Combining Speech with Sight and Touch; Investigating the Benefits of Audio-Visual and Audio-Tactile Speech on Speech Intelligibility and Cortical Speech-Envelope Tracking Accuracy

Brandon Lee O'Hanlon, MSc

Department of Psychology

Lancaster University

Thesis submitted for the degree of

Ph.D., Doctor of Philosophy

25th October 2024

Table of Contents

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, in whole or in part, for a degree at this, or any other university. For transparency, a small portion of the data presented in Chapter 2 were collected during my MSc, which comprised of pilot work for my thesis; however, Chapter 2 bears no resemblance to my MSc thesis. This thesis does not exceed the maximum permitted word length of 80,000 words including tables, figures, and footnotes, but excluding the thesis abstract, bibliography, and appendices.

Name: Brandon Lee O'Hanlon

Date: 27th October, 2024

Acknowledgements

Without a doubt, I would not have made it this far in my academic education and career without the wonderful support of my supervisors, Helen Nuttall and Christopher Plack.

To Helen, you have guided my learning and development since my early years as an undergraduate student – from providing my first research experience on the fabled ‘Sling Study’ to being there with the NoSA lab during my first conference trip to Denmark. Through thick and thin, the ups and the downs, I have had the pleasure of knowing you as a supervisor, a mentor, and a friend for seven years. I will be ever grateful for your support, and I can only hope that I am able to repay your kindness for years to come with continued collaboration. To Chris, I thank you for being a brilliant wealth of knowledge, always ready and willing to teach and to guide with your expertise. You strive to be as best as you can be with your work and have motivated me to improve time and time again. Thank you for your endless wisdom.

To the wider NoSA lab group, thank you for the brilliant memories and the fantastic support throughout my many years with you all. To Jess Pepper, my dear friend and fellow postgraduate researcher. I have no doubt in my mind that you will achieve the greatest of achievements with your work, again and again and again. I thank you massively for all the support, be it through our first conference trips, experimental code and analysis, or our inevitable trips to Fylde bar for pints after work. You are a shining example of a best friend and colleague. I cannot wait to see what you achieve next, Pepper. And to Kate Slade, thank you for all your assistance and support. You never cease to cheer up all those around you, be it at a lab meeting or through stand-up comedy. To my wonderful officemates and fellow postgraduate members of NoSA: Jessica Andrew, Haydn Farrelly, and Nez Sharp. Thank you for your guidance and support, for our office chats and breaks to play neuroscience-based games on the whiteboard. To my great research assistants, Neve Ferrie and Jessica Downes, who assisted me with the long lab hours on our lengthy training study. I thank you both for your hard work and dedication. The NoSA lab is a wonderful, amazing team who is always there to support one another. I am so grateful to have been a part of it and wish to continue to help this lab grow and flourish.

I would also like to thank all the incredible minds who have collaborated on my research projects, offering crucial insight from new perspectives that have elevated my own

understanding of Psychology and Neuroscience. To Barrie Usherwood, whose genius as a research technician gave rise to the tactile stimulation device that became my life for two whole years of research. To Lars Hausfeld, for your expertise on neural tracking and for teaching me frisbee golf in the fields of Nyborg. I will one day be semi-decent at it. To Lars Riecke, for your expertise in neural tracking and audio-tactile integration. And finally, to Sana Hannan, for your expertise in invasive methodologies and interictal activity.

To my loving partner, Dale Hargrove, who means the world to me and more. Your constant encouragement has kept me pushing forward and forward. Your awe and interest in my work – be it on a technical or base level – has pushed my motivation for research to new heights. To have such a strong pillar of support in my life has been such a boon – and the benefits of which I hope to come through in my work. To my parents, Sean O'Hanlon and Sonya Prem. Despite not always understanding the deeper neuroscience I sometimes ramble about at family meals; you have always listened and been proud. Your encouragement and support have gotten me to where I am now. To all my family, from my cousins to my aunties and uncles, my brothers and my grandparents. Thank you for checking in on me and wishing me luck, especially in the final stretch of completing this thesis.

For living with me these many, many years, I thank my dear friends and housemates Holly Pritchard, Kyle Stonehouse, Yumna Patel, and Marie Ceesay. We have been through it all, supporting one another as we need to and making our lives at Lancaster that much sweeter. My work here would not be the same without your support. Holly, I thank you for being with me throughout all the hills we have climbed together, both metaphorically and physically. Kyle, I thank you for constant reassurance and perseverance – your success has driven me to find my own, be it through my academic work or on the climbing wall. Yumna, I thank you for all the time spent enjoying our hobbies together, from cosy gaming, to live orchestras, and to discovering new interests together. And Marie, I thank you for your fresh perspectives and your encouragement to learn from new places and ideas.

And lastly, to my wonderful bunch of loveable friends: to Millie, Benji, Aoki, Joe, and Ran. For all the nights spent playing games and watching movies, chatting and engaging in our hobbies and interests. To all the wonderful days spent with you all at various events and get-togethers. These cherished moments of respite and your ever-present words of encouragement for my work have meant the world to me.

I could not have done this without you all.

Abstract

Combining Speech with Sight and Touch; Investigating the Benefits of Audio-Visual and Audio-Tactile Speech on Speech Intelligibility and Cortical Speech-Envelope Tracking Accuracy

Brandon Lee O'Hanlon, MSc

Speech plays a critical role in communication in our everyday lives. Understanding how speech is processed in the brain, from mechanoelectrical transduction in the ear to processing in the subcortical and cortical regions of the auditory pathway, provides insight into how we may improve speech perception during difficult listening conditions which may hinder us. As discussed in Chapter 1, multisensory integration is one such method in which we may restore intelligibility to speech-in-noise. This thesis aims to contribute to the current multisensory speech integration literature through new investigations into audio-visual and audio-tactile speech and their benefits to cortical speech-envelope tracking in the auditory cortex and speech intelligibility.

Chapter 2 reassessed the audio-visual benefits to speech perception in-noise when phoneme stimuli are selected from different visually distinct viseme categories. It found that, when visemes are considered, the benefits of audio-visual integration were reduced compared to previous literature, with some phonemes providing more benefit to intelligibility than others. Chapters 3 and 4 explored potential benefits of both bottom-up and top-down short-term audio-tactile training on cortical speech-envelope tracking accuracy and speech intelligibility. Within these chapters, findings indicated an initial enhancement on neural tracking accuracy but with no associated benefit to speech intelligibility. This unexpected

outcome was further investigated with Chapter 5, which returned to audio-visual speech to examine the potential role of neural tracking in the prediction of oncoming speech.

In all, this thesis provided evidence of multisensory benefits to speech perception, especially in difficult listening conditions. The thesis also contributes to the growing literature around neural tracking in the brain, with further evidence to suggest that tracking accuracy is not intrinsically linked to intelligibility. Chapter 6 discusses these findings, alongside future directions, to finalise this thesis.

Chapter 1

1 Introduction

Speech, alongside gesture, plays a large role in communication between people in social settings. However, speech signals are highly complex and are not always easily processed and decoded in the brain when trying to understand what someone is saying, particularly in difficult listening conditions. Understanding the neural underpinnings of speech processing and perception will allow us to develop new methods of improving communication in everyday life. This is especially true for those with hearing loss, as understanding how the speech signal is transformed throughout the auditory pathway lends great insight into where and how speech processing may be affected. This loss of hearing can be alleviated with new hearing technologies, such as cochlear implants and hearing aids, but may not necessarily mean that speech processing ability is comparable to normal-hearing populations. Difficult listening conditions, such as an environment where many people are talking, further affect those with hearing loss, but also normal-hearing populations too, with reductions seen in speech intelligibility and neural tracking accuracy. Multisensory integration naturally occurs between the visual and auditory systems, with visual lipreading assisting with speech when difficult listening arises using viseme identification. Despite audio-visual integration playing a large role in the restoration of speech intelligibility, visual lipreading is not always possible. It may be possible to train audio-tactile integration in speech perception to enhance the utility of speech-relevant tactile information in difficult listening conditions when lipreading is inaccessible.

This thesis explores: the benefits of audio-visual integration to speech perception when visually distinct phonemes are selected as stimuli; the potential benefits of audio-tactile integration to neural representations in the auditory cortex and to speech intelligibility, both

when training is bottom-up and top-down; and finally the potential role of neural tracking in the prediction of oncoming speech through the lens of audio-visual integration in the posterior superior temporal gyrus. The following literature review will discuss the topics of the components of a speech signal, neural speech processing, difficult listening conditions, multisensory integration, and sensory training, providing an overview of the relevant literature that has built up the aims and objectives of this thesis.

1.1 The Speech Signal

Speech is made up of two primary components: the temporal fine structure and the speech envelope. The temporal fine structure relates to the rapid, individual pressure variations in the air that reach the ear. This provides information about pitch (Moore, 2008; Santurette & Dau, 2011; Moon & Hong, 2014), frequency discrimination (Hopkins & Moore, 2010; Moon & Hong, 2014), and spatial location of speech in our environment (Borjigan & Bharadwaj, 2019). The speech envelope, on the other hand, relates to the slow temporal fluctuations in the overall amplitude of a speech signal over time. This envelope provides critical information for understanding speech (Shannon *et al*., 1995; Ahissar *et al*., 2001; Apoux & Healy, 2013). Shannon and colleagues (1995) extracted speech envelopes at multiple frequency bands and presented these envelopes to participants, finding that the intelligibility of speech increased with the addition of each band. Despite the speech presented being highly degraded and lacking in fine structure, participants could still recognise consonants, vowels, and words with high rates of accuracy, demonstrating the importance of the speech envelope in our understanding of speech. Studies using auditory chimaeras (where the fine structure from one sentence is combined with the envelope of another) have further found distinctions between envelope and fine structure processing. In the case of Smith and colleagues (2002), when chimaeras were presented to participants the envelope of the chimaera was more influential on that participant's understanding of the

speech heard than the fine structure. The more frequency bands were introduced, the more the envelope of the chimaera influenced their understanding of the speech. The fine structure on the other hand provided further clarity of the speech, such as source location, consistent with other studies (Heinz & Swaminathan, 2009). Further chimerical evidence from Smith *et al.* (2002) indicated that sound location was defined by the temporal fine structure alone, whereas word identification is derived from the speech envelope.

This finding has been supported by more recent work by Warnecke *et al.* (2020), who investigated the role of the fine structure and envelope in processing non-stationary sounds in the environment. Warnecke and colleagues (2020) found that the removal of low-frequency temporal fine structure reduced participants' sensitivity to capturing sound motion in the environment, leading to worsened source location ability. Conversely, when the speech envelope fluctuated, participants were more likely to assume a sound was stationary in the environment. As shown, the wealth of research into the temporal fine structure and the speech envelope over the past few decades has highlighted key differences between the two components of speech in their purpose in speech processing. Understanding the difference between the speech envelope and the temporal fine structure provides a clear focus for neuroscientific research to understand how each component is processed in the brain, and how decoded information from both is combined to provide understanding of the speech perceived. Of particular interest is the speech envelope and its apparent importance in speech segmentation and speech understanding (Shannon et al., 1995; Ahissar et al., 2001; Apoux & Healy, 2013). However, it is also important to consider how the processing of both speech components may occur in the brain in parallel, with each providing crucial information for segmentation of speech units to assist with speech processing (Teng *et al.* 2019). For example, Teng *et al*. (2019) showed evidence of both the temporal fine structure and the speech envelope entraining cortical responses at low frequencies, with the fine structure

providing comparable temporal information for speech processing as the envelope. Here, it is argued that both components are responsible for aiding speech segmentation and suggest that research into speech segmentation should be with both components of speech in mind.

To further understand the importance of these components, it is crucial to understand the auditory processing pathway detailing the complex process of listening in the brain.

1.2 Listening in the Brain

Auditory processing in the brain begins with stimulation from transduction of inner hair cells in the cochlea as they respond to sound, leading to electrical activity passing through the auditory nerve and reaching the cochlear nucleus in the brainstem (Plack, 2018). From here, the cochlear nucleus sends auditory information along afferent pathways towards the inferior colliculus, either through the dorsal, anterior ventral, or posterior ventral cochlear nucleus (Pickles, 2015; Kunchur, 2023). Distinctions in purpose between these differing regions of the cochlear nucleus have been found. The dorsal cochlear nucleus assists with sound localisation (May, 2000; Trussell & Oertel, 2018) and some early multisensory integration (Shore, 2005; Balmer & Trussell, 2021). The anterior ventral cochlear nucleus, on the other hand, plays a crucial role in the encoding of the speech envelope through T-stellate and bushy cells (Rubio, 2018), which act as time-coding cells to track the onset and offset of dynamic, more complex auditory experiences whilst further assisting sound localisation (Kuenzel, 2019; Kunchur, 2023). Finally, the posterior ventral cochlear nucleus acts as the branching, early afferent pathway responsible for extracting information from the temporal fine structure and amplitude modulations via octopus cells (Rebhan & Leibold, 2021; Kunchur, 2023). The cochlear nucleus is also thought to engage with efferent pathways, both from the inferior colliculus back to the cochlear nucleus and from the cochlear nucleus back to the ear, via medial and lateral olivocochlear neurons (Balmer & Trussell, 2022; Farhadi, 2023). This olivocochlear system may play a role in selective attentional processes in which information regarding noise in an

environment is detected and the cochlea is informed for noise suppression or intensity gain (Jennings, 2020; Hockley *et al.*, 2022; Lauer *et al.*, 2022). However, conflicting evidence does suggest that attention is not processed in the olivocochlear system, prompting further study (Kikuchi *et al.*, 2023; Gafoor *et al.*, 2023; Jedrzejczak *et al.*, 2020).

After processing at the cochlear nucleus, sound then reaches the inferior colliculus. Many roles of the inferior colliculus have been discussed, providing a general view of the inferior colliculus as a central hub in the auditory system for integrating sound information from afferent and efferent pathways by utilising tonotopic arrangement of neurons (Plack, 2018; Driscoll & Tadi, 2020; Kunchur, 2023). This tonotopic arrangement allows information from other nuclei – such as the various branches of the cochlear nucleus or the medial geniculate body – to arrive to specific layers of the inferior colliculus dependent on neuronal frequencies. It is likely that this auditory region, alongside the superior colliculus, plays a role in communicating to other midbrain and other sensory cortical regions to enact multisensory integration, like the sensorimotor and visual cortices (T. Ito *et al.*, 2021; S. Ito *et al.,* 2021; X. Liu *et al.*, 2022; Bean *et al.*, 2023). Indeed, an example of this for audio-visual integration is the inferior-superior colliculi circuit, which is thought to take auditory and visual cues in the environment for the driving of selective attention tin the primary visual cortex (Hu & Dan, 2022). Furthermore, evidence does suggest that the inferior colliculus plays a part in pitch perception (Braun, 2000) and spatial coding (Palmer & Kuwada, 2005; Litovsky *et al.*, 2002), suggesting that the temporal fine structure – given its importance in understanding the pitch and source of a speaker - is heavily processed at this region. However, it may be that the inferior colliculus binds this information together after they have been extracted and processed by other areas. It is important to note that the inferior colliculus in recent study has been shown to contain sub-regions in a similar manner to the cochlear nucleus, with both lemniscal (central nucleus) and non-lemniscal (dorsal cortex, and external cortex) layers (M. Liu *et al.*, 2022).

Here, it appears that the specific sub-region of the inferior colliculus that processes non-auditory information for multisensory integration is the external cortex, whereas the dorsal cortex focuses on auditory information projected from and to the medial geniculate body of the thalamus. This medial body is the next relay region along the auditory pathway following the inferior and superior colliculi, with three sub-regions (the dorsal, ventral, and medial) connecting to previous sub-regions of said colliculi (see Bartlett, 2012). As a thalamic body, it has been suggested that the medial geniculate acts simply as a region of relay for information to travel to cortical regions, although this may not be the case (see Winer, 1984; Bartlett, 2012). Evidence suggests that sub-regions of the medial geniculate body are modulated by corticofugal projections from the auditory cortices (Ojima & Rouiller, 2010; Winer *et al.*, 2001). For a more specific example, the modulation of the medial sub-region by the secondary auditory cortex has been highlighted by Luo *et al.* (2022) as a means by which auditory threat memory can be accessed and potential threatening sounds in the environment may be assessed with attentional priority. Other thalamic bodies are present that play a similar role for different sensory processes, such as the lateral geniculate nucleus experiencing corticofugal projection from the primary visual cortex (O'Connor *et al.*, 2002; Meng & Schneider, 2022). Though, the extent to which these thalamic regions communicate locally through thalamic projections during multisensory integration is difficult to discern.

The auditory cortex is the final higher-level area on the auditory neural pathway and is situated on the superior temporal gyrus and sulcus. Sub-regions of the auditory cortex include the primary cortex located largely on – though not exclusively - Heschl's gyrus, and secondary cortex regions like the planum temporale and the planum polare (see Poeppel *et al.,* 2012). Whilst these sub-regions have been evidenced to engage in notably different stimulus contexts – with primary regions mostly processing simplistic acoustic features and nonprimary regions complex semantic context of sounds, such as in speech (Ahveninen *et al.*, 2006; de Heer *et al.*,

2017) – these processing streams do appear to work in parallel, shown with intracranial electrocortical stimulation work by Hamilton *et al.* (2021). Notably, many neuroanatomical studies of the auditory cortex (and other subcortical regions as well) are animal-based studies. The human auditory cortex is exceptionally complex and vast, with new sub-regions being discovered in recent work (Zachlod *et al.*, 2020) that may not be present in animal models in the same manner (see Brewer & Barton, 2016). Still, knowledge of the functional role of the human auditory cortex is well documented. The auditory processing includes complex auditory working memory, speech processing, recognition, and prediction (Yu *et al.,* 2021; King & Schnupp, 2007; King *et al.*, 2018), as well as influences on pitch perception, motion, frequency discrimination, spatial location, and more (Hall *et al*., 2003; Hall & Barker, 2012), the outputs of which can be used in top-down descending corticofugal, corticollicular, and corticothalamic projections to influence previous subcortical regions as discussed above. Interestingly, the auditory cortex does comprise of lateralisation differences in organisation and function, with the left-hemisphere's superior temporal sulcus being shallower than the right-hemisphere's, in example (Moerel *et al.*, 2014; Tzourio-Mazoyer *et al.*, 2020). Functionally, Zatorre and Belin (2001) demonstrated that the left hemisphere shows greater responses to stimuli that showed temporal change in frequency whilst the right hemisphere shows greater responses to stimuli with spectral changes, though both hemispheres still showed some response to both temporal and spectral features. This lateralisation is guided by handedness, with right-handed individuals showing left-hemisphere bias in speech processing and left-handed individuals showing bilateral activity (Papadatou-Pastou, 2011; Potdevin *et al.*, 2023). This handedness and lateralisation link is not just exclusive to speech and language centres in the brain, as sensorimotor function also exhibits lateralisation (Sainburg, 2014). This is further shown in multisensory contexts too, with Koskinen *et al.* (2020) showing left-lateralisation for

contextual influences on predicting oncoming speech, with regions involving the auditory and sensorimotor cortices.

In all, the processing of sound in our environment in the brain follows a complex path through subcortical and cortical systems, with descending and ascending projections. In relation to the specific components and units of speech, however, further investigation is required to understand exactly how the speech envelope and temporal fine structure are decoded within the brain. As a crucial element of speech intelligibility (Shannon *et al*., 1995; Ahissar *et al.,* 2001; Apoux & Healy, 2013), the decoding of the speech envelope remains of particular interest. A further finding from Koskinen *et al.* (2020)'s work highlighting speech-processing lateralisation was that theta frequency coupling to the speech envelope was reflective of the syllabic rate of perceived speech, whilst low-delta frequency coupling was reflective of speech prosody. Oscillatory activity in the auditory system may help further uncover more detail about the neural underpinnings of speech processing and how it differs from nonspeech auditory processing.

1.2.1 Oscillatory Activity in the Auditory Cortex

Neuronal oscillations occur across the brain and the sensory cortices and can be banded by specific frequency speeds (see Jensen *et al.*, 2019). These frequencies of activity can be as low as in the delta (0.1 – 4 Hz; Steinmetzger & Rosen, 2017; Morillon *et al.*, 2019), theta (4 – 8 Hz; Steinmetzger & Rosen, 2017), and alpha ranges (8 – 12 Hz; Becker, *et al.*, 2018), extending to higher frequencies with beta (13 – 30 Hz; Morillon *et al.*, 2019; Cabral *et al.*, 2022), gamma (30 – 90 Hz; Crone *et al.,* 2011), and even as high as 200 Hz with broadband high-gamma activity (Crone *et al.,* 2011). The boundaries for each frequency band are not entirely consistent in the literature, however. For example, some have determined theta to be between 4 and 7 Hz and beta between 13 and 25 Hz (von Stein & Sarnthein, 2000; Beste *et al.*, 2023), whilst others have debated alpha oscillating between 7 and 15 Hz (Tripathi, 2022).

Indeed, some definitions of lower frequency bands also indicate variance, with Boucher *et al.* (2019) categorising theta activity as between 3 Hz and 10 Hz, with delta residing below 3 Hz. This lack of standardisation in the literature does mean that care should be taken when making direct comparisons between studies about specific frequency bands as to exactly which boundaries are being suggested. Regardless of the specificities of the boundaries, a common understanding about oscillatory activity is that each band represents a different contribution to brain function (Gourévitch *et al*., 2020). As a general example, alpha-band activity is thought to play a role in inhibition across the brain as well as an indicator of resting state (Jensen & Mazaheri, 2010; Scheeringa *et al.*, 2012; Lombardi *et al.*, 2023). Oscillatory waves in neuronal clusters of specific brain regions are non-sinusoidal (Cole & Voytek, 2017), and, as detailed by Schroeder *et al.* (2008), fluctuate between states of high and low excitability as the oscillatory activity continues (see also Lakatos *et al.*, 2005). Separate brain regions can communicate and integrate with one another, despite operating different frequency oscillations, through hierarchical phase-amplitude coupling (Esghaei *et al.*, 2022). For example, beta-gamma phase-amplitude coupling would modulate the power of gamma (the higher of the two frequencies) along the phase of beta (the lower of the two frequencies), allowing both to transmit oscillatory information without disrupting the other. This phase-coupling also exists locally, such as in the processing of more complex speech stimuli (Schroeder, *et al.*, 2008).

In the context of speech, oscillatory bands are seen to have slightly different contributions to speech perception. For example, phoneme-level processing seems to occur in the delta and theta band ranges (Di Liberto *et al.*, 2015), yet rhythm and intelligibility of speech are reflected separately between delta and theta frequencies respectively (Ding & Simