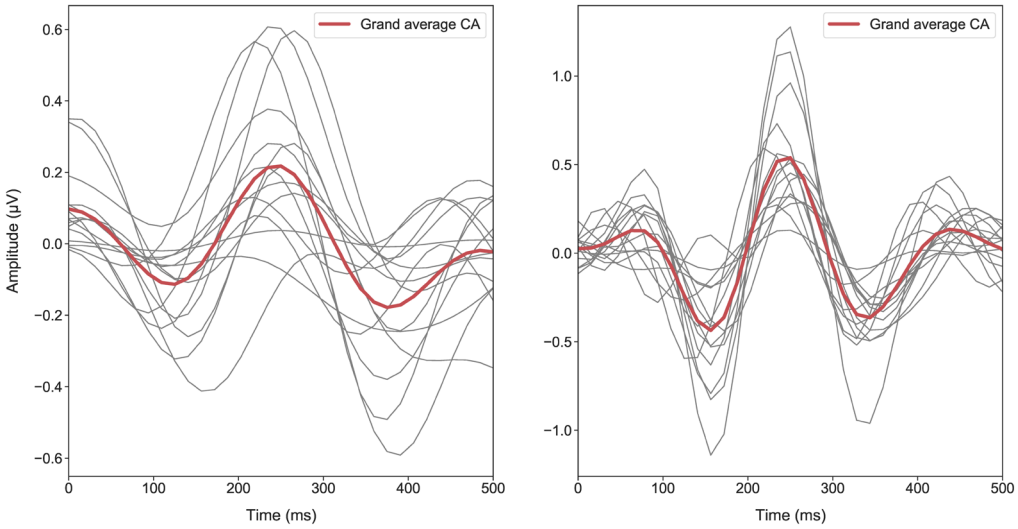

Figure 5.7. The individual (grey) and grand average (red) of coherent averages of cortical responses to /da/ in the delta (left) and theta band (right).

Figure 5.7 shows the individual and grand averaged of the coherent average waveforms within the 0 to 500 ms latency following the onset of stimulus from 15 subjects (1 subject was discarded because the EEG data was missing) in the delta and theta frequency band. HT2 analysis indicated that coherent average TVMs significantly different from zero for all subjects. The amplitude of P2 peak across individuals is greater in the theta band than in the delta band (Wilcoxon Signed Ranks Test, $p < 0.001$). However, the latency of the P2 peak, is not statistically different (Wilcoxon Signed Ranks Test, $p = 0.612$)

5.3.3 Latency of P2 in TRF-model of cortical responses to continuous speech and CAEP to /da/ in the theta band

Table 5.1. Mean and SD of P2 latency in TRF-model of cortical responses to continuous speech and CAEP to /da/ in the theta band

Condition	Mean latency (ms)	Standard deviation (ms)
TRF – Natural speech	207.93	61.24
TRF – Short pauses	133.86	71.70
TRF – Long pauses	175.13	53.27
CAEP to /da/	241.99	12.43

Table 5.1 shows the mean and standard deviation of P2 latency in TRF-model of cortical responses to continuous speech and CAEP to /da/ in the theta band. P2 latency in TRF-model was significantly different compared to CAEP (Wilcoxon Signed Ranks Test, $p<0.001$), however, P2 was not consistently significant in every individual and the latency is considerably variable across individuals. Although examples are given for the P2 wave here, it should be noted that neither P2, P1 nor N2 were consistently detected in all subjects.

5.3.4 Comparison of detection method sensitivity

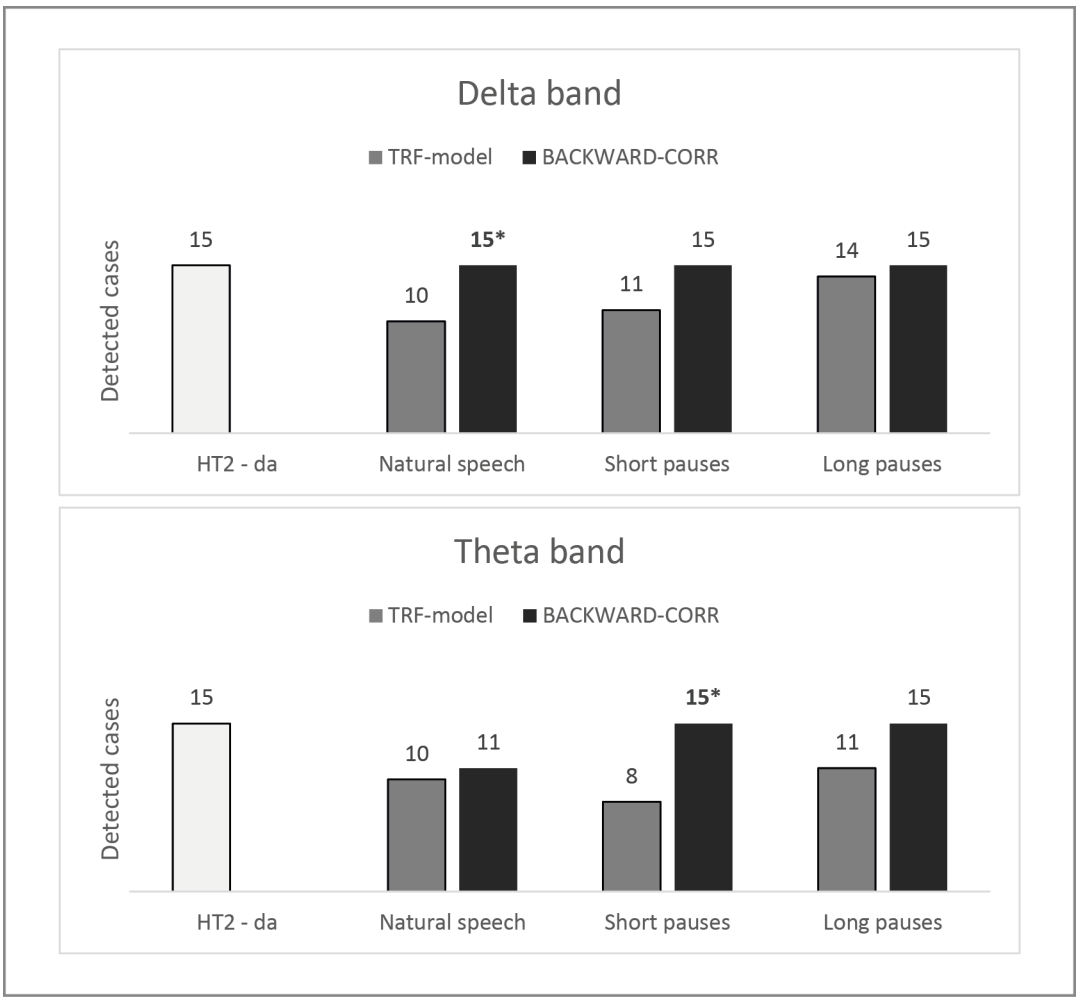

Figure 5.8. **Number of responses detected from the total of 15 subjects (subject 2 excluded) for the Hotelling's T-squared (white), TRF-model (black), and BACKWARD-CORR (grey) detection parameters in the delta and theta band.** The number of detected cases from different detection parameters and speech pauses condition are shown on top of each bar. Stars denote condition where there is a significant difference in number of detected cases between the TRF-model and BACKWARD-CORR only ($p \leq 0.05$, McNemar test) between the TRF-model and BACKCORR for the natural speech condition in the delta band ($p \leq 0.008$) and the speech with short pauses in the theta band ($p \leq 0.031$).

Figure 5.8 shows the number of responses detected from 15 individuals (subject 2 excluded due to missing of EEG in the /da/ condition) using the HT2 on CAEP to /da/ and using TRF-model and BACKWARD-CORR on cortical responses to continuous speech in the delta and theta band. The BACKWARD-CORR appears to outperform the TRF-model, showing consistently greater number of responses detected. However, the difference in number of detected cases was only statistically significant in the natural speech condition in the delta band ($p \leq 0.031$) and in the speech with short pauses condition in the theta band ($p \leq 0.008$), as denoted by the stars in Figure 5.8.

5.3.5 TRF against CA in detecting CAEP to /da/

Figure 5.9 and Figure 5.10 shows the detection of significant absolute peak in the TRF of cortical responses to /da/ (TRF-da) in the delta and theta band respectively. Figure 5.11 and Figure 5.12 shows the detection of significant absolute peaks in the coherent average of CAEP to /da/ (CA-da) in the delta and theta band respectively. For both detection methods, individuals with significant absolute peak in the TRF-da or CA-da is marked with a star. McNemar Test indicates no significant difference in number of detected responses between TRF-da and CA-da. Number of responses detected of TRF-da against CA-da (from total of 15); 13 vs. 12 in the delta band and 14 vs. 14 in the theta band.

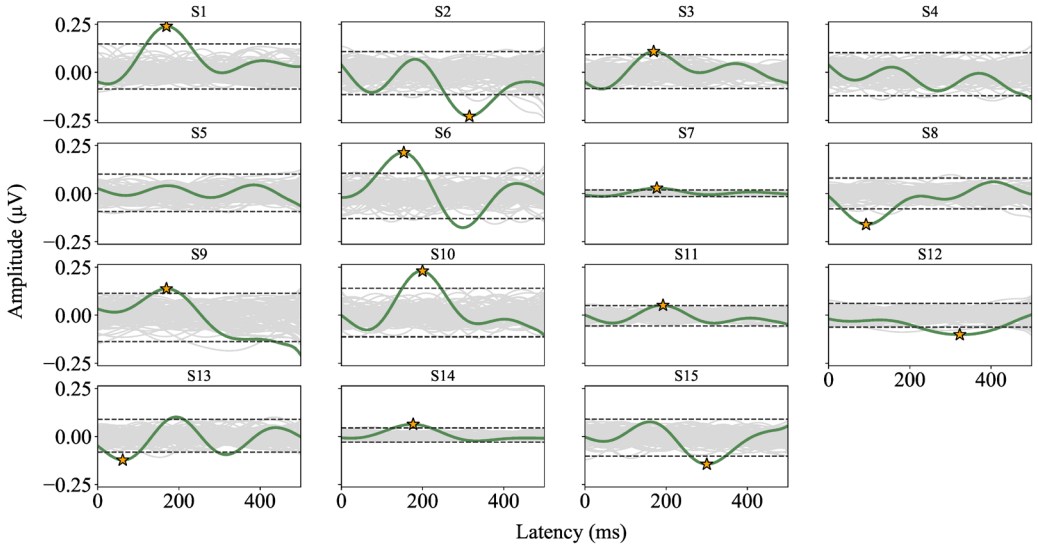

Figure 5.9. Detection of maximum absolute peak with statistical significance in the TRF of CAEP to /da/ (TRF-da) in the delta band for each individual. Light grey plots show the bootstrapped cortical responses to /da/ TRF-models.

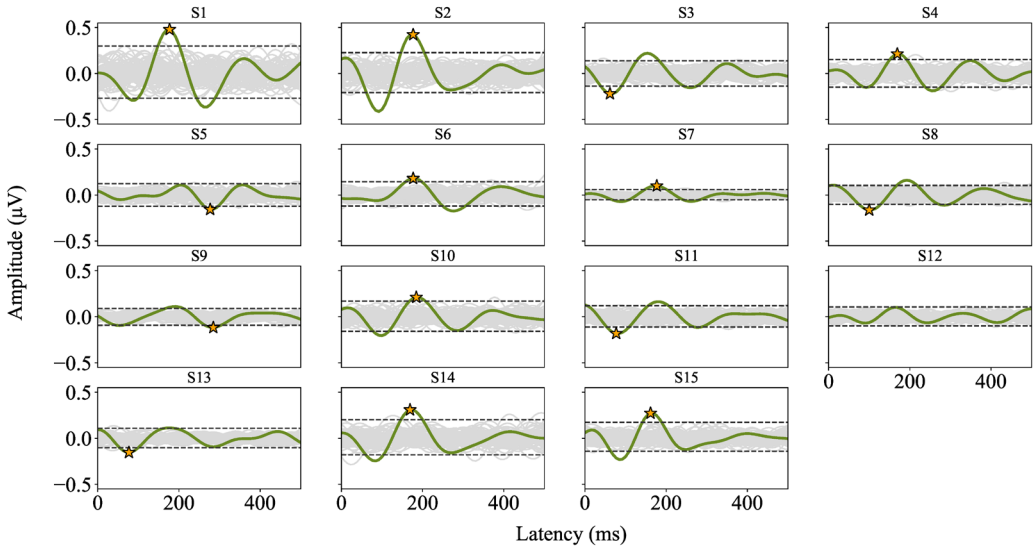

Figure 5.10. Detection of maximum absolute peak with statistical significance in the TRF of CAEP to /da/ (TRF-da) in the theta band for each individual. Light grey plots show the bootstrapped cortical responses to /da/ TRF-models.

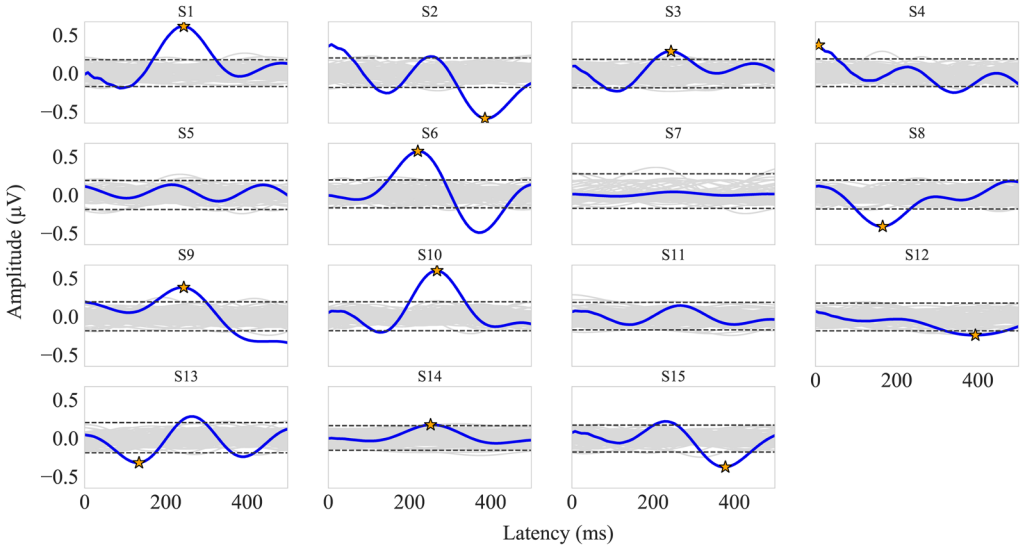

Figure 5.11. Detection of maximum absolute peak with statistical significance in the coherent averages of CAEP to /da/ (CA-da) in the delta band. Light grey plots show the bootstrapped coherent averaged of CAEP to /da/.

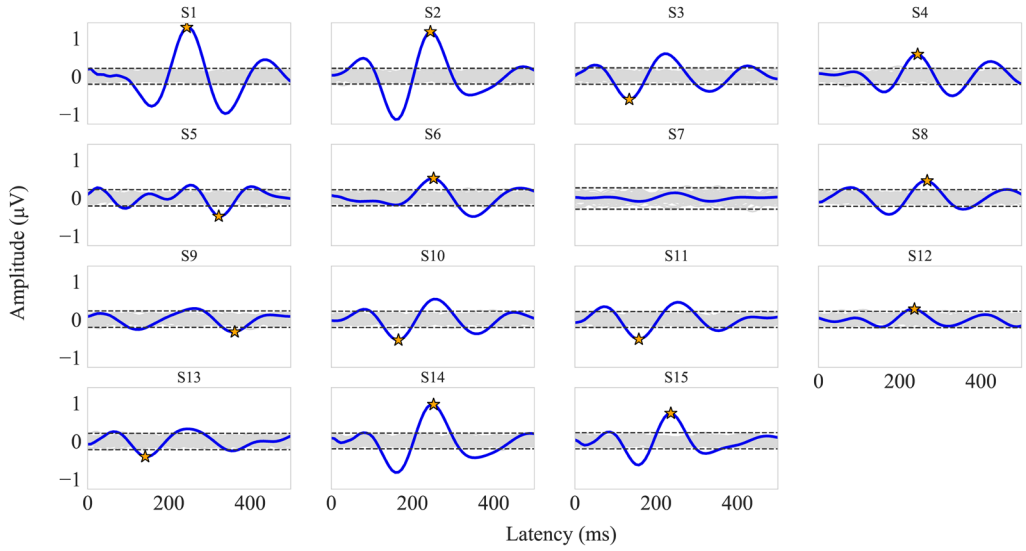

Figure 5.12. Detection of maximum absolute peak with statistical significance in the coherent averages of CAEP to /da/ (CA-da) in the theta band. Light grey plots show the bootstrapped coherent averaged of CAEP to /da/.

5.3.6 TRF of responses to /da/ and CAEP

Figure 5.13 and Figure 5.14 shows that the TRF-da shape appears to be similar to the CA-da waveform, as measured through Pearsons' correlation coefficient (approximately r=0.99 for almost every individual except for subject 5). However, the TRF-da needs to be shifted (or CA-da shifted) by approximately 70 ms before measuring the correlation to obtain a strong correlation coefficient. The 70 ms time shifting was due to the 50 ms delay between stimulus trigger and actual /da/ sound. The remaining 20 ms difference is presumably due to the delay where the stimulus intensity reach it peaks during the vowel /a/.

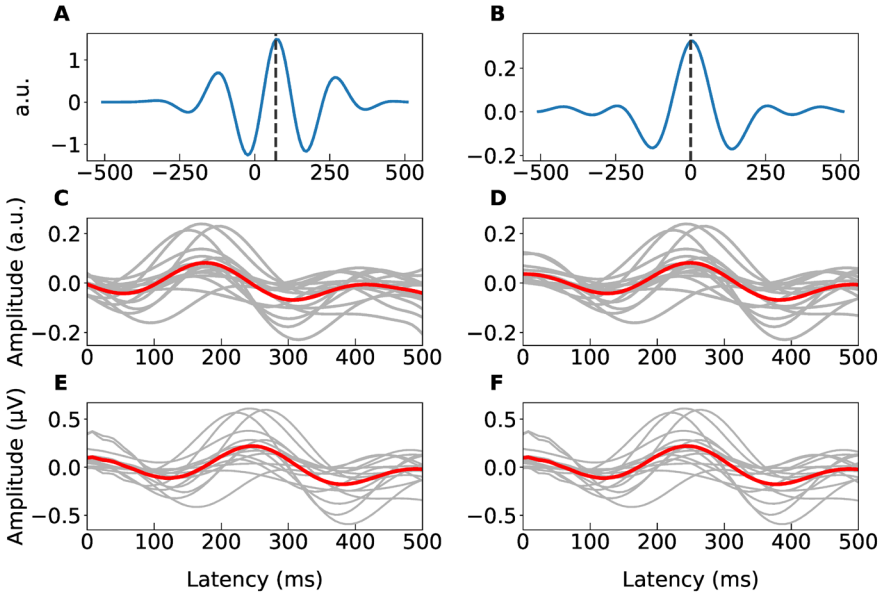

Figure 5.13. Latency shift between the TRF-da and coherent average of CAEP to /da/ in the delta band. Plots in the left-hand column shows the introduced time lag of approximately 70 ms (A) and the misalignment between the TRF-da (C) and the coherent average (E) of CAEP to /da/. Plots in the right-hand column shows the corrected time lag (B) and the alignment between the TRF-da (D) and the coherent average (F) of CAEP to /da/. The TRF-da and CAEP from individuals are shown in grey plots, red plots are the averaged across the cohort.

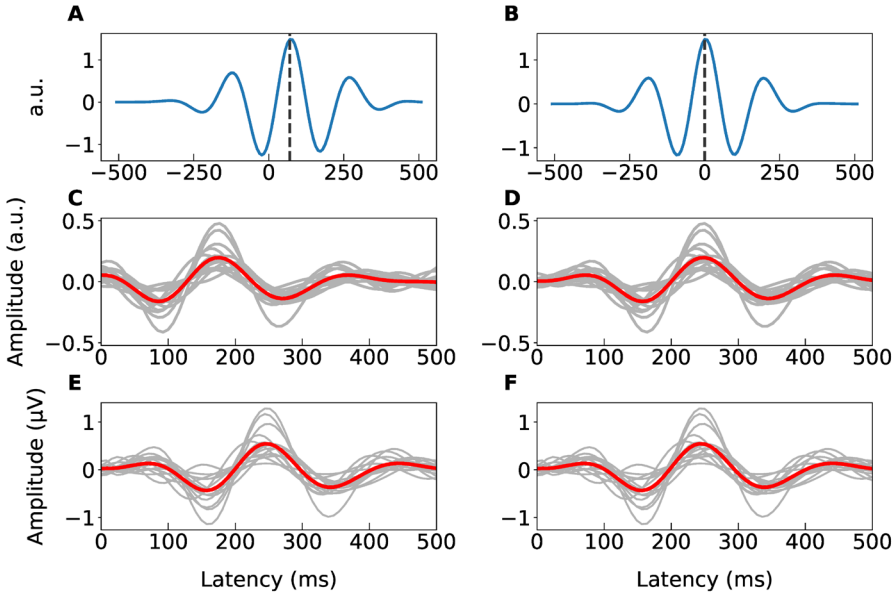

Figure 5.14. Latency shift between the TRF-da and coherent average of CAEP to /da/ in the theta band. Plots in the left-hand column shows the introduced time lag of approximately 70 ms (A) and the misalignment between the TRF-da (C) and the coherent average (E) of CAEP to /da/. Plots in the right-hand column shows the corrected time lag (B) and the alignment between the TRF-da (D) and the coherent average (F) of CAEP to /da/. The TRF-da and CAEP from individuals are shown in grey plots, red plots are the averaged across the cohort.

5.4 Discussion

The primary aim of this study was to determine whether the BACKWARD-CORR or TRF-model is more sensitive for detecting cortical responses to continuous speech. This study clearly showed that the BACKWARD-CORR was more sensitive as a detection method for cortical responses to continuous speech than the forward TRF model peaks. Moreover, the BACKWARD-CORR is more sensitive to the effect of pauses in the continuous speech than the TRF-model, where the significant change in cortical response is shown even when short pauses were added to the speech stream. So, the BACKWARD-CORR will be chosen for detection of cortical responses to continuous speech in the next study (chapter 6). It is worth noting that the number of detected cases for cortical responses to natural speech is similar to the CAEP to /da/ when using BACKWARD-CORR but the CAEP to /da/ measurement time is considerably shorter (~3.5 minutes against ~15 minutes). This suggest that cortical responses to simple repeating speech sound may contain clearer/stronger onset responses and hence are easier to be detected than cortical responses to natural speech.

The secondary aim was to examine the characteristics of cortical responses to continuous speech and CAEP to /da/. The main finding is that although in the group level the TRF-models showed significant increase in peaks amplitude when additional pauses were added to the continuous speech, individual TRF-models did not appear to show significant increase in peaks amplitude for all participants as in the grand averaged TRF-model. The coherent averages of CAEP to /da/ stimulus appears to exhibit more consistent morphology across individuals. This demonstrated that measurement of cortical responses to continuous speech appears to be more variable across individuals than CAEP to /da/, and interpretation of model might be misleading.

Taken together, the results from the primary and secondary aim showed that the BACKWARD-CORR is more sensitive in detecting cortical entrainment to speech envelope than the TRF-model if the goal is only to detect a response. However, the choice for analysis method hugely depends on the experimental context. For assessment of speech comprehension, the TRF-model may be needed as a complement to indicate whether there is a significant difference in response characteristics between cortical responses to sound in different experimental conditions or group of subjects (Holdgraf *et al.*, 2017; Van Canneyt, Wouters and Francart, 2021) . The BACKWARD-CORR may be more preferable for determining individual's selective attention to a speaker of interest, also known as cocktail party problem (O'Sullivan *et al.*, 2015; Etard *et al.*, 2019).

5.4.1 TRF-model and BACKWARD-CORR to detect cortical responses to continuous speech

Detection of cortical responses to continuous speech based on BACKWARD-CORR was better than significant peak in the TRF-model might be due to the advantage of controlling the covariance between stimulus feature and multichannel EEG rather than single channel. The backward TRF modelling may give more weights to the model over several time lag, due to the difference in response delay across EEG channels. Where these weights may not directly have physiological relevance, unlike the TRF-model where the model weight and time lag are physiologically related (Haufe *et al.*, 2014). One possibility why TRF-model performed worse is that the EEG channel with the strongest cortical response vary across the individuals, it is not necessarily channel Cz. An alternative method is to construct the TRF-model for each EEG channel to find the channel with the greatest peak amplitude then apply the bootstrapping test to determine the significance of that TRF-model. Other studies suggested that the best channels for measuring cortical envelope entrainment are the 6-20 frontal-central channels neighbouring channels around Cz (Montoya-Martinez *et al.*, 2021; Aljarboa, Bell and Simpson, 2022). The TRF-model may be improved for response measurement by including multidimensional stimulus features instead of a univariate broadband speech envelope. Di Liberto, O'Sullivan and Lalor (2015) and Drennan and Lalor (2019) shown that other representations in the continuous speech, for example, the time-aligned sequence of phonemes or phonetic features and the multivariate envelope representing different intensity level can improve the TRF modelling.

5.4.2 TRF and CA on cortical responses to /da/

The TRF and CA waveform of CAEP to /da/ showed identical response morphology, this agrees with the result from Lalor *et al.* (2009) where they applied the two methods showed identical CAEP to repeating tones. This study further assessed the sensitivity of response detection for the two detection methods. It is found that the difference in number of detected cases was not statistically significant. This suggests that the stimulation methods (repeating short stimuli or continuous stimuli) and, as a result, the nature of the response makes much more difference to the detection of responses than the choice of detection method.

It is possible that number of detected responses were lower with continuous speech using TRF-model because there were less/weaker onset responses compared to CAEP and

cortical response to speech with pauses. However, not many studies have investigated the exact contribution of onset, non-onset, off-set, or sustained segments in the continuous speech to the measurement of cortical envelope entrainment to be able establish a firm conclusion (Brodbeck, Presacco and Simon, 2018; Weissbart, Kandylaki and Reichenbach, 2020).

The assumption of linear relationship between the response and stimulus feature may be another reason that cortical envelope entrainment is modelled poorly through the TRF approach using either forward or backward direction. Some studies showed that non-linear modelling approaches, such as neural networks, can improve the modelling of cortical envelope entrainment especially for neural responses at higher frequency (≥16Hz) (Pasley *et al.*, 2012; Yang *et al.*, 2015). Non-linear models also outperform linear models in backward modelling applications, such as identifying the attended speaker (de Taillez, Kollmeier and Meyer, 2020) and predicting behavioural speech intelligibility using EEG responses (Accou *et al.*, 2021). A non-linear forward model proposed by Keshishian *et al.* (2020) provides improved stimulus-response modelling performance against the conventional forward linear model and maintains interpretability of the model. However, the main drawback of non-linear modelling is the amount of data that is needed for model training, approximately 40 minutes of EEG responses (Thornton, Mandic and Reichenbach, 2022), which is considerably greater than for linear models. The amount of data required for model training may be reduced by removing less important parameters in the non-linear model (Accou *et al.*, 2023). For invasive electrophysiological measurements, a linear model appears to be sufficient for modelling the response and stimulus due to the direct measure on the source of response resulting in a high SNR recording (Hamilton, Edwards and Chang, 2018; Oganian and Chang, 2019).

In some case, the use of forward TRF and CA on CAEP leads to different experimental conclusion. For example, in a study by Reetzke, Gnanateja and Chandrasekaran (2021), the forward TRF on CAEP to tones was not sensitive to group difference in language proficiency but was significantly sensitive to attention task (attending to speech ignoring tones and vice versa), while CA showed significant effect in both conditions. They suggested that this is due to the regularisation in the forward TRF which affects the amplitude of the model weights to be lower than CA.

5.4.3 Interpreting TRF-model of cortical responses to continuous speech

The grand averaged TRF-model of cortical responses to continuous speech showed overall increase in amplitude and decrease in latency for the dominant peaks when the duration of pauses is longer (as shown Figure 2). The individual TRF-model, however, showed considerably variable amplitude and latency in the dominant peaks especially in the natural speech condition, with many of them not being statistically significant. Other studies have not presented individual cortical TRF-models as it is done in this study, normally the grand averaged TRF-model is presented which is difficult to relate the current findings to other studies. Maddox and Lee (2018) presented individual TRF-models but for auditory brainstem responses (ABR). They demonstrated that the ABR to continuous speech can be measured through the TRF-model and the morphology of ABR to continuous speech is more variable compared to ABR to click sounds.

Although the coherent average waveform and TRF-model shows similar morphology, CAEP and cortical response to continuous sound stimulus are suggested to be generated from different neural sources. An invasive electrophysiological study by Hamilton, Edwards and Chang (2018) has shown that there are distinct onset and sustained responses to continuous speech in the Superior Temporal Gyrus. Obviously, surface EEG measurements would not be as sensitive as invasive electrophysiological measures to clearly observe these two types of responses. Hence, the TRF-model is likely to represent limited components of the underlying neural responses (Lalor *et al.*, 2009), and those responses may mostly be the onset response, as shown in the previous and current chapter. The choice of stimulation method also depends on the goal of the research. If the goal is only to detect a response, the /da/ stimulus is more effective for response generation and may be sufficient. The question then is whether cortical responses to /da/ represents speech perception (which is unknown at this point). For more advance electrophysiological studies involving language and comprehension, a more natural stimulus such as sentences, phrases, or narrative speech is probably needed over monosyllables or words.

5.4.4 Current findings and considerations for speech in noise study

Since the next study will involve measurement of cortical responses to sound stimuli with varying level of background noise, the findings in this chapter led to a consideration to choose the best method in measuring the responses. It is shown in this chapter that the BACKWARD-CORR was more efficient in detecting cortical responses to continuous

speech and more sensitive to the effect of additional pauses in speech than the TRF-model. Hence the BACKWARD-CORR will be selected as the main detection method in the next study, with the anticipation that the method will be more sensitive to detect cortical responses to continuous speech with background noise presented.

The repeating /da/ stimulus will also be included in the next study since it was seen to be highly efficient in generating detectable responses. This suggests that CAEP to /da/ might also be an effective method to measure cortical responses to sound with background noise. If the SRT mainly depends on audibility, then the complexity of using a natural speech stimulus as opposed to a short speech sound may not be needed.

A study by Muncke, Kuruvila and Hoppe (2022) conducted a study using parameters in the TRF-model as an objective measure to predict SRT in normal hearing adults. They showed that the TRF-model was sensitive to changes in cortical responses to continuous speech at SNR close to the speech reception threshold. The root mean square value in the TRF-model could be used to predict SRT with prediction error less than ±2 dB in 16 out of 18 participants. However, the TRF-model was not effective at the individual level for the dataset used in the current study. If the cortical response cannot be detected even in condition without background noise, the speech reception threshold might not be possible to predict or predicted poorly.

There is a considerable inconsistency on how parameters from both forward and backward TRF modelling relate to behavioural speech intelligibility, although the relation between TRF parameters and some experimental conditions are better established, such as attention to target stimulus (O'Sullivan *et al.*, 2015; Biesmans *et al.*, 2017; Etard *et al.*, 2019). A number of studies report that attention to a target stimulus increases the cortical entrainment to speech envelope but the stronger cortical envelope entrainment does not necessarily indicate better speech-in-noise performance (Song and Iverson, 2018; Zou *et al.*, 2019).

5.5 Conclusions

This chapter explored whether the TRF-model is more sensitive than the BACKWARD-CORR for detecting cortical responses to continuous speech. The number of detected responses was greater for the BACKWARD-CORR than the TRF-model in every speech pause condition both in the delta and theta band. Therefore, the BACKWARD-CORR will

be the main detection method with the anticipation that the method will be effective for measuring cortical responses to continuous speech with varying SNR.

The TRF-model of cortical responses to continuous speech in the theta band showed similar P1-N1-P2 complex which also appears in the CAEP to /da/. Significant difference between P2 in the TRF-model and CAEP to /da/ was found, but due to the variability of peaks amplitude and latency in the TRF-model, it was decided that conclusion regarding the characteristics between responses to continuous speech and /da/ stimulus should not be made from this dataset as they might be misleading.

For clinical applications, if the goal is to just detect responses, using BACKWARD-CORR is more effective than using TRF-model and it is sufficient as the interpretation of response model is not necessary. However, the interpretability of the TRF-model would provide more insights on how change in stimulus properties affects the encoding of sound by the auditory system, as reflected by amplitude and latency of peaks in the waveform. Using simple /da/ stimuli may be more efficient than using continuous speech for detecting auditability, but it is not yet clear how much the stimulus needs to be speech-like in order to predict speech intelligibility.

Chapter 6 Experiment 2: Predicting behavioural speech reception threshold using cortical responses to continuous sound

6.1 Introduction

It has been demonstrated that the strength of cortical responses is correlated with stimulus SNR, stimulus with lower SNR generates weaker responses, and speech-in-noise performance can be predicted using objective measures of cortical responses to either speech and non-speech stimuli (either presented as a continuous signal or repetition of discrete stimulus) (Du *et al.*, 2011; Billings *et al.*, 2013; Vanthornhout *et al.*, 2018). However, it is not clear whether using cortical responses to continuous speech will be more beneficial (e.g., providing more accurate prediction of SRT) than using responses to other types of stimuli. The literature normally suggests continuous speech to be preferable to short repeating sound for studies relating cortical responses to sound and speech-in-noise performance, but this suggestion is based on logic that speech is more face valid in terms of a real-world stimuli rather than providing empirical evidence to support. Moreover, the prediction of speech-in-noise performance appears to be accurate at a group level but uncertain if this is also true at an individual level.

In previous studies in this thesis, improvements have been made in terms of modifying speech stimuli and selecting parameters in order to best detect responses to speech. Building on the work of Vanthornhout *et al.* (2018), Lesenfants *et al.* (2019), and Verschueren, Vanthornhout and Francart (2020) who showed that the behavioural SRT can be predicted from the objective measures to cortical responses to continuous speech, the current study focussed on the question of how the complexity of the stimulus used links to prediction of speech intelligibility. Three stimuli were used varying in complexity, comprising of continuous speech (most complex), broadband noise which has the same envelope as the continuous speech but unintelligible, and finally repeating /da/ stimuli which are the simplest stimuli but have some properties of speech. It is expected that the current study would answer the question as to whether the measurement of cortical responses to speech or non-speech stimuli could be simply a measurement of audibility (in which case using either stimuli may not make a difference to the inference), or the measurement of speech intelligibility does depend on the type of stimuli used.

The aim of this study was to assess the applicability of using objective measures of cortical responses to sound to predict the behavioural speech reception threshold (SRT) for individuals with no requirement for participant's attention on the stimuli. The reason for the focus on non-attended stimuli is for potential application to infants or children in the longer term, where alertness is easier to control, for example with a video, than attention to a stimulus. The applicability of using cortical responses to sound to predict SRT will be assess through 1.) the absolute prediction error between the SRT and threshold estimated from the objective measures or correlation threshold (CT) and 2.) number of cases where the SRT could not be predicted from CT. The findings from this study could help bridge the connection between the use of cortical responses to speech and non-speech stimuli to objectively predict behavioural speech-in-noise performance in individuals.

If the use of cortical entrainment to continuous speech and modulated noise envelope showed no difference in SRT prediction performance, then the cortical entrainment to continuous sound envelope may simply be a representation of sound perception (similar to PTA). Therefore, it may not be the best feature to indicate speech intelligibility in a more diverse group of subjects such as those who do not understand a particular form of speech. Then the use of modulated noise could be considered advantageous over the use of continuous speech, as it is not language specific. The modulated noise could then be used with no requirement for language proficiency (applicable in children) and may be easier to ignore (less confounded by cognitive functions). The use of CAEP to /da/ would further emphasise this investigation to show if the objective measure to a much simpler form of speech will be sufficient to predict the SRT.

6.2 Methods

6.2.1 Participants

Twenty native English speakers aged between 18 to 40 years old (an average of 23 years old) participated in this study (15 male and 7 female). All participants had self-reported normal hearing (hearing threshold below 25 dB HL for 250 Hz until 8 kHz) and no history of neurological disorder. The study was approved by the University of Southampton Ethics Committee (ethics reference number ERGO: 52472). All participants provided written informed consent prior to the experiment.

6.2.2 Behavioural experiment

The British version of the Oldenburg Matrix Test was used to evaluate the participant's behavioural speech-in-noise performance (Kollmeier *et al.*, 2015). Each Sentence in the British Matrix test contains 5 words with the fixed structure of "Name – Verb – Numeral – Adjective – Object" (e.g., Peter got three large desks). Each position in the sentence has 10 alternative words which are randomly selected to form a testing sentence. Five signal-to-noise ratio levels for testing were -15, -10, -5, 0, and 20 dB (without background noise or in quiet). The Matrix sentences were presented at 65 dBA in quiet. A speech-shaped noise was used as background noise also presented at 65 dBA. The level of noise remains constant while the level of speech varies according to the SNR level.

Each participant was presented with 20 sentences at each SNR. Scoring for each sentence was based on the number of correct words in percentage (word scores). The SRT was estimated by fitting the averaged percentage of correct words at each SNR with a sigmoid function in equation (6.1) (Vanthornhout, Decruy and Francart, 2019) .

$$S(SNR) = \frac{\gamma}{1 + e^{-\frac{SNR-\alpha}{s_{50}}}} \tag{6.1}$$

where S is the word scores as a function of SNR, γ is the maximum averaged score, α is the midpoint of SNR, and s_{50} is the slope at α.

6.2.3 Stimuli

In this study, three types of sound stimulus were used, a continuous speech, speech envelope-modulated noise, and /da/ as the signal. The purpose of using a modulated noise that has a similar intensity envelope to the continuous speech is to create a stimulus as identical as possible to the continuous speech, e.g., similar temporal waveform and long-term spectrum, but unintelligible. By having the same amount and duration of pauses in the speech and modulated noise, the strength of cortical envelope entrainment should not be biased towards any stimulus condition. As it remains unclear how speech with additional pauses inserted between words affects different aspects of speech perception (e.g., intelligibility and comprehensibility), the interpretation of results and comparison with findings from other studies that use natural continuous speech may be confounded. Therefore, speech with additional pauses was not employed in the current study.

The continuous speech stimulus was extracted from of part 1, chapter 2 and chapter 5 of 'The Children of Odin – by Pádraic Colum' audiobook read by a female narrator (available to download for free on https://librivox.org/the-children-of-odin-by-padraic-colum/). The original audiobook file sampling rate was 22,050 Hz in an mp3 format (energy band limited to 0 – 12,000 Hz) and was converted to a wav file with a new sampling rate of 44,100 Hz.

A speech-shaped noise, the carrier signal, was created by generating a Gaussian white noise filtered by the long-term average spectral shape of all continuous speech stimulus used in this study, with the bandwidth of 0 – 12,000 Hz. The modulated noise was then generated by multiplying the speech-shaped noise with the amplitude envelope of the continuous speech, the modulating signal. The amplitude envelope of the continuous speech contains the series of amplitude values in the continuous speech stimulus, the amplitude envelope was smoothed by weighted averaging each amplitude sample over their speeling samples. All of these procedures were done using standard functions in Praat software (Boersma and Weenink, 2001). The speech-shaped noise was also used as the background noise.

A monosyllable /da/ stimulus with a duration of 40 ms was used for the measurement of CAEP. The stimulus was presented to each participant 200 times, alternating between rarefaction and condensation polarisation (100 per polarity), with an inter-stimulus interval of 1.11 s. The total duration of 200 repeating /da/ was approximately 4 minutes.

All stimuli were calibrated using Bruel and KjAERs Type 4230 sound level calibrator (Nærum, Denmark). Presentation of stimuli to the participants were controlled by a MATLAB script. Stimuli were presented through RME Fireface UC audio interface and Etymotic ER-2A inserted earphones (Etymotic Research, Inc., Illinois, USA).

6.2.4 EEG experiment

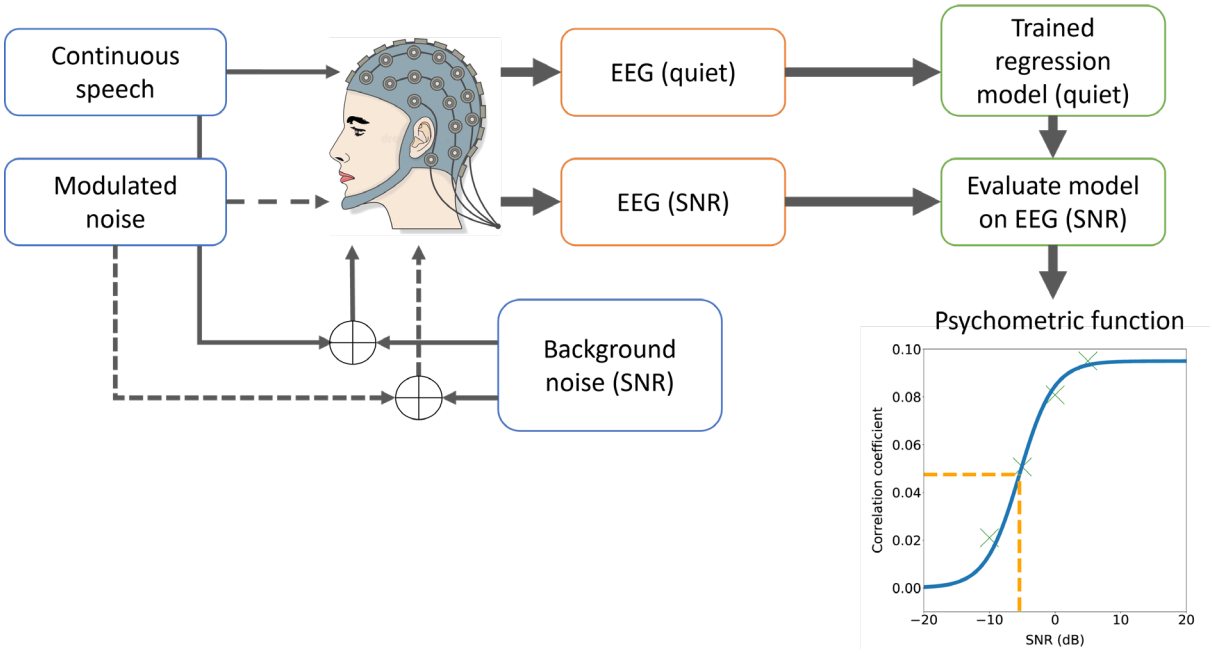

Figure 6.1. Overview of the EEG experimental setup for specifically for the continuous speech and modulated noise condition. Cortical responses in the quiet condition were used to train the backward TRF model. The trained model was then tested on cortical responses at 5 SNR conditions to obtain the BACKWARD-CORR. A sigmoid function was fitted to the 5 BACKWARD-CORR and the SRT was predicted at the steepest gradient of the fitted function (indicated by the dashed orange lines).

EEG measurement was conducted after the behavioural test. EEG data were recorded using the BioSemi ActiveTwo 32-channel system (BioSemi, Amsterdam, Netherlands). The sampling rate was 2,048 Hz. An antialiasing filter applied during the recording; the EEG data bandwidth was 417 Hz.

Prior to the experiment, participants were instructed to avoid listening to the speech signals and pay attention to a muted documentary with subtitles presented on a screen. The choice of a documentary over other genre was based on other studies suggesting that movies where lip reading is possible can significantly degrade the quality of response measurement and that the auditory processing is not significantly affected while reading subtitles (Navarra, 2003; O'Sullivan *et al.*, 2016). No questions related to the speech story were asked after the stimulus presentation. No questions about the documentary were asked to encourage the participants to attend. The EEG experiment consist of three experimental conditions varying in terms of speech stimuli, continuous speech, modulated noise, and /da/. The 5 testing SNR levels here were the same as in the behavioural experiment. In this EEG experiment, the level of signal was fixed and instead the background noise varied according to the SNR.

An overview of the EEG experiment for the continuous speech and modulated noise conditions is shown in Figure 6.1. The /da/ condition was not included in the figure because it did not involve the TRF approach to obtain the objective measure. In the first part of the experiment, the continuous speech and modulated noise were presented only in quiet at 65 dBA with the duration of 15 minutes for each stimulus. The /da/ stimulus were presented at 65 dBA in quiet and other 4 SNR levels. The order of conditions was randomised and preassigned to the participants, so the number of participants starting with different stimulus conditions are as equal as possible (Dettori, 2010). The second part of the EEG experiment consisted of measurements of responses to continuous speech and modulated noise for 2 minutes with 4 repetitions at each of 5 SNR levels (8 minutes of EEG measurement at each SNR level; -15, -10, -5, 0, and 20 dB SNR). The type of stimulus and level of SNR were randomly presented to the participants. The total duration of the behavioural and EEG experiment together was 3 hours.

6.2.5 Data analysis

The EEG data from each participant in all stimulus conditions were referenced to channel T7 and T8, leaving 30 channels available to be analysed. They were then normalised to give a mean of zero and a standard deviation of 1. EEG data were band-pass filtered using a zero-phase (non-causal) FIR filter (filter length was 6.6 times the reciprocal of the shortest transition band) over the range of 1-4 Hz (delta frequency band) and 4-8 Hz (theta frequency band) and resampled at 128 Hz.

The stimulus envelope of the continuous speech and modulated noise were obtained using the Hilbert transform applied on the original signal at the sampling rate of 44.1 kHz. The Hilbert transform signal was then band-pass filtered using the same filter settings as in the EEG over the range of 1-4 Hz and 4-8 Hz, matching the frequency bands used in the EEG, and resampled at 128 Hz.

6.2.6 Temporal response function of responses to continuous speech and modulated noise

This study utilised both the forward and the backward TRF approach for measuring the cortical envelope entrainment. The correlation coefficient between the actual and reconstructed envelope from the backward TRF (BACKCORR-CORR) was the objective measure for predicting the SRT using cortical responses to continuous speech and modulated noise. The forward TRF model weight (TRF-model) was only used to further

assess the characteristics of the cortical responses but no parameters from the TRF-model was used for predicting the SRT. A detailed explanation for the forward and the backward TRF approach is in section 3.2. This section will only describe the main TRF modelling parameters this study.

For the backward TRF, the cortical responses to continuous speech and modulated noise in quiet were used as the training data set for the backward model. The backward model train on cortical responses in quiet condition was then used to train on the four 2 minutes responses measured at 5 SNR levels to obtain four BACKWARD-CORR at each SNR level. The lambda value (λ) for the regularisation was fixed at 355. No cross-validation was performed to adjust this model parameter, as it was usually done in previous chapters, for equality between measurements. The BACKWARD-CORR was calculated over the range of 0-300 ms time lag. The prediction of SRT was done by using the same sigmoid function (see equation 1) that was used for the behavioural SRT estimation to fit the BACKWARD-CORR over the 5 SNR levels. The sigmoid function was fitted to the 5 BACKWARD-CORR values using the nonlinear trust region reflective algorithm. Bound constraint for α (the midpoint of predicted SNR) was between -100 and 100. Henceforth, the predicted SRT from the backward TRF approach will be refer to as the correlation threshold (CT).

The TRF-model was applied on cortical responses in the quiet conditions only to explore the characteristics of responses to stimuli with identical amplitude envelope. The λ for the regularisation was fixed at 355, same value used in Chapter 5. The TRF-model was calculated over the time lag range of -100 to 500 ms.

6.2.7 CAEP to /da/

The CAEP to /da/ at each SNR level was obtained through the coherent averaging method. The objective parameter for predicting the SRT from CAEP to /da/ was the amplitude of P2 peak, a positive peak occurring within the 150 to 250 ms latency range. The P2 peak was selected as it was the most prominent peak, largest in amplitude, in the previous study (chapter 5) and it is found to be strongly correlated with the SNR. The amplitude of P2 peak at each SNR was fitted with the sigmoid function as in equation (6.1) to obtain the predicted SRT.

6.2.8 Statistical analysis

For the forward and backward TRF analysis on cortical responses to continuous and modulated noise, the significance level at $\alpha = 0.05$ for TRF-model and BACKWARD-CORR was estimated using the bootstrapping method, modelling misaligned response and stimulus envelope only from the quiet condition (same as in chapter 4 and 5). Friedman tests were used to explore the difference in BACKWARD-CORR across the SNR levels for the continuous speech and modulated noise stimulus condition. Post hoc pairwise comparisons were performed using Wilcoxon singed rank tests.

6.3 Results

6.3.1 Behavioural SRT

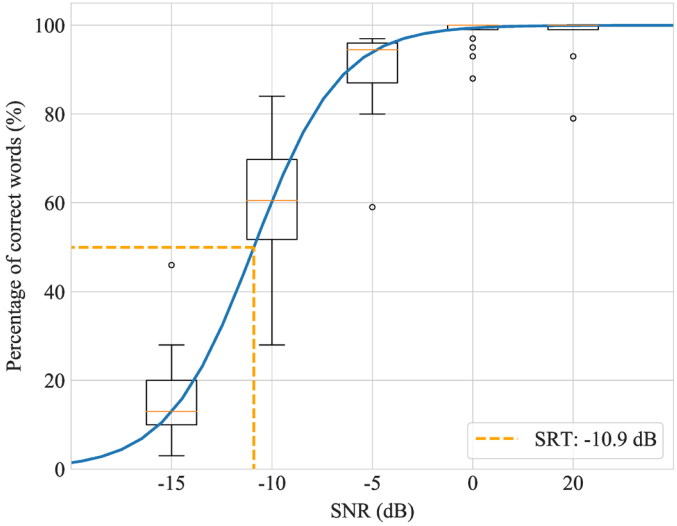

Figure 6.2. Boxplots of percentage of correct words from all participants obtained from the Matrix test as a function of SNR levels fitted with a sigmoid function (blue solid line) through the median values at each SNR. The orange dashed line indicates the estimated SRT at the steepest gradient of the sigmoid function.

The mean behavioural SRT across 20 participants was -10.8 dB with a standard deviation of 1.1 dB, The SRT ranged from a minimum of -12.3 dB to a maximum of -9.3 dB. Figure 6.2 shows the percentage of correct words from all participants as a function of SNR levels. The sigmoid function was fitted to the median of the percentage of corrected words at each SNR to estimate the group level SRT, which is -10.9 dB SNR almost identical to the SRT averaged across individuals. The orange dashed lines indicates the estimated SRT from the maximum value of the derivative of sigmoid function (steepest gradient).

6.3.2 Group and individual level decoder correlation as a function of SNR

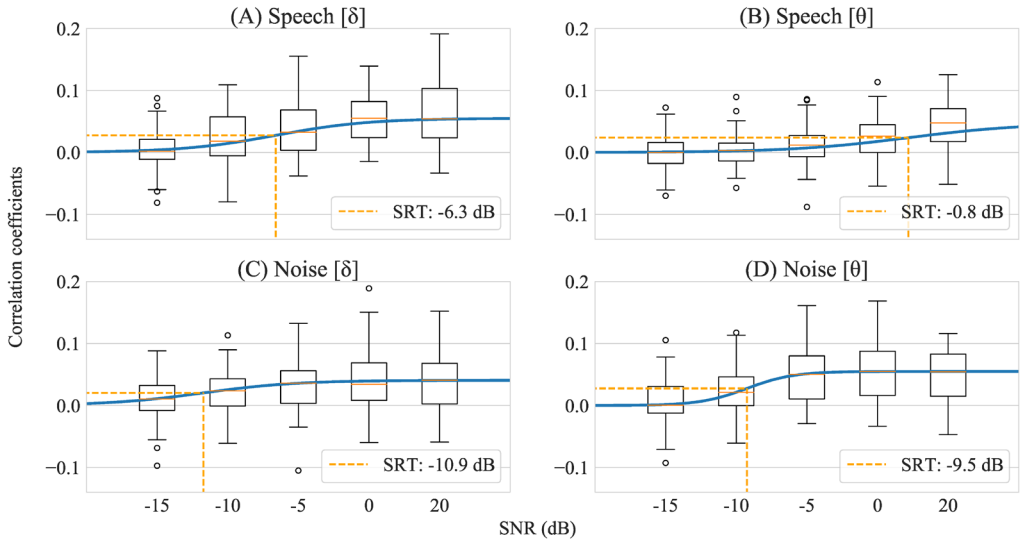

Figure 6.3. **Boxplots of the backward TRF correlation coefficients from all subjects in the speech and modulated noise condition in delta (A and C) and theta (B and D) frequency band as a function of SNR levels fitted with a sigmoid function (blue solid line) through the median values at each SNR.** The orange dashed line indicates the estimated CT at the steepest gradient of the sigmoid function.

Figure 6.3 shows the boxplot of BACKWARD-CORR obtained from all the participants in the speech and modulated noise condition in the delta and theta frequency bands as a function of SNR levels fitted with a sigmoid function (blue solid line) through the median value of each box. Overall, the median of BACKWARD-CORR increases with SNR for both stimulus conditions and both EEG frequency bands. Friedman test showed significant difference in correlation coefficient across 5 levels of SNR ($p \leq 0.001$) in all stimulus conditions and EEG frequency bands.

6.3.3 Significant differences in BACKWARD-CORR between cortical responses to continuous speech and modulated noise at each SNR level

Table 6.1. Statistical significance in BACKWARD-CORR difference between cortical responses to continuous speech and modulated noise at each SNR level. Each cell contains a p-value from the Wilcoxon signed rank test and an indication of stimulus condition with greater BACKWARD-CORR mean in the parenthesis, S for continuous speech and N for modulated noise.

	Significance of difference between BACKWARD-CORR in continuous speech and modulated noise condition: p – value (condition with greater BACKWARD-CORR mean)	
SNR level (dB)	**Delta band**	**Theta band**
20	0.252	0.599
0	0.151	**0.008 (N)**
-5	0.715	**0.004 (N)**
-10	0.978	0.026
-15	0.762	0.720

From Table 6.1, Wilcoxon sign rank test shows significant difference in correlation coefficient between the speech and modulated noise condition at 0 ($p \leq 0.008$) and 5 dB SNR ($p \leq 0.004$) only in the theta band, where the averaged BACKWARD-CORR from cortical responses to modulated noise is greater than from cortical responses to continuous speech.

6.3.4 Predicting SRT using cortical responses to continuous speech and modulated noise

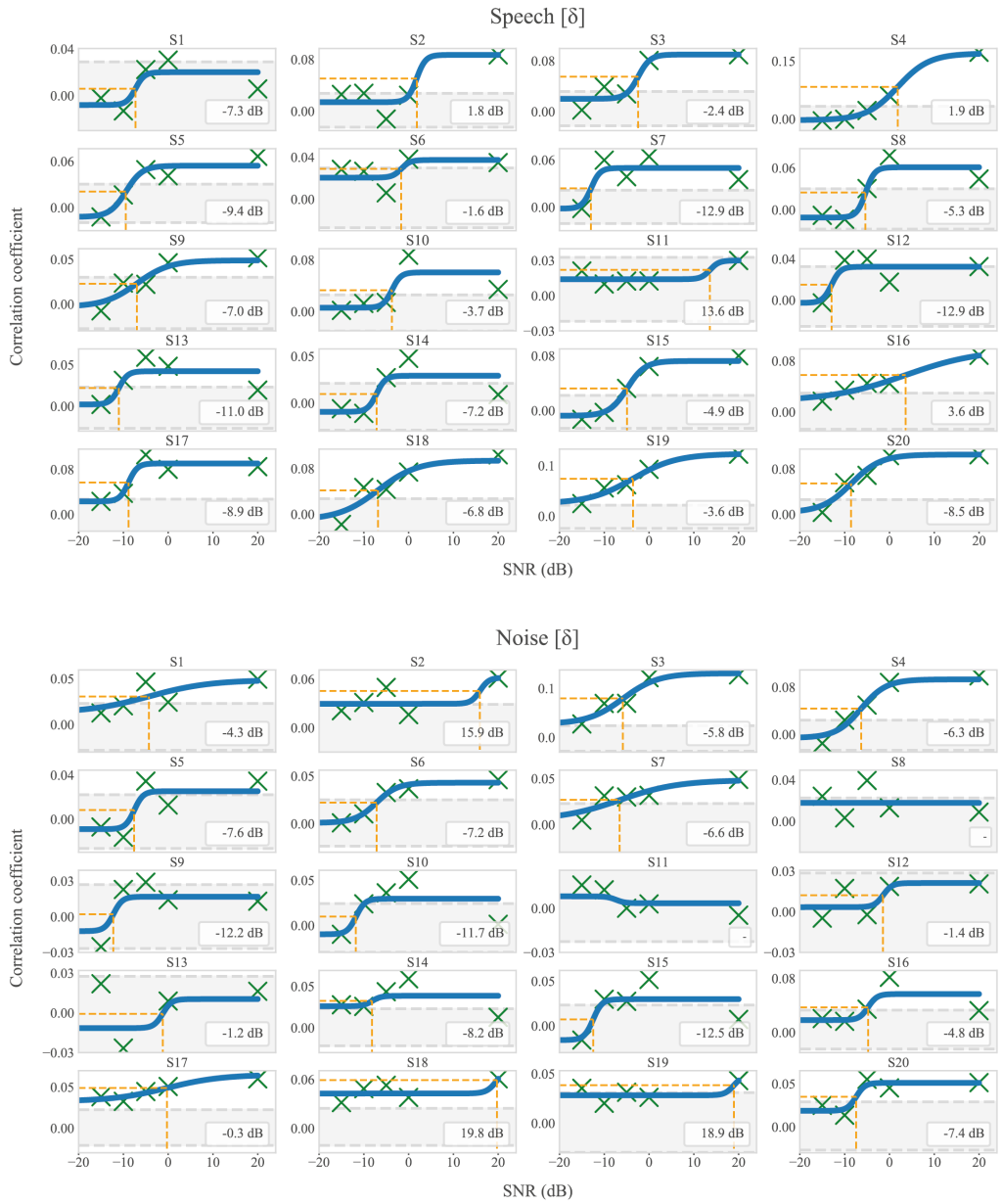

Figure 6.4. Correlation threshold (CT) from 20 participants estimated from the BACKWARD-CORR using cortical responses to continuous speech (top array) and modulated noise (bottom array) in the delta band. Green X points are the averaged BACKWARD-CORR as a function of SNR level. The blue solid line is the sigmoid function fitted to the BACKWARD-CORR. The orange dashed line indicates the CT at the steepest gradient of the sigmoid function. The grey shaded area is the BACKWARD-CORR critical band estimated from cortical responses in quiet condition only.

Figure 6.5. Correlation threshold (CT) from 20 participants estimated from the BACKWARD-CORR using cortical responses to continuous speech (top array) and modulated noise (bottom array) in the theta band. Green X points are the averaged BACKWARD-CORR as a function of SNR level. The blue solid line is the sigmoid function fitted to the BACKWARD-CORR. The orange dashed line indicates the CT at the steepest gradient of the sigmoid function. The grey shaded area is the BACKWARD-CORR critical band estimated from cortical responses in quiet condition only.

Figure 6.4 and Figure 6.5 shows individuals' prediction of SRT using cortical responses to continuous speech and modulated noise in the delta band and theta band respectively. Generally, CT tends to give an overestimation of the SRT of individuals, greater number in dB suggesting worse speech-in-noise performance than indicated from behavioural SRT. However, the CT varies considerably across individuals.

6.3.5 Absolute prediction error and number of individuals that were unable to use correlation threshold (CT) to predict SRT

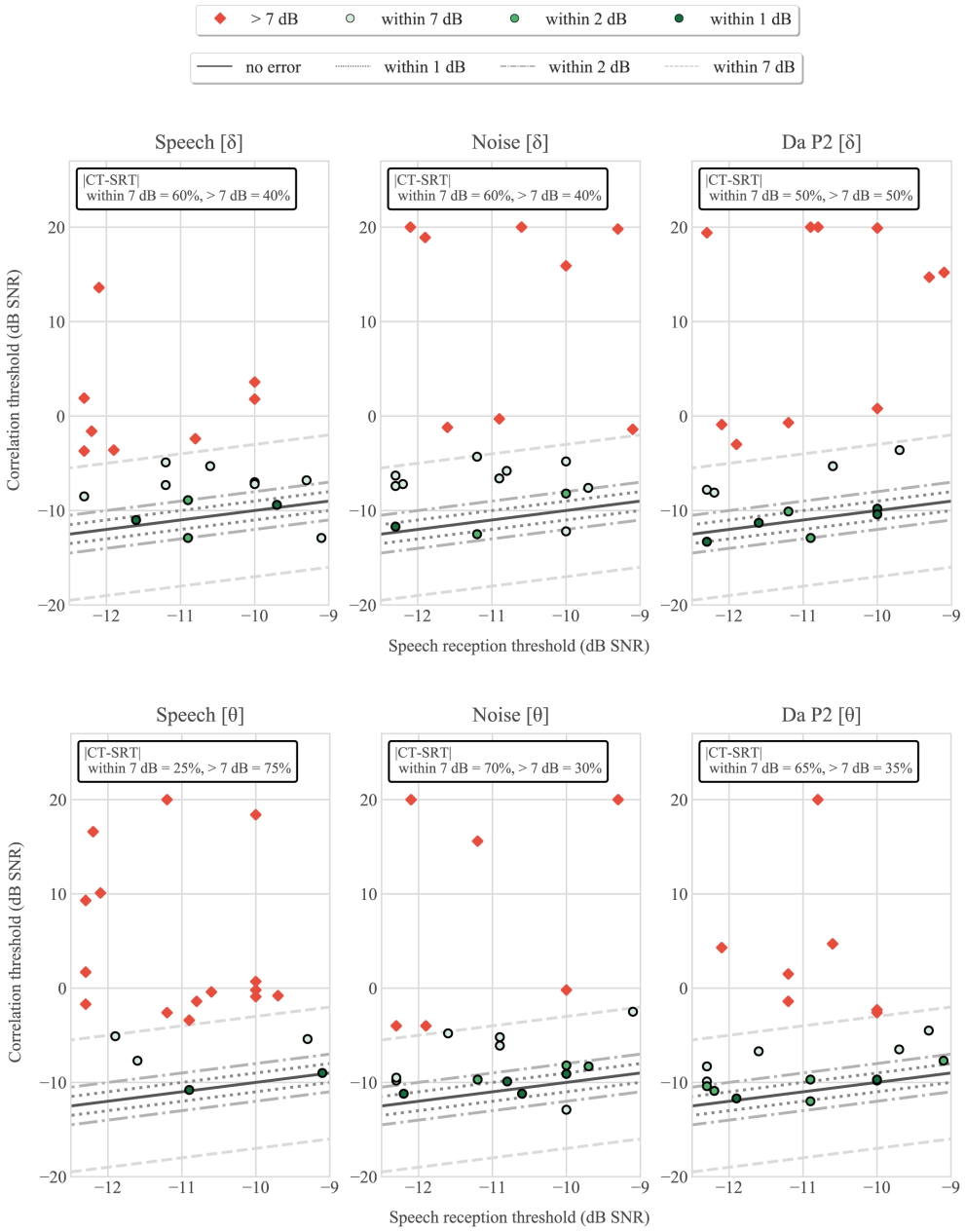

Figure 6.6. Absolute difference between the behavioural speech reception threshold (SRT) and the correlation threshold (CT) obtained from EEG response in the delta [δ] and theta [θ] band, in the top row and bottom row, respectively. Circles and diamonds represent the paired SRT and CT for each participant. Dark green, green, and light green circles indicate an absolute difference within 1, 2, and 7 dB, respectively. Diamonds indicate an absolute difference greater than 7dB. The box labelled |CT-SRT| in each plot indicates the proportion of participants who had an absolute difference within 7 dB and the proportion of participants who had an absolute difference greater than 7 dB. The dark solid line represents zero absolute difference between SRT and CT.

Figure 6.6 the points representing the absolute differences between SRT and CT, within ranges of 1, 2, and 7 dB, across the three stimulus conditions in the delta and theta bands. Additionally, Figure 6.6 shows the points representing absolute differences greater than 7

dB between SRT and CT for each participant. Maximum CT was limited to 20 dB for the purpose of visualisation, to ensure that data from 20 participants are displayed together.

In the delta band, the difference in number of participants with absolute SRT error within 7 dB between the continuous speech and modulated noise conditions (p=0. 623, McNemar test). The same comparison was also not statistically significant between continuous speech and /da/ and modulated noise and /da/ condition was not statistically significant ($p \leq 0.344$ and $p \leq 0.363$ respectively, McNemar test). The continuous speech condition showed identical prediction of SRT in the modulated noise and slightly better than /da/ condition. However, only up to 12 participants could obtain an objective prediction of SRT within 7 dB range of the behavioural result. The /da/ condition performed worst in the delta band, absolute SRT prediction error within 7 dB for 10 participants.

In the theta band, the difference in number of participants with absolute SRT error within 7 dB between the continuous speech and modulated noise, and continuous speech and /da/ conditions was statistically significant ($p \leq 0.019$ and $p \leq 0.008$ respectively, McNemar test). The same comparison was not statistically significant between the modulated noise and /da/ conditions (p=0.5, McNemar test). The modulated noise and the /da/ condition showed better prediction of SRT than in the continuous speech condition, where SRT could be predicted within 7 dB range for 13-14 participants. The continuous speech condition could predict SRT within 7 dB range for only 5 participants.

The number of non-applicable SRT prediction (NA), considered when absolute difference between SRT and CT is greater than 7 dB, was lowest in the modulated noise condition in the theta band (30% from the total number of participants) and highest in the continuous speech condition in the delta band (75%).

6.3.6 CAEP to /da/ P2 peak amplitude as a function of SNR

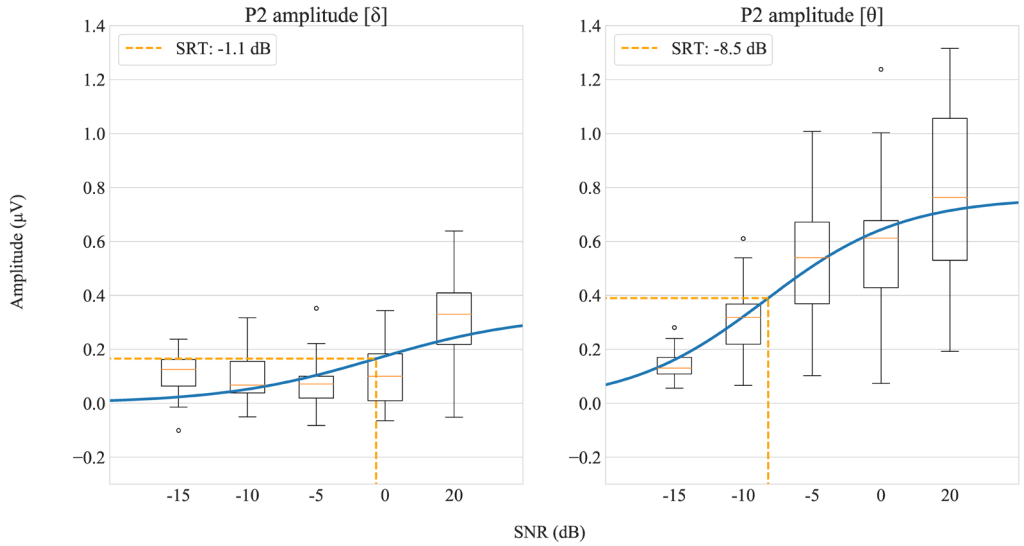

Figure 6.7. Amplitude of P2 peak from the CAEP to /da/ from all participants as a function of SNR level (dB) in the delta (left) and theta band (right). The median P2 amplitudes across SNR levels are fitted with the sigmoid function (blue solid line). The predicted group level SRT is indicate by the orange dashed lines (steepest gradient).

Figure 6.7 shows the amplitude of the P2 peak from the CAEP to /da/ at different SNR level in the delta and theta bands. Friedman test showed significant difference in correlation coefficient across 5 levels of SNR ($p \leq 0.004$ and $p < 0.001$ for the delta and theta band respectively).

6.3.7 Group averaged TRF-model of cortical response to speech and modulated noise in quiet

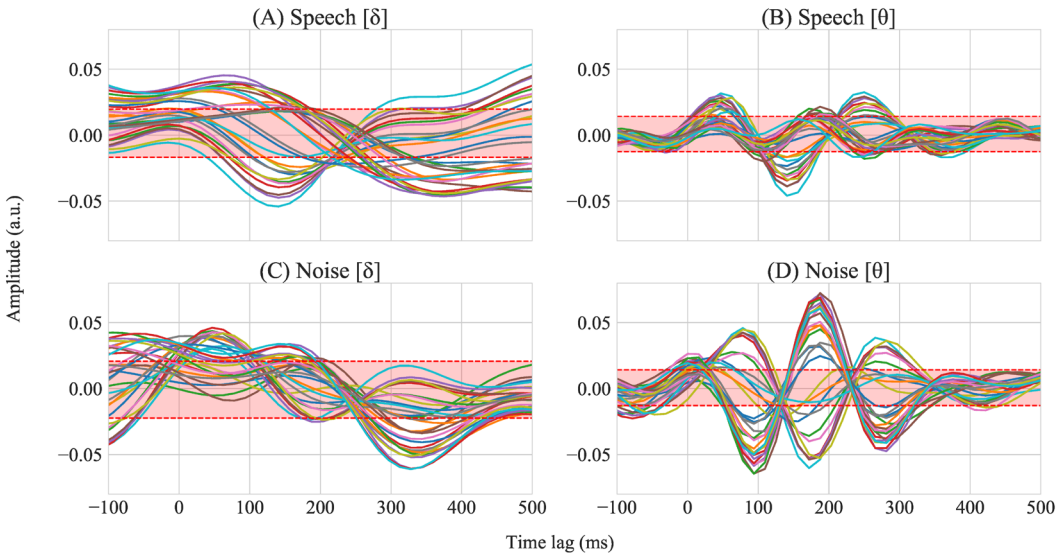

Figure 6.8. The group grand averaged TRF-models from each of the 30 EEG channels in the continuous speech and modulated noise conditions in the delta (A and C) and theta (B and D) frequency band in quiet condition. The red shaped area is the significant threshold estimated through the bootstrapping test.

Figure 6.8 shows TRF-models from each of the 30 EEG channels in the continuous speech and modulated noise conditions in quiet. Each TRF-model was averaged across 20 participants. The shaded area indicates the group averaged, across EEG channels and individuals, TRF-model threshold band obtain from the bootstrapping test where samples in the TRF-model are considered not statistically significant. In the theta band (Figure 6 B and D), a notable difference in peak amplitude between the TRF-models in the continuous speech and modulated noise was at approximately 190 ms.

6.3.8 CAEP to /da/

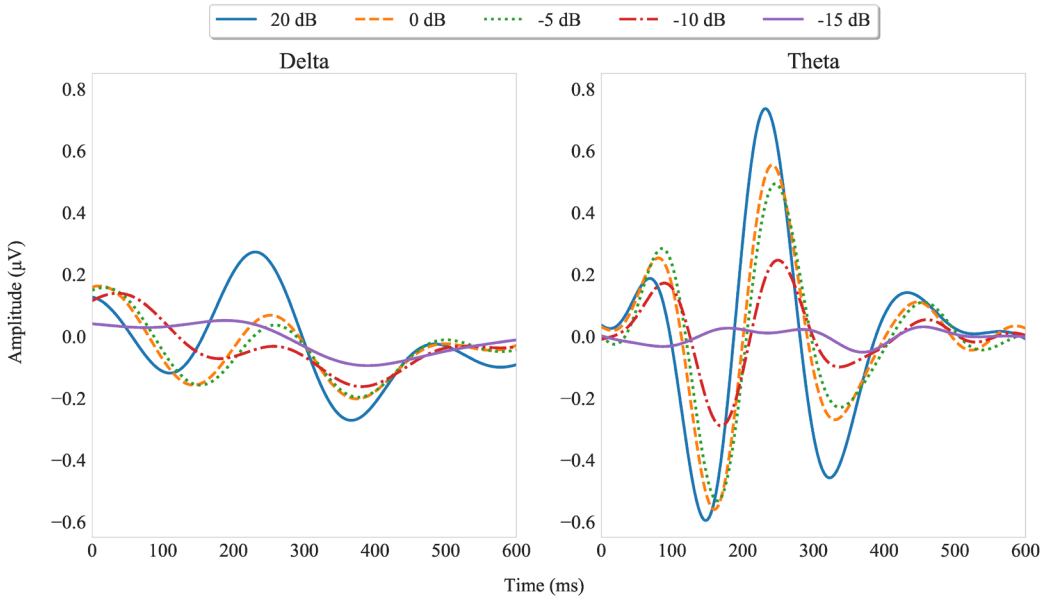

Figure 6.9. The CAEP to /da/ averaged across 20 participants at each SNR level in the delta (left) and theta (right) band.

Figure 6.9 shows the grand average CAEP to /da/ at each SNR level in the delta and theta band. CAEP to /da/ showed decrease in amplitude with decreasing SNR. The trend in decreasing in dominant peaks amplitude is more consistent in the theta band than in the delta band.

6.3.9 Correlation between behavioural intelligibility scores and objective measures

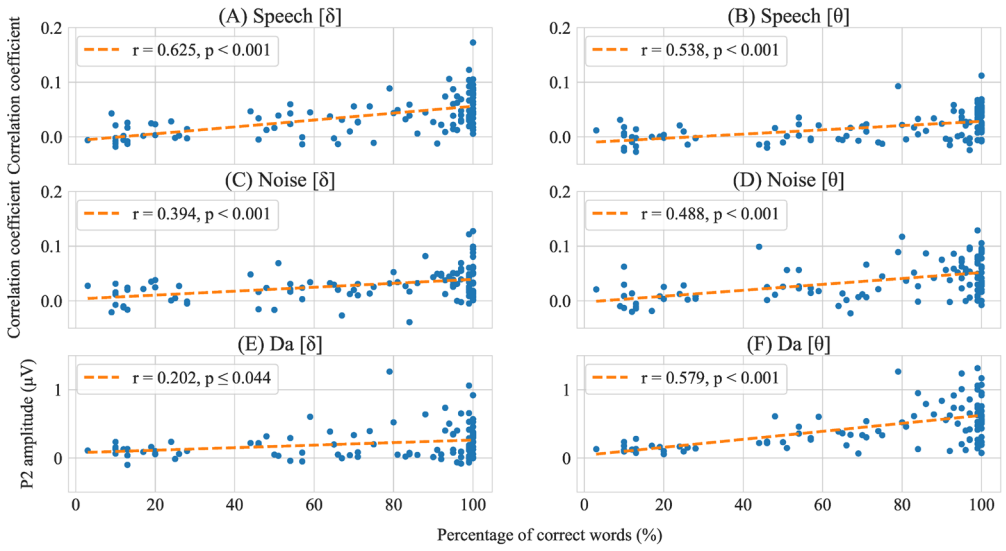

Figure 6.10. Spearman correlation between the percentage of words correct from the behavioural Matrix test and the objective measure of cortical responses to each type of stimuli in the delta and theta band. Blue points represent the paired objective measure and the percentage of words correct from each individual at each SNR. The objective measures show statistically significant correlation with the percentage word correct ($p \leq 0.001$, Wilcoxon sign ranked test), except for the /da/ condition in the delta band.

Figure 6.10 shows the correlation between the percentage word correct from the behavioural Matrix test and the objective measure of cortical responses to each type of stimuli in the delta and theta band. Generally, the objective measures show statistically significant correlation with the percentage word correct ($p < 0.001$), except for the /da/ condition in the delta band. Across the group there is a trend for the objective response parameters to increase as intelligibility increases. Given that the behavioural SRT only range from -12.3 to -9.3 dB, as expected, the correlation between the SRT obtain from the behavioural Matrix test and the CT from the objective measures of cortical responses in the delta and theta band were not statistically significant.

6.4 Discussions

This study investigated the use of cortical responses to continuous speech, modulated noise, and repeating /da/ stimulus to predict the SRT in normal hearing participants without paying attention to the stimuli. Overall, all objective measure showed statistically significant correlation with the intelligibility of speech (word recognition scores). Both objective measures in the speech and modulated noise condition can provide an estimate of

SRT on the group level even in the absence of attention. However, the prediction of SRT may be challenging in individuals. The method of fitting a sigmoid function to the individual objective data does not work very well and did not consistently provide acceptable predicted SRT in all individuals, so it is unlikely to be sufficient for clinical application.

Although, the BACKWARD-CORR showed significant correlation with the percentage word correct on the group level, errors in objective SRT predictions compared to behavioural SRTs was generally high. For some subjects, it was hard to fit sigmoid curves to objective data in order to predict SRTs, probably due to the variability of measurements as a result of noise. In some subjects, the BACKWARD-CORRs were not significant at many SNR levels, thus no monotonic relationship between the correlation and the SNR is present for the sigmoid function to fit. This is shown in Figure **6.4** and Figure **6.5**, where many of the BACKWARD-CORR were not statistically significant compared to bootstrapped BACKWARD-CORR. The BACKWARD-CORR was also considerably variable within SNR level on the group level, leading to low precision of SRT prediction across subjects. The best performing objective methods to predict the behavioural SRTs of individuals were modulated noise and /da/ responses in the theta band. However, even in the best stimulus conditions, errors in SRT predictions of greater than ±7 dB (defined as CT not applicable in the current study) were seen in 30% of the participants. This suggest that the measurement is not reliable for use in individuals.

A study by Lesenfants *et al.* (2019) reported only 37% and 11% of the participant that showed SRT prediction error greater than ±3.5 dB (defined as CT not applicable in their study) for cortical entrainment of speech envelope in the delta and theta band respectively. Note that the number portion of participants with CT not applicable in Lesenfants *et al.* (2019) is lower within a narrower range than in the current study. The cause of this difference might be due to the difference in the backward TRF approach. Lesenfants *et al.* (2019) applied a generic backward TRF model (model averaged from all other participant's model) to each participant with the purpose to obtain the BACKWARD-CORR using minimum amount of EEG data, whereas the current study applied subject-specific backward TRF. The BACKWARD-CORR obtained through the generic model might be less variable than the ones from the subject-specific model (Jessen *et al.*, 2019). However, in some cases, the generic model may underperform the subject-specific model if the cortical responses across the cohort does not meet the homogeneity assumption (Di Liberto and Lalor, 2017). The selection of optimal EEG electrodes which generate

significantly high BACKWARD-CORR were also made in Lesenfants *et al.* (2019). The removal of EEG electrodes contributing less to the stimulus reconstruction approach can significantly increase the BACKWARD-CORR (Montoya-Martinez *et al.*, 2021). Another factor which may cause the SRT prediction in this study to be worse than other studies is the participant's attention to the stimulus. Vanthornhout, Decruy and Francart (2019) showed that SRT can be predicted using cortical responses to sound in both conditions with and without attention to the stimulus. However, the strength of cortical envelope entrainment is generally stronger in attention condition, thus SRT prediction in the current study might suffer from the low SNR in the EEG due to the lack of attention to the stimulus.

In the condition using continuous speech, it is possible that participants constantly switching their attention between watching the documentary and listening to the story. Although, the instruction was given for the participants to not pay attention to the story, it cannot be guaranteed that they followed the instruction throughout the whole session. Possibly, this might be one reason that the TRF-models from the continuous speech condition showed significantly lower amplitude compared to the TRF-models from the modulated noise condition (see Figure **6.8** B and D). Another possibility is the order of conditions in which the participants undergo. Generally assumed that participants would mostly be alert and capable of maintaining attention to the documentary without being distract by the stimulus in the first condition they undergo. The TRF-model from the first stimulus condition is likely to have higher SNR than the cortical responses from following stimulus conditions. However, the order of condition for each participant were pre-allocated (not randomly allocated), so that the number of participants starting the experiment with the speech condition will be similar to number of participants starting with different condition. Therefore, the bias towards obtaining better cortical responses from any condition should be negligible.

The experimental design and task involved could possibly be a factor affecting the SRT prediction accuracy. In the current study, the behavioural Matrix test measures how well the listener can discriminate/identify the sound in background noise not just verifying detection of sound, whereas the EEG experiment did not include such a task. The two experiments clearly involve speech perception at different processing stage. This might be a limitation for using measurements from a non-attending experiment to predict measurement from a more complex experiment as the attention and language skill were not taken into account. However, when considering from a clinical diagnosis point of view,

measurement with less or no confounding might be desirable (Cacace and McFarland, 2013). For example, when attention and encoding of sound by the auditory system both correlates with the cortical envelope entrainment, it may be challenging to identify whether difficulty in speech comprehension is from auditory processing in the brain or the auditory pathway. This study showed that experiment involving non meaningful stimuli might be better for the investigation of deficits in the auditory pathway because experiments involving speech, whether meaningful or not, are more susceptible to the effects of attention. Attention is indeed an important cognitive function for listening to speech-in-noise, but the measurement of attention is beneficial for diagnosis when the type of attention deficit is clearly categorised (Jafari, Malayeri and Rostami, 2015; Stavrinos *et al.*, 2018). The cortical envelope entrainment, however, suffers from the confounding auditory sensory and several cognitive factors (Nejime and Moore, 1998; Zou *et al.*, 2019; Reetzke, Gnanateja and Chandrasekaran, 2021).

For a clinical application, /da/ responses might have most application as they are faster to measure than responses to continuous speech or modulated noise and give similar performance in terms of prediction of SRT to modulated noise but considerably better prediction than using continuous speech. However, even for /da/ responses, the errors seen in the prediction of SRT for individual subjects are probably too large for clinical application (with the approaches used in the current study).

6.5 Conclusion

This study showed that it is possible to predict SRT using cortical responses to continuous speech, modulated noise and /da/ when participants are not required to pay attention to the stimuli. Particularly in the theta band, modulated noise and /da/ responses provided more accurate SRT prediction than using cortical responses to continuous speech. The method used in this study, however, does not seem to be applicable in clinics due to the inconsistent prediction and large prediction error in individuals. Given that the SRT can be predicted from cortical responses to modulated noise, the cortical envelope entrainment may not be a measurement which specifically reflects speech intelligibility in the human brain. For normal hearing people, the SRT appears to be mostly related to measurements of audibility or hearing threshold (PTA, CAEP, etc.).

Chapter 7 **Discussion and conclusion**

The aim of this thesis was to assess the applicability of using cortical responses to speech-based stimuli as an objective measure to predict individual's behavioural speech intelligibility in normal hearing adults, potentially for clinical use. For this, a number of questions regarding the detection of response and functional role of the cortical envelope entrainment need to be addressed. This thesis addressed some questions, which are:

1. Concerning the issue of low SNR in AERs introduced in section 3.1, can the detection of cortical responses to natural speech be improved by adding pauses between words in the speech stream without affecting the meaning of speech? (In chapter 4)

2. Between the forward and backward TRF, which approach is more sensitive for detecting cortical responses to natural speech and how effective they are in terms of response detection compared to detection of CAEP to /da/? (In chapter 5)

3. Do cortical responses to natural speech provide more accurate prediction of SRT than cortical responses to modulated noise or /da/? (In chapter 6)

As a summary, the findings are that the detection of cortical responses to natural speech was improved by inserting pauses to the speech stream and response detection was more sensitive trough the backward TRF approach, however, CAEP to /da/ appears to be easier and faster (less test time) to detect than cortical responses to natural speech with and without additional pauses. Although the best choices for detecting cortical responses to speech-based stimulus were utilised, cortical responses to natural speech as well as to modulated noise and /da/ provides high SRT prediction error in individuals. These findings suggest that the method of using cortical responses to sound to predict SRT in individuals is unlikely to be applicable for clinical use due to low SNR and variability in responses measurement that may cause high prediction error. More importantly, the cortical envelope entrainment may not specifically represent speech intelligibility in normal hearing adults, as it can be driven by both stimulus properties and attention to the target sound.

Discussion made in chapter 4, 5, and 6, are more related to each specific study. This chapter will discuss the findings from the three studies altogether, mainly concerning the detection of cortical responses to speech using different type of stimulus and the relationship between cortical responses and speech intelligibility. This chapter then concludes with some suggestions for future study.

7.1 Effectiveness of speech and non-speech stimulus in generating cortical responses

Two studies were carried out to improve the detection of cortical responses to continuous speech by 1.) adding pauses between words in the continuous speech stimulus (chapter 4) and 2.) compare the sensitivity of response detection between the forward and backward TRF (chapter 5). These studies were aimed to explore the best way to measure the cortical envelope entrainment using the TRF approach, so that the speech SRT could be predicted more efficiently. In chapter 4, additional pauses inserted to the continuous speech stimulus generates stronger onset responses in the EEG. This stimulus modification led to a greater number of responses detected. This finding also raised a question of whether the cortical envelope entrainment is reflecting the acoustic processing or comprehension of speech. In chapter 5, the number of detected responses to continuous speech were greater via the backward TRF compared to the forward TRF. The number of detected cortical responses to continuous speech is similar to CAEP to /da/ but requires longer measurement time. If the goal is to detect access to speech-like sounds and not examining speech comprehension, CAEP to /da/ might be a better option, since the response is greater in SNR and require shorter measurement time.

As it is shown throughout in this thesis, cortical responses to natural speech are generally weaker and lower in number of detected responses than cortical responses to all other type of stimuli that has been used throughout this thesis (continuous speech with additional pauses inserted between words, broadband noise modulated by the natural speech intensity envelope, repeating monosyllable /da/). This shows that natural speech was not the best stimulus to use as a tool for assessing whether a person has access to speech-like sounds. Findings in chapter 5 suggested that this may be due to the difference in response characteristics rather than the choice of response detection method. Cortical responses to natural speech contain less onset responses, only where pauses occur between phrases or sentences, compared to cortical responses to continuous speech with additional pauses and repeating /da/. The forward TRF-model of cortical responses to /da/ showed response morphology very similar to the CA waveform. The number of detected /da/ responses via bootstrapping to find statistically significant peaks is also similar between the forward TRF-model and CA. Cortical responses to onsets may exhibit stronger linear properties than non-onset responses and better modelled by the TRF approach, thus number of response detected were greater when using continuous speech with pauses and /da/ than using natural speech.

7.2 Relation between cortical envelope entrainment and behavioural speech intelligibility

In chapter 2 and 3, a challenge of which stage in the process of speech comprehension (1. detection, 2. discrimination, 3. identification, and 4. comprehension) is being measured or quantify through behavioural test or cortical responses to sound has been highlighted. With the choice of stimuli and the non-attention condition, the studies in this thesis may only involve the detection stage due to the following reasons. By using /da/ and modulated noise stimulus, it is clear that the stage of comprehension cannot be examined, as the stimuli are not meaningful. For the use of natural speech stimulus, although it is comprehensible, the participants were only passively listening to the stimulus. It remains as a challenge to determine whether the stage of discrimination between different sounds can be measured through the cortical envelope entrainment.

The relation between cortical envelope entrainment and speech intelligibility has been extensively studied. Vanthornhout *et al.* (2018) was the first to establish a framework of using cortical envelope entrainment to predict the behavioural SRT. Lesenfants *et al.* (2019) then expanded this framework further by using cortical responses to stimulus envelope and other features in the natural speech stimulus such as spectrogram, phonemes, phonetic features and combination of spectrogram and phonetic features to predict SRT. Vanthornhout, Decruy and Francart (2019) showed that SRT can be predicted from the cortical envelope entrainment either with or without the participant's attention to the stimulus. In their study, the SRT prediction error at group level from both attention and non-attention conditions were within ±2 dB, which is similar to the results in chapter 6. However, they did not show how well the SRT prediction is in individuals. The study in chapter 6 built on these previous studies aimed to investigate whether the cortical envelope entrainment specifically reflects individual's speech intelligibility or merely a measurement affect by the stimulus SNR. This thesis showed that the cortical envelope entrainment may not specifically represent human speech comprehension and prone to confounding stimulus property and attention. This was achieved by using cortical responses to three stimuli ranging from a simpler repeating monosyllable to a more complex modulated noise and finally natural speech to be used as an objective measure for predicting behavioural speech-in-noise performance without the need of participant's attention to the stimuli.

The functional role of cortical envelope entrainment is often suggested to be a combination between auditory processing of acoustic cues and cognitive abilities to combine the

acoustic information to comprehend speech (Ding and Simon, 2014). The interaction between these two processes may cause complexity in interpreting the relationship between cortical envelope entrainment and speech intelligibility. The two measures are expected to be positively correlated if it is assumed that they are directly related to each other. However, studies by Song and Iverson (2018); Zou *et al.* (2019); Reetzke, Gnanateja and Chandrasekaran (2021) conducted a study involving native and non-native English speakers to investigate how the participant's attention towards a story narrated in English affects the cortical envelope entrainment. The native English speaking group had significantly higher scores in English proficiency test. Their study showed that non-native English speaker exhibit stronger cortical envelope entrainment than native English speakers. These studies demonstrated that attention may have more influence on the cortical envelope entrainment than the auditory processing. The study in chapter 6 minimised the involvement of attention by instructing participants to watch a documentary with closed captions and ignore the stimulus, and that the measured cortical envelope entrainment may mostly be driven by the acoustic processing. The results point out the importance of minimising interaction between the two processes when drawing conclusion from the EEG and behavioural measurement to prevent possible misconceptions. The cortical envelope entrainment may still be useful for the assessment of hearing sensitivity similar to PTA, if the cognitive abilities do not confound. Further research building upon the framework of using cortical responses to natural speech to predict SRT should include analysis on other stimulus feature, for example phonemes and phonetic features (Lesenfants *et al.*, 2019), in addition to the stimulus envelope to assess on the auditory processing specific to speech not simply any sound.

The cortical envelope entrainment in the delta band was also suggested to be closely related to speech comprehension (Etard *et al.*, 2019) and that SRT was predicted more accurately than in the theta band (Vanthornhout *et al.*, 2018). The cortical envelope entrainment in the theta band was suggested to be more related to clarity of the speech material, can be defined as a level (in percentage) of speech sound in background noise that a native speaker can understand (Etard *et al.*, 2019), than in the delta band. In this thesis, the cortical responses to speech and non-speech stimulus in the delta frequency band generate stronger response than in the theta band, as indicated by the correlation coefficient between the actual and the estimated stimulus envelope. This agrees with previous studies reporting stronger responses to continuous speech in the delta band including better SRT prediction than in the theta band (Vanthornhout *et al.*, 2018;

Lesenfants *et al.*, 2019). However, the results from individuals do not fully support suggestions made by other studies because the cortical responses to both natural speech and modulated noise in the delta band can be used to predict SRT. This may be due to the exclusion of attention task in chapter 6, so the cortical envelope entrainment only reflects the acoustic processing for responses to both natural speech and modulated noise. Results in chapter 6 also do not fully support the suggestion of cortical envelope entrainment in the theta band being related to clarity of perceived speech. In the theta band, both cortical envelope entrainment in both natural speech and modulated noise condition showed a trend of decreasing entrainment strength with lower stimulus SNR. So, it is clearly not specifically an indication of clarity of perceived speech. Overall, this thesis would suggest that, when cognitive abilities are not greatly involved, the cortical envelope entrainment might mostly be related to the SNR of stimulus (stimulus intensity against stationary speech-shaped noise) in both the delta and theta band.

A further way in which the relationship between speech processing and the EEG could be explored is to consider the spatial location in the brain, where the EEG signal may be originating. This can be achieved most easily through invasive measurements. For non-invasive EEG, one of the alternative methods is through the source mapping cortical responses to determine whether speech is processed distinctively from other sounds in certain brain regions, which was not done in this thesis. Studies using invasive EEG reported that the superior temporal gyrus and the posterior inferior frontal gyrus (Broca's regions) were found to be more responses to natural speech (Kubanek *et al.*, 2013; Hamilton, Edwards and Chang, 2018). Speech might indeed be processed differently from other sounds, but the encoded speech features might not be reliably extract from non-invasive EEG due to low SNR compared to invasive EEG. This type of study is very limited as it is normally done in people who have epilepsy. If further study is possible, it may be interesting to see whether auditory system of native and non-native speakers encode certain language differently or not. Then it may be further explored whether processing of speech sound will be different at the auditory or cognitive processing stage for the two group of subjects.

One factor that may be considered as a challenge when examining the correlation between cortical envelope entrainment and speech intelligibility in normal hearing participants only is the narrow range of behavioural SRT. This was a problem found in chapter 6, where there is no significant correlation between the behavioural SRT, less variable (within 3 dB range), and predicted SRT using cortical responses, highly variable (range greater than 10

dB). This contradicts with the results from Vanthornhout *et al.* (2018). Note that the range of behavioural SRT in chapter 6 (-12.3 to -9.3 dB) is narrower and the mean SRT (-10.8 dB) is lower than other similar studies, -9.9 to -4.7 dB (mean -7.4 dB) in Vanthornhout *et al.* (2018); Lesenfants *et al.* (2019) and -10.3 to -7.7 dB (mean -8.5 dB) in Lesenfants *et al.* (2019). This may be due to the difference in language and testing SNR levels in the behavioural matrix test. The variability in the cortical envelope entrainment both between and within subjects can also give rise to the problem. In other group of subjects, such as cochlea implant users, the range of behavioural SRT may be greater than normal hearing people (Abdel-Latif and Meister, 2021), this may also lead to a stronger and significant correlation between the actual and predicted SRT.

7.3 The applicability of cortical responses to continuous and repeating short stimulus measurement in clinics

Although previous studies have suggested that the cortical envelope entrainment would be useful as an objective measure to predict SRT, it appears that this need to be taken with caution due to the complexity of its relationship to individual's speech intelligibility. More recent research including this thesis suggest that much clearer understanding is needed for the auditory functional role that the cortical envelope entrainment is representing. This thesis also shows that this type of response through the TRF approach also seems to be less sensitive in detecting cortical responses to speech-like sound than conventional CAEP. The detection of cortical responses to continuous speech could be improve using some suggested models such as canonical correlation analysis based approaches (de Cheveigne *et al.*, 2018; Dmochowski *et al.*, 2018) for more efficient response detection. These approaches might improve the correlation between the EEG and the stimulus feature because an optimal linear transformation is applied on both signals, while the TRF approach applies to the input signal only. However, one limitation of the canonical corelation analysis is that the output correlation coefficients from the model will always be positive (Zhuang, Yang and Cordes, 2020). While this may not be problematic for detection of responses, it may be more difficult to determine the direction of relationship between the cortical response and the stimulus, thus the neurophysiological function might be misinterpreted.

CAEP to /da/ has shown to be a more sensitive method for detecting cortical responses to speech-like sounds compared to natural speech and continuous speech with additional pauses inserted between words. In chapter 6, it was also the best option as an objective

measure to predict SRT with significantly greater number of participants which SRT can be predicted within 7 dB difference than other conditions. The strong correlation between the CAEP peaks amplitude and speech intelligibility was reported by Billings *et al.* (2013), however, to the best of the authors knowledge, there are no study that apply the method of fitting a psychometric function to the CAEP peaks amplitude to predict SRT. Ultimately, this may be due to the stronger or more onset responses in the CAEP. As it is shown in chapter 5 and 6, responses to /da/ have relatively high amplitude and hence SNR than cortical envelope entrainment. From a clinical point of view, in terms of prediction of SRT, CAEP may be preferable over the measurement of cortical entrainment to stimulus envelope. This is due to the shorter test time, higher number of detected responses, and importantly the better-established way of interpreting the peaks amplitude and latency in the coherent average waveform than for cortical responses to continuous speech. However, due to its unnatural stimulation method, it is probably not able to reflect comprehension of the speech stimulus, which is an area for further study.

7.4 Suggestions for future study

7.4.1 Cortical responses to continuous speech with pauses added between words

The study in chapter 4 has demonstrated that the strength of cortical envelope entrainment can be affected by the amount and duration of pauses inserted to the continuous speech stream. A remaining question from the study is how does the stronger onset responses in cortical responses to speech with additional pauses relate to speech intelligibility. This would include behavioural speech intelligibility test, which was not done previously. The aim here is to investigate whether the more pause in continuous speech improves only the detection of responses or these pauses may contribute to stronger encoding of speech sound in the brain, though it might not improve intelligibility of perceived speech.

7.4.2 Relation between cortical responses to sounds and speech intelligibility

For the research are on relating cortical responses to sound and speech intelligibility, suggestions for future study can be divided into two lines of research. The first line of research is to improve the framework of using cortical responses as objective measures to predict SRT. The second line of research concerns the bigger picture of stages in speech comprehension introduced in chapter 2.

For the first line of research, the ultimate goal of the framework is to provide an accurate SRT prediction for each individual. However, this thesis showed that this framework does not work well individuals as the prediction error is considerably high, can be greater than 7 dB difference from the actual SRT. One clear problem is that the measurement of cortical responses may not always be detected, especially for responses to continuous speech even in quiet condition. This may be addressed by including more EEG pre-processing and analysis methods for better detection of responses and more successful prediction of SRT. It would be important to manage the sessions of experiment well so that the participant will not get too tired because of long experiments, which can hugely affect the quality of the EEG. To make this framework more applicable for clinical use, further work is needed to assess the repeatability of the test. As it is shown across this thesis that cortical responses to continuous speech is considerably variable between participants, it is not yet clear whether it is also variable within the same participants or not.

For the second line of research, future study may focus on extracting cortical responses to speech based on other stimulus features other than the stimulus envelope. As stated, that the measurement of cortical responses to speech and non-speech sounds in this thesis may mostly be a confirmation of access to sound (detection) but not reflecting further processes (discrimination, identification, and discrimination) in speech comprehension. This is true for non-speech sounds but uncertain for speech sounds. A study by Di Liberto, O'Sullivan and Lalor (2015) has shown that cortical responses to other stimulus feature more specific to speech, such as phonemes and phonetic features, can be decoded from the EEG. Cortical responses to these features also showed stronger correlation with speech intelligibility than responses to speech envelope. However, it is unclear if relating cortical responses to these more speech specific features reflects specific speech encoding or cognitive processing. It is also questionable whether what aspect in speech processing is being concluded from the framework of using cortical responses to sound to predict SRT. For normal hearing people, speech or lexical processing might not be measured from the behavioural test, due to the simplicity of words used in the test, it may be dependent mostly on SNR of sound.

Since this thesis has shown that acoustic and cognitive processing can confound, it might worth consider towards using more objective measure to disentangle the measurements of encoding of acoustic information and cognitive processing related to speech intelligibility. One possible method is to decode speech specific features from cortical responses to speech in languages that the listener can and cannot understand. The aim of this is to determine whether the measurements of successful encoding of speech sounds in the brain

alone can be used to indicate speech comprehension or it is again a confirmation of access to sound in general. If successful, a measurement to confirm the encoding of speech sound in the brain without confounding cognitive processing might be useful for diagnosing the cause of language impairments in children, where it is not always or entirely related to hearing problem alone (Tuller and Delage, 2014).

Appendix A

A complete analysis of testing each of the three decoders (trained on full envelope, onsets, and non-onsets) on all speech features, nine pairs in total for each speech pause condition.

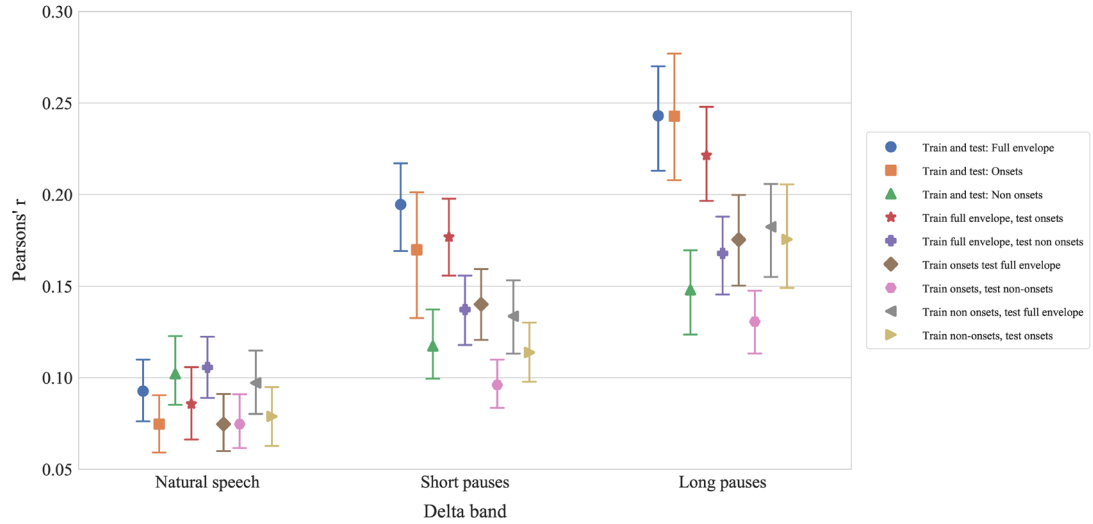

Figure A1. The correlation coefficient of each decoder training and testing combination in the delta band across three speech pause conditions. Each point indicates the average Pearsons' r across sixteen participants. Error bars indicate the 95% confidence interval for the mean.

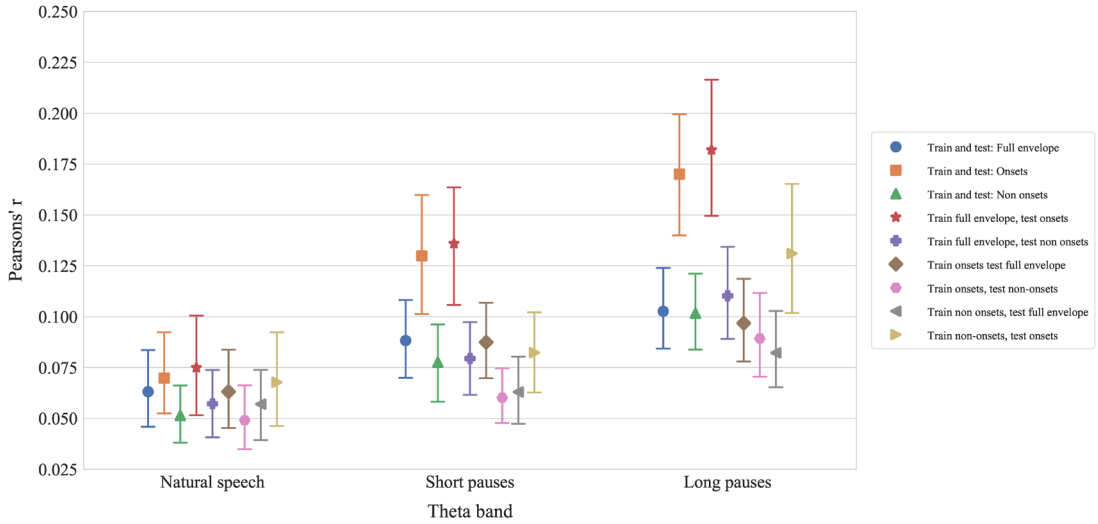

Figure A2. The correlation coefficient of each decoder training and testing combination in the theta band across three speech pause conditions. Each point indicates the average Pearsons' r across sixteen participants. Error bars indicate the 95% confidence interval for the mean.

Bibliography

Abdel-Latif, K.H.A. and Meister, H. (2021) 'Speech Recognition and Listening Effort in Cochlear Implant Recipients and Normal-Hearing Listeners', *Front Neurosci*, 15, p. 725412.

Accou, B. *et al.* (2021) 'Predicting speech intelligibility from EEG in a non-linear classification paradigm *', *Journal of Neural Engineering*, 18(6).

Accou, B. *et al.* (2023) 'Decoding of the speech envelope from EEG using the VLAAI deep neural network', *Sci Rep*, 13(1), p. 812.

Ahissar, E. *et al.* (2001) 'Speech comprehension is correlated with temporal response patterns recorded from auditory cortex', *Proc Natl Acad Sci U S A*, 98(23), pp. 13367-72.

Aiken, S.J. and Picton, T.W. (2008a) 'Envelope and spectral frequency-following responses to vowel sounds', *Hearing Research*, 245(1-2), pp. 35-47.

Aiken, S.J. and Picton, T.W. (2008b) 'Human cortical responses to the speech envelope', *Ear Hear*, 29(2), pp. 139-57.

Akeroyd, M.A. (2008) 'Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults', *Int J Audiol*, 47 Suppl 2, pp. S53-71.

Akhoun, I. *et al.* (2008) 'The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults', *Clin Neurophysiol*, 119(4), pp. 922-33.

Alemei, R. and Lehmann, A. (2019) 'Middle Latency Responses to Optimized Chirps in Adult Cochlear Implant Users', *Journal of the American Academy of Audiology*, 30(5), pp. 396-405.

Aljarboa, G.S., Bell, S.L. and Simpson, D.M. (2022) 'Detecting cortical responses to continuous running speech using EEG data from only one channel', *International Journal of Audiology*.

Anderson, S. *et al.* (2010a) 'Cortical-evoked potentials reflect speech-in-noise perception in children', *Eur J Neurosci*, 32(8), pp. 1407-13.

Anderson, S. and Kraus, N. (2010) 'Objective neural indices of speech-in-noise perception', *Trends Amplif*, 14(2), pp. 73-83.

Anderson, S. *et al.* (2010b) 'Brainstem correlates of speech-in-noise perception in children', *Hear Res*, 270(1-2), pp. 151-7.

Armstrong, R.A. (2014) 'When to use the Bonferroni correction', *Ophthalmic and Physiological Optics*, 34(5), pp. 502-508.

Auerbach, B.D., Rodrigues, P.V. and Salvi, R.J. (2014) 'Central gain control in tinnitus and hyperacusis', *Frontiers in Neurology*, 5.

Avan, P., Giraudet, F. and Buki, B. (2015) 'Importance of binaural hearing', *Audiol Neurootol*, 20 Suppl 1, pp. 3-6.

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(3), pp. 549-63.

Bibliography

Bamiou, D.E., Musiek, F.E. and Luxon, L.M. (2001) 'Aetiology and clinical presentations of auditory processing disorders--a review', *Arch Dis Child*, 85(5), pp. 361-5.

Beattie, R.C. (1988) 'Interaction of click polarity, stimulus level, and repetition rate on the auditory brainstem response', *Scand Audiol*, 17(2), pp. 99-109.

Bell, S.L. *et al.* (2004) 'Recording the middle latency response of the auditory evoked potential as a measure of depth of anaesthesia. A technical note', *British Journal of Anaesthesia*, 92(3), pp. 442-445.

Bench, J., Kowal, A. and Bamford, J. (1979) 'The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children', *Br J Audiol*, 13(3), pp. 108-12.

Bess, F.H., Dodd-Murphy, J. and Parker, R.A. (1998) 'Children with minimal sensorineural hearing loss: prevalence, educational performance, and functional status', *Ear Hear*, 19(5), pp. 339-54.

Bidelman, G.M. (2015) 'Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR', *Hear Res*, 323, pp. 68-80.

Bidelman, G.M. and Bhagat, S.P. (2016) 'Objective detection of auditory steady-state evoked potentials based on mutual information', *International Journal of Audiology*, 55(5), pp. 313-319.

Bieser, A. and Muller-Preuss, P. (1996) 'Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds', *Exp Brain Res*, 108(2), pp. 273-84.

Biesmans, W. *et al.* (2017) 'Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario', *IEEE Trans Neural Syst Rehabil Eng*, 25(5), pp. 402-412.

Billings, C.J. *et al.* (2011) 'Cortical Encoding of Signals in Noise: Effects of Stimulus Type and Recording Paradigm', *Ear and Hearing*, 32(1), pp. 53-60.

Billings, C.J. *et al.* (2013) 'Predicting perception in noise using cortical auditory evoked potentials', *J Assoc Res Otolaryngol*, 14(6), pp. 891-903.

Billings, C.J. *et al.* (2015) 'Electrophysiology and Perception of Speech in Noise in Older Listeners: Effects of Hearing Impairment and Age', *Ear Hear*, 36(6), pp. 710-22.

Bishop, D.V. and Snowling, M.J. (2004) 'Developmental dyslexia and specific language impairment: same or different?', *Psychol Bull*, 130(6), pp. 858-86.

Boersma, P. and Weenink, D. (2001) 'PRAAT, a system for doing phonetics by computer', *Glot international*, 5, pp. 341-345.

Bond, Z.S. and Moore, T.J. (1994) 'A Note on the Acoustic-Phonetic Characteristics of Inadvertently Clear Speech', *Speech Communication*, 14(4), pp. 325-337.

Bradlow, A.R., Torretta, G.M. and Pisoni, D.B. (1996) 'Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics', *Speech Commun*, 20(3), pp. 255-272.

Bramhall, N. *et al.* (2015) 'Speech Perception Ability in Noise is Correlated with Auditory Brainstem Response Wave I Amplitude', *J Am Acad Audiol*, 26(5), pp. 509-517.

Brigell, M. *et al.* (2003) 'Guidelines for calibration of stimulus and recording parameters used in clinical electrophysiology of vision', *Documenta Ophthalmologica*, 107(2), pp. 185-193.

Brodbeck, C., Presacco, A. and Simon, J.Z. (2018) 'Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension', *Neuroimage*, 172, pp. 162-174.

Byrne, D. *et al.* (1994) 'An International Comparison of Long-Term Average Speech Spectra', *Journal of the Acoustical Society of America*, 96(4), pp. 2108-2120.

Cacace, A.T. and McFarland, D.J. (2013) 'Factors Influencing Tests of Auditory Processing: A Perspective on Current Issues and Relevant Concerns', *Journal of the American Academy of Audiology*, 24(7), pp. 572-589.

Ceponiene, R., Cheour, M. and Naatanen, R. (1998) 'Interstimulus interval and auditory event-related potentials in children: evidence for multiple generators', *Electroencephalogr Clin Neurophysiol*, 108(4), pp. 345-54.

Ceponiene, R., Rinne, T. and Naatanen, R. (2002) 'Maturation of cortical sound processing as indexed by event-related potentials', *Clin Neurophysiol*, 113(6), pp. 870-82.

Cerella, J. (1990) 'Aging and Information-Processing Rate', in Birren, J.E. and Schaie, K.W. (eds.) *Handbook of the Psychology of Aging*. Academic Press, pp. 201-221.

Chait, M. *et al.* (2015) 'Multi-time resolution analysis of speech: evidence from psychophysics', *Front Neurosci*, 9, p. 214.

Cherry, E.C. (1953) 'Some Experiments on the Recognition of Speech, with One and with Two Ears', *The Journal of the Acoustical Society of America*, 25(5), pp. 975-979.

Chesnaye, M.A. *et al.* (2018) 'Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time', *Int J Audiol*, 57(6), pp. 468-478.

Clopper, C.G., Pisoni, D.B. and Tierney, A.T. (2006) 'Effects of open-set and closed-set task demands on spoken word recognition', *J Am Acad Audiol*, 17(5), pp. 331-49.

Comon, P. (1994) 'Independent Component Analysis, a New Concept', *Signal Processing*, 36(3), pp. 287-314.

Crosse, M.J. *et al.* (2016) 'The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli', *Front Hum Neurosci*, 10, p. 604.

Culling, J.F. (2016) 'Speech intelligibility in virtual restaurants', *J Acoust Soc Am*, 140(4), p. 2418.

Cunningham, J. *et al.* (2001) 'Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement', *Clin Neurophysiol*, 112(5), pp. 758-67.

Curry, E.T. (1949) 'A study of the relationship between speech thresholds and audiometric results in perception deafness', *J Speech Disord*, 14(2), pp. 104-10.

Curtin, S. *et al.* (2017) 'Speech Perception: Development ☆', in *Reference Module in Neuroscience and Biobehavioral Psychology*.

Das, N., Bertrand, A. and Francart, T. (2018) 'EEG-based auditory attention detection: boundary conditions for background noise and speaker positions', *J Neural Eng*, 15(6), p. 066017.

Davis, H. *et al.* (1966) 'The slow response of the human cortex to auditory stimuli: recovery process', *Electroencephalogr Clin Neurophysiol*, 21(2), pp. 105-13.

de Cheveigne, A. and Nelken, I. (2019) 'Filters: When, Why, and How (Not) to Use Them', *Neuron*, 102(2), pp. 280-293.

de Cheveigne, A. *et al.* (2018) 'Decoding the auditory brain with canonical component analysis', *Neuroimage*, 172, pp. 206-216.

de Melo, A. *et al.* (2016) 'Cortical auditory evoked potentials in full-term and preterm neonates', *Codas*, 28(5).

de Taillez, T., Kollmeier, B. and Meyer, B.T. (2020) 'Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech', *European Journal of Neuroscience*, 51(5), pp. 1234-1241.

Desjardins, J.L. and Doherty, K.A. (2013) 'Age-related changes in listening effort for various types of masker noises', *Ear Hear*, 34(3), pp. 261-72.

Dettori, J. (2010) 'The random allocation process: two things you need to know', *Evid Based Spine Care J*, 1(3), pp. 7-9.

Di Liberto, G.M. and Lalor, E.C. (2017) 'Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research', *Hearing Research*, 348, pp. 70-77.

Di Liberto, G.M., O'Sullivan, J.A. and Lalor, E.C. (2015) 'Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing', *Curr Biol*, 25(19), pp. 2457-65.

Dimitrijevic, A., John, M.S. and Picton, T.W. (2004) 'Auditory steady-state responses and word recognition scores in normal-hearing and hearing-impaired adults', *Ear Hear*, 25(1), pp. 68-84.

Ding, N. and Simon, J.Z. (2012a) 'Emergence of neural encoding of auditory objects while listening to competing speakers', *Proc Natl Acad Sci U S A*, 109(29), pp. 11854-9.

Ding, N. and Simon, J.Z. (2012b) 'Neural coding of continuous speech in auditory cortex during monaural and dichotic listening', *J Neurophysiol*, 107(1), pp. 78-89.

Ding, N. and Simon, J.Z. (2013) 'Adaptive temporal encoding leads to a background-insensitive cortical representation of speech', *J Neurosci*, 33(13), pp. 5728-35.

Ding, N. and Simon, J.Z. (2014) 'Cortical entrainment to continuous speech: functional roles and interpretations', *Front Hum Neurosci*, 8, p. 311.

Dingemanse, J.G. and Goedegebure, A. (2019) 'The Important Role of Contextual Information in Speech Perception in Cochlear Implant Users and Its Consequences in Speech Tests', *Trends Hear*, 23, p. 2331216519838672.

Dmochowski, J.P. *et al.* (2018) 'Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity', *Neuroimage*, 180, pp. 134-146.

Doelling, K.B. *et al.* (2014) 'Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing', *Neuroimage*, 85 Pt 2, pp. 761-8.

Drennan, D.P. and Lalor, E.C. (2019) 'Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity', *eNeuro*, 6(3).

Dryden, A. *et al.* (2017) 'The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis', *Trends in Hearing*, 21.

Du, Y. *et al.* (2011) 'Auditory frequency-following response: a neurophysiological measure for studying the "cocktail-party problem"', *Neurosci Biobehav Rev*, 35(10), pp. 2046-57.

Edwards, J., Fox, R.A. and Rogers, C.L. (2002) 'Final consonant discrimination in children: effects of phonological disorder, vocabulary size, and articulatory accuracy', *J Speech Lang Hear Res*, 45(2), pp. 231-42.

Ellis, L. *et al.* (1996) 'Effects of gender on listeners' judgments of speech intelligibility', *Percept Mot Skills*, 83(3 Pt 1), pp. 771-5.

Epstein, M. and Florentine, M. (2012) 'Binaural loudness summation for speech presented via earphones and loudspeaker with and without visual cues', *J Acoust Soc Am*, 131(5), pp. 3981-8.

Erber, N.P. (1976) 'Use of Audio Tape-Cards in Auditory Training for Hearing Impaired Children', *Volta Review*, 78(5), pp. 209-218.

Etard, O. *et al.* (2019) 'Decoding of selective attention to continuous speech from the human auditory brainstem response', *Neuroimage*, 200, pp. 1-11.

Etard, O. and Reichenbach, T. (2019) 'Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise', *J Neurosci*, 39(29), pp. 5750-5759.

Ferguson, S.H. (2004) 'Talker differences in clear and conversational speech: vowel intelligibility for normal-hearing listeners', *J Acoust Soc Am*, 116(4 Pt 1), pp. 2365-73.

Ferman, L., Verschuure, J. and Van Zanten, B. (1993) 'Impaired speech perception in noise in patients with a normal audiogram', *Audiology*, 32(1), pp. 49-54.

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(4), pp. 1725-36.

Fletcher, H. (1950) 'A method of calculating hearing loss for speech from an audiogram', *Acta Otolaryngol Suppl*, 90, pp. 26-37.

Ghitza, O. and Greenberg, S. (2009) 'On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence', *Phonetica*, 66(1-2), pp. 113-26.

Giraud, A.L. and Poeppel, D. (2012) 'Cortical oscillations and speech processing: emerging computational principles and operations', *Nat Neurosci*, 15(4), pp. 511-7.

Golding, M. *et al.* (2009) 'The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection', *Int J Audiol*, 48(12), pp. 833-42.

Goossens, T. *et al.* (2018) 'Neural envelope encoding predicts speech perception performance for normal-hearing and hearing-impaired adults', *Hear Res*, 370, pp. 189-200.

Goossens, T. *et al.* (2019) 'The association between hearing impairment and neural envelope encoding at different ages', *Neurobiol Aging*, 74, pp. 202-212.

Gordon-Salant, S. (2005) 'Hearing loss and aging: new research findings and clinical implications', *J Rehabil Res Dev*, 42(4 Suppl 2), pp. 9-24.

Gramfort, A. *et al.* (2013) 'MEG and EEG data analysis with MNE-Python', *Front Neurosci*, 7, p. 267.

Greenberg, S. *et al.* (2003) 'Temporal properties of spontaneous speech - a syllable-centric perspective', *Journal of Phonetics*, 31(3-4), pp. 465-485.

Guest, H. *et al.* (2018) 'Impaired speech perception in noise with a normal audiogram: No evidence for cochlear synaptopathy and no relation to lifetime noise exposure', *Hearing Research*, 364, pp. 142-151.

Hagerman, B. (1982) 'Sentences for testing speech intelligibility in noise', *Scand Audiol*, 11(2), pp. 79-87.

Hagerman, B. (1984) 'Clinical measurements of speech reception threshold in noise', *Scand Audiol*, 13(1), pp. 57-63.

Hambrook, D.A., Soni, S. and Tata, M.S. (2018) 'The effects of periodic interruptions on cortical entrainment to speech', *Neuropsychologia*, 121, pp. 58-68.

Hamilton, L.S., Edwards, E. and Chang, E.F. (2018) 'A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus', *Curr Biol*, 28(12), pp. 1860-1871 e4.

Harris, J.D., Haines, H.L. and Myers, C.K. (1960) 'The importance of hearing at 3 kc for understanding speeded speech', *Laryngoscope*, 70, pp. 131-46.

Haufe, S. *et al.* (2014) 'On the interpretation of weight vectors of linear models in multivariate neuroimaging', *Neuroimage*, 87, pp. 96-110.

Heald, S. and Nusbaum, H. (2014) 'Speech perception as an active cognitive process', *Frontiers in Systems Neuroscience*, 8.

Heeringa, A.N. and van Dijk, P. (2019) 'Neural coding of the sound envelope is changed in the inferior colliculus immediately following acoustic trauma', *European Journal of Neuroscience*, 49(10), pp. 1220-1232.

Herdman, A.T. *et al.* (2002) 'Intracerebral sources of human auditory steady-state responses', *Brain Topogr*, 15(2), pp. 69-86.

Hertrich, I. *et al.* (2012) 'Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal', *Psychophysiology*, 49(3), pp. 322-34.

Hickok, G., Farahbod, H. and Saberi, K. (2015) 'The Rhythm of Perception: Entrainment to Acoustic Rhythms Induces Subsequent Perceptual Oscillation', *Psychol Sci*, 26(7), pp. 1006-13.

Hirsh, I.J. (1948) 'The Influence of Interaural Phase on Interaural Summation and Inhibition', *The Journal of the Acoustical Society of America*, 20(4), pp. 536-544.

Holdgraf, C.R. *et al.* (2017) 'Encoding and Decoding Models in Cognitive Electrophysiology', *Front Syst Neurosci*, 11, p. 61.

Howard, M.F. and Poeppel, D. (2010) 'Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension', *J Neurophysiol*, 104(5), pp. 2500-11.

Iotzov, I. and Parra, L.C. (2019) 'EEG can predict speech intelligibility', *J Neural Eng*, 16(3), p. 036008.

Jafari, Z., Malayeri, S. and Rostami, R. (2015) 'Subcortical encoding of speech cues in children with attention deficit hyperactivity disorder', *Clin Neurophysiol*, 126(2), pp. 325-32.

Janse, E. (2009) 'Processing of fast speech by elderly listeners', *J Acoust Soc Am*, 125(4), pp. 2361-73.

Jessen, S. *et al.* (2019) 'Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie', *Neuroimage*, 202, p. 116060.

Johnson, E.M. and Ferguson, S.H. (2016) 'Gender and rate effects on speech intelligibility', *The Journal of the Acoustical Society of America*, 139(4), pp. 2124-2124.

Jorgensen, S., Decorsiere, R. and Dau, T. (2015) 'Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility', *J Acoust Soc Am*, 137(3), pp. 1401-10.

Kaandorp, M.W. *et al.* (2015) 'Assessing speech recognition abilities with digits in noise in cochlear implant and hearing aid users', *Int J Audiol*, 54(1), pp. 48-57.

Kalikow, D.N., Stevens, K.N. and Elliott, L.L. (1977) 'Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability', *J Acoust Soc Am*, 61(5), pp. 1337-51.

Kayser, S.J. *et al.* (2015) 'Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha', *J Neurosci*, 35(44), pp. 14691-701.

Kemper, S. and Harden, T. (1999) 'Experimentally disentangling what's beneficial about elderspeak from what's not', *Psychology and Aging*, 14(4), pp. 656-670.

Kerr, C.C., Rennie, C.J. and Robinson, P.A. (2008) 'Physiology-based modeling of cortical auditory evoked potentials', *Biol Cybern*, 98(2), pp. 171-84.

Keshishian, M. *et al.* (2020) 'Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models', *Elife*, 9.

Kilic, M.A. and Ogut, F. (2004) 'The effect of the speaker gender on speech intelligibility in normal-hearing subjects with simulated high frequency hearing loss', *Rev Laryngol Otol Rhinol (Bord)*, 125(1), pp. 35-8.

Killion, M.C. *et al.* (2004) 'Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners', *J Acoust Soc Am*, 116(4 Pt 1), pp. 2395-405.

Klimesch, W. (1999) 'EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis', *Brain Res Brain Res Rev*, 29(2-3), pp. 169-95.

Koerner, T.K. and Zhang, Y. (2018) 'Differential effects of hearing impairment and age on electrophysiological and behavioral measures of speech in noise', *Hearing Research*, 370, pp. 130-142.

Kollmeier, B. *et al.* (2015) 'The multilingual matrix test: Principles, applications, and comparison across languages: A review', *International Journal of Audiology*, 54, pp. 3-16.

Kong, Y.Y., Mullangi, A. and Ding, N. (2014) 'Differential modulation of auditory responses to attended and unattended speech in different listening conditions', *Hear Res*, 316, pp. 73-81.

Krause, J.C. and Braida, L.D. (2002) 'Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility', *J Acoust Soc Am*, 112(5 Pt 1), pp. 2165-72.

Krueger, M. *et al.* (2017) 'Relation Between Listening Effort and Speech Intelligibility in Noise', *Am J Audiol*, 26(3S), pp. 378-392.

Kubanek, J. *et al.* (2013) 'The Tracking of Speech Envelope in the Human Cortex', *Plos One*, 8(1).

Kujawa, S.G. and Liberman, M.C. (2009) 'Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss', *J Neurosci*, 29(45), pp. 14077-85.

Lalor, E.C. *et al.* (2006) 'The VESPA: a method for the rapid estimation of a visual evoked potential', *Neuroimage*, 32(4), pp. 1549-61.

Lalor, E.C. *et al.* (2009) 'Resolving precise temporal processing properties of the auditory system using continuous stimuli', *J Neurophysiol*, 102(1), pp. 349-59.

Lash, A. *et al.* (2013) 'Expectation and entropy in spoken word recognition: effects of age and hearing acuity', *Exp Aging Res*, 39(3), pp. 235-53.

Le Prell, C.G. and Clavier, O.H. (2017) 'Effects of noise on speech recognition: Challenges for communication by service members', *Hear Res*, 349, pp. 76-89.

Lee, J.Y. *et al.* (2015) 'Speech Recognition in Real-Life Background Noise by Young and Middle-Aged Adults with Normal Hearing', *J Audiol Otol*, 19(1), pp. 39-44.

Lehiste, I. and Peterson, G.E. (1959) 'Linguistic Considerations in the Study of Speech Intelligibility', *The Journal of the Acoustical Society of America*, 31(3), pp. 280-286.

Leigh-Paffenroth, E.D. and Murnane, O.D. (2011) 'Auditory steady state responses recorded in multitalker babble', *International Journal of Audiology*, 50(2), pp. 86-97.

Lesenfants, D. *et al.* (2019) 'Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations', *Hear Res*, 380, pp. 1-9.

Lewis, E.R. and Henry, K.R. (1989) 'Transient responses to tone bursts', *Hear Res*, 37(3), pp. 219-39.

Luke, R., De Vos, A. and Wouters, J. (2017) 'Source analysis of auditory steady-state responses in acoustic and electric hearing', *Neuroimage*, 147, pp. 568-576.

Lv, J., Simpson, D.M. and Bell, S.L. (2007) 'Objective detection of evoked potentials using a bootstrap technique', *Med Eng Phys*, 29(2), pp. 191-8.

Maddox, R.K. and Lee, A.K.C. (2018) 'Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners', *eNeuro*, 5(1).

Makary, C.A. *et al.* (2011) 'Age-related primary cochlear neuronal degeneration in human temporal bones', *J Assoc Res Otolaryngol*, 12(6), pp. 711-7.

Makhdoum, M.J. *et al.* (1998) 'Intra- and interindividual correlations between auditory evoked potentials and speech perception in cochlear implant users', *Scandinavian Audiology*, 27(1), pp. 13-20.

Manju, V., Gopika, K.K. and Arivudai Nambi, P.M. (2014) 'Association of auditory steady state responses with perception of temporal modulations and speech in noise', *ISRN Otolaryngol*, 2014, p. 374035.

Maris, E. and Oostenveld, R. (2007) 'Nonparametric statistical testing of EEG- and MEG-data', *Journal of Neuroscience Methods*, 164(1), pp. 177-190.

Martin, F.N., Champlin, C.A. and Perez, D.D. (2000) 'The question of phonetic balance in word recognition testing', *J Am Acad Audiol*, 11(9), pp. 489-93; quiz 522.

Mayo, C., Aubanel, V. and Cooke, M. (2012) 'Effect of prosodic changes on speech intelligibility', *13th Annual Conference of the International Speech Communication Association 2012 (Interspeech 2012), Vols 1-3*, pp. 1706-1709.

Meyer, J., Dentel, L. and Meunier, F. (2013) 'Speech recognition in natural background noise', *PLoS One*, 8(11), p. e79279.

Middelweerd, M.J., Festen, J.M. and Plomp, R. (1990) 'Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram', *Audiology*, 29(1), pp. 1-7.

Milaine Dominici, S. *et al.* (2019) 'The Frequency Following Response: Evaluations in Different Age Groups', in Stavros, H., Andrea, C. and Piotr, H.S. (eds.) *The Human Auditory System*. Rijeka: IntechOpen, p. Ch. 7.

Miller, G.A. (1947) 'The masking of speech', *Psychol Bull*, 44(2), pp. 105-29.

Millman, R.E. *et al.* (2017) 'Magnified Neural Envelope Coding Predicts Deficits in Speech Perception in Noise', *J Neurosci*, 37(32), pp. 7727-7736.

Montoya-Martinez, J. *et al.* (2021) 'Effect of number and placement of EEG electrodes on measurement of neural tracking of speech', *Plos One*, 16(2).

Muncke, J., Kuruvila, I. and Hoppe, U. (2022) 'Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise', *Frontiers in Neuroscience*, 16.

Munro, K.J. *et al.* (2020) 'Recording Obligatory Cortical Auditory Evoked Potentials in Infants: Quantitative Information on Feasibility and Parent Acceptability', *Ear and Hearing*, 41(3), pp. 630-639.

Musiek, F. and Nagle, S. (2018) 'The Middle Latency Response: A Review of Findings in Various Central Nervous System Lesions', *Journal of the American Academy of Audiology*, 29(9), pp. 855-867.

Muthukumaraswamy, S.D. (2013) 'High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations', *Frontiers in Human Neuroscience*, 7.

Naatanen, R. (1990) 'The Role of Attention in Auditory Information-Processing as Revealed by Event-Related Potentials and Other Brain Measures of Cognitive Function', *Behavioral and Brain Sciences*, 13(2), pp. 201-232.

Nahum, M., Nelken, I. and Ahissar, M. (2008) 'Low-level information and high-level perception: the case of speech in noise', *PLoS Biol*, 6(5), p. e126.

Navarra, J. (2003) *Visual speech interference in an auditory shadowing task: the dubbed movie effect*. Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona:).

Nejime, Y. and Moore, B.C. (1998) 'Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss', *J Acoust Soc Am*, 103(1), pp. 572-6.

Nilsson, M., Soli, S.D. and Sullivan, J.A. (1994) 'Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise', *J Acoust Soc Am*, 95(2), pp. 1085-99.

Noordhoek, I.M., Houtgast, T. and Festen, J.M. (2001) 'Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz frequency region', *J Acoust Soc Am*, 109(3), pp. 1197-212.

Nourski, K.V. *et al.* (2009) 'Temporal envelope of time-compressed speech represented in the human auditory cortex', *J Neurosci*, 29(49), pp. 15564-74.

Nuesse, T. *et al.* (2019) 'Measuring Speech Recognition With a Matrix Test Using Synthetic Speech', *Trends Hear*, 23, p. 2331216519862982.

O'Sullivan, A.E. *et al.* (2016) 'Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading', *Front Hum Neurosci*, 10, p. 679.

O'Sullivan, J.A. *et al.* (2015) 'Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG', *Cereb Cortex*, 25(7), pp. 1697-706.

Oganian, Y. and Chang, E.F. (2019) 'A speech envelope landmark for syllable encoding in human superior temporal gyrus', *Sci Adv*, 5(11), p. eaay6279.

Paludetti, G. *et al.* (1991) 'Relationships between Middle Latency Auditory Responses (Mlr) and Speech-Discrimination Tests in the Elderly', *Acta Oto-Laryngologica*, pp. 105-109.

Papakonstantinou, A., Strelcyk, O. and Dau, T. (2011) 'Relations between perceptual measures of temporal processing, auditory-evoked brainstem responses and speech intelligibility in noise', *Hear Res*, 280(1-2), pp. 30-7.

Pasley, B.N. *et al.* (2012) 'Reconstructing speech from human auditory cortex', *PLoS Biol*, 10(1), p. e1001251.

Peelle, J.E. (2018) 'Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior', *Ear and Hearing*, 39(2), pp. 204-214.

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(6), pp. 1378-87.

Pena, M. and Melloni, L. (2012) 'Brain oscillations during spoken sentence processing', *J Cogn Neurosci*, 24(5), pp. 1149-64.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1989) 'Speaking clearly for the hard of hearing. III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech', *J Speech Hear Res*, 32(3), pp. 600-3.

Picton, T.W. and Hillyard, S.A. (1974) 'Human Auditory Evoked-Potentials .2. Effects of Attention', *Electroencephalography and Clinical Neurophysiology*, 36(2), pp. 191-199.

Plomp, R. (1994) 'Noise, amplification, and compression: considerations of three main issues in hearing aid design', *Ear Hear*, 15(1), pp. 2-12.

Plomp, R. and Mimpen, A.M. (1979) 'Improving the reliability of testing the speech reception threshold for sentences', *Audiology*, 18(1), pp. 43-52.

Plyler, P.N. and Ananthanarayan, A.K. (2001) 'Human frequency-following responses: representation of second formant transitions in normal-hearing and hearing-impaired listeners', *J Am Acad Audiol*, 12(10), pp. 523-33.

Power, A.J. *et al.* (2012) 'At what time is the cocktail party? A late locus of selective attention to natural speech', *Eur J Neurosci*, 35(9), pp. 1497-503.

Presacco, A., Simon, J.Z. and Anderson, S. (2016) 'Effect of informational content of noise on speech representation in the aging midbrain and cortex', *J Neurophysiol*, 116(5), pp. 2356-2367.

Purcell, D.W. *et al.* (2004) 'Human temporal auditory acuity as assessed by envelope following responses', *J Acoust Soc Am*, 116(6), pp. 3581-93.

Reetzke, R., Gnanateja, G.N. and Chandrasekaran, B. (2021) 'Neural tracking of the speech envelope is differentially modulated by attention and language experience', *Brain Lang*, 213, p. 104891.

Rennies, J. *et al.* (2018) 'Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements', *J Acoust Soc Am*, 144(4), p. EL315.

Rennies, J. *et al.* (2014) 'Listening effort and speech intelligibility in listening situations affected by noise and reverberation', *J Acoust Soc Am*, 136(5), pp. 2642-53.

Renz, T., Leistner, P. and Liebl, A. (2018) 'Auditory distraction by speech: Sound masking with speech-shaped stationary noise outperforms-5 dB per octave shaped noise', *Journal of the Acoustical Society of America*, 143(3), pp. El212-El217.

Ringach, D. and Shapley, R. (2004) 'Reverse correlation in neurophysiology', *Cognitive Science*, 28(2), pp. 147-166.

Roberts, B., Summers, R.J. and Bailey, P.J. (2011) 'The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes', *Proceedings of the Royal Society B-Biological Sciences*, 278(1711), pp. 1595-1600.

Rogers, C.S., Jacoby, L.L. and Sommers, M.S. (2012) 'Frequent false hearing by older adults: the role of age differences in metacognition', *Psychol Aging*, 27(1), pp. 33-45.

Rompelman, O. and Ros, H.H. (1986) 'Coherent averaging technique: a tutorial review. Part 2: Trigger jitter, overlapping responses and non-periodic stimulation', *J Biomed Eng*, 8(1), pp. 30-5.

Ronnberg, J. *et al.* (2016) 'Hearing impairment, cognition and speech understanding: exploratory factor analyses of a comprehensive test battery for a group of hearing aid users, the n200 study', *Int J Audiol*, 55(11), pp. 623-42.

Rosen, S. (1992) 'Temporal information in speech: acoustic, auditory and linguistic aspects', *Philos Trans R Soc Lond B Biol Sci*, 336(1278), pp. 367-73.

Rosen, S. *et al.* (2013) 'Listening to speech in a background of other talkers: effects of talker number and noise vocoding', *J Acoust Soc Am*, 133(4), pp. 2431-43.

Schadler, M.R. *et al.* (2015) 'Matrix sentence intelligibility prediction using an automatic speech recognition system', *Int J Audiol*, 54 Suppl 2, pp. 100-7.

Schestatsky, P., Morales-Quezada, L. and Fregni, F. (2013) 'Simultaneous EEG Monitoring During Transcranial Direct Current Stimulation', *Jove-Journal of Visualized Experiments*, (76).

Schmitt, J.F. (1983) 'The effects of time compression and time expansion on passage comprehension by elderly listeners', *J Speech Hear Res*, 26(3), pp. 373-7.

Schroeder, C.E. and Lakatos, P. (2009) 'Low-frequency neuronal oscillations as instruments of sensory selection', *Trends Neurosci*, 32(1), pp. 9-18.

Sek, A. *et al.* (2015) 'Modulation masking within and across carriers for subjects with normal and impaired hearing', *J Acoust Soc Am*, 138(2), pp. 1143-53.

Shamma, S. (2001) 'On the role of space and time in auditory processing', *Trends Cogn Sci*, 5(8), pp. 340-348.

Shannon, R.V. *et al.* (1995) 'Speech recognition with primarily temporal cues', *Science*, 270(5234), pp. 303-4.

Skoe, E. and Kraus, N. (2013) 'Musical training heightens auditory brainstem function during sensitive periods in development', *Front Psychol*, 4, p. 622.

Small, J.A., Kemper, S. and Lyons, K. (1997) 'Sentence comprehension in Alzheimer's disease: effects of grammatical complexity, speech rate, and repetition', *Psychol Aging*, 12(1), pp. 3-11.

Small, S.A. *et al.* (2018) 'The Effect of Signal to Noise Ratio on Cortical Auditory-Evoked Potentials Elicited to Speech Stimuli in Infants and Adults With Normal Hearing', *Ear Hear*, 39(2), pp. 305-317.

Smiljanic, R. and Bradlow, A.R. (2009) 'Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes', *Lang Linguist Compass*, 3(1), pp. 236-264.

Smits, C. *et al.* (2022) 'The one-up one-down adaptive (staircase) procedure in speech-in-noise testing: Standard error of measurement and fluctuations in the track', *Journal of the Acoustical Society of America*, 152(4), pp. 2357-2368.

Smoorenburg, G.F. (1992) 'Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram', *J Acoust Soc Am*, 91(1), pp. 421-37.

Snik, A.F. *et al.* (1997) 'Speech perception performance of children with a cochlear implant compared to that of children with conventional hearing aids. I. The "equivalent hearing loss" concept', *Acta Otolaryngol*, 117(5), pp. 750-4.

Song, J. and Iverson, P. (2018) 'Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents', *Cognition*, 179, pp. 163-170.

Stapells, D.R., Gravel, J.S. and Martin, B.A. (1995) 'Thresholds for auditory brain stem responses to tones in notched noise from infants and young children with normal hearing or sensorineural hearing loss', *Ear Hear*, 16(4), pp. 361-71.

Stavrinos, G. *et al.* (2018) 'The Relationship between Types of Attention and Auditory Processing Skills: Reconsidering Auditory Processing Disorder Diagnosis', *Frontiers in Psychology*, 9.

Steinschneider, M., Nourski, K.V. and Fishman, Y.I. (2013) 'Representation of speech in human auditory cortex: is it special?', *Hear Res*, 305, pp. 57-73.

Story, B.H. and Bunton, K. (2015) 'Formant measurement in children's speech based on spectral filtering', *Speech Commun*, 76, pp. 93-111.

Svec, J.G. and Granqvist, S. (2018) 'Tutorial and Guidelines on Measurement of Sound Pressure Level in Voice and Speech', *Journal of Speech Language and Hearing Research*, 61(3), pp. 441-461.

Tanaka, A., Sakamoto, S. and Suzuki, Y. (2011) 'Effects of pause duration and speech rate on sentence intelligibility in younger and older adult listeners', *Acoustical Science and Technology*, 32(6), pp. 264-267.

Thornton, M., Mandic, D. and Reichenbach, T. (2022) 'Robust decoding of the speech envelope from EEG recordings through deep neural networks', *Journal of Neural Engineering*, 19(4).

Tomblin, J.B. *et al.* (2015) 'Language Outcomes in Young Children with Mild to Severe Hearing Loss', *Ear Hear*, 36 Suppl 1, pp. 76S-91S.

Tsipouras, M. (2019) 'Spectral information of EEG signals with respect to epilepsy classification', *Eurasip Journal on Advances in Signal Processing*.

Tuller, L. and Delage, H. (2014) 'Mild-to-moderate hearing loss and language impairment: How are they linked?', *Lingua*, 139, pp. 80-101.

Tun, P.A., O'Kane, G. and Wingfield, A. (2002) 'Distraction by competing speech in young and older adult listeners', *Psychol Aging*, 17(3), pp. 453-67.

Uchanski, R.M. (2005) 'Clear Speech', *The Handbook of Speech Perception*, pp. 207-235.

Valderrama, J.T. *et al.* (2018) 'Effects of lifetime noise exposure on the middle-age human auditory brainstem response, tinnitus and speech-in-noise intelligibility', *Hear Res*, 365, pp. 36-48.

Van Canneyt, J., Wouters, J. and Francart, T. (2021) 'Cortical compensation for hearing loss, but not age, in neural tracking of the fundamental frequency of the voice', *J Neurophysiol*, 126(3), pp. 791-802.

van Hoesel, R.J. (2004) 'Exploring the benefits of bilateral cochlear implants', *Audiol Neurootol*, 9(4), pp. 234-46.

Vanthornhout, J., Decruy, L. and Francart, T. (2019) 'Effect of Task and Attention on Neural Tracking of Speech', *Front Neurosci*, 13, p. 977.

Vanthornhout, J. *et al.* (2018) 'Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope', *J Assoc Res Otolaryngol*, 19(2), pp. 181-191.

Vaughan, N.E. *et al.* (2002) 'Time-expanded speech and speech recognition in older adults', *J Rehabil Res Dev*, 39(5), pp. 559-66.

Verschueren, E., Vanthornhout, J. and Francart, T. (2020) 'The Effect of Stimulus Choice on an EEG-Based Objective Measure of Speech Intelligibility', *Ear Hear*, 41(6), pp. 1586-1597.

Verschueren, E., Vanthornhout, J. and Francart, T. (2021) 'The effect of stimulus intensity on neural envelope tracking', *Hearing Research*, 403.

Weissbart, H., Kandylaki, K.D. and Reichenbach, T. (2020) 'Cortical Tracking of Surprisal during Continuous Speech Comprehension', *J Cogn Neurosci*, 32(1), pp. 155-166.

Wingfield, A. (1996) 'Cognitive factors in auditory performance: context, speed of processing, and constraints of memory', *J Am Acad Audiol*, 7(3), pp. 175-82.

Wingfield, A. *et al.* (1999) 'Regaining lost time: adult aging and the effect of time restoration on recall of time-compressed speech', *Psychol Aging*, 14(3), pp. 380-9.

Wong, L.L., Cheung, C. and Wong, E.C. (2008) 'Comparison of hearing thresholds obtained using pure-tone behavioral audiometry, the Cantonese Hearing in Noise Test (CHINT) and cortical evoked response audiometry', *Acta Otolaryngol*, 128(6), pp. 654-60.

Yang, M.D. *et al.* (2015) 'Speech Reconstruction from Human Auditory Cortex with Deep Neural Networks', *16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Vols 1-5*, pp. 1121-1125.

Yanhong, O., Chandra, M. and Venkatesh, D. (2013) 'Mild cognitive impairment in adult: A neuropsychological review', *Ann Indian Acad Neurol*, 16(3), pp. 310-8.

Zaltz, Y. *et al.* (2020) 'Listening in Noise Remains a Significant Challenge for Cochlear Implant Users: Evidence from Early Deafened and Those with Progressive Hearing Loss Compared to Peers with Normal Hearing', *Journal of Clinical Medicine*, 9(5).

Zekveld, A.A. *et al.* (2006) 'Top-down and bottom-up processes in speech comprehension', *Neuroimage*, 32(4), pp. 1826-36.

Zhang, X.C., Gong, Q. and Zhang, T. (2016) 'Cortical Auditory Evoked Potentials (CAEPs) Represent Neural Cues Relevant to Pitch Perception', *2016 38th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (Embc)*, pp. 1628-1631.

Zhuang, X.W., Yang, Z.S. and Cordes, D. (2020) 'A technical review of canonical correlation analysis for neuroscience applications', *Human Brain Mapping*, 41(13), pp. 3807-3833.

Zou, J.J. *et al.* (2019) 'Auditory and language contributions to neural encoding of speech features in noisy environments', *Neuroimage*, 192, pp. 66-75.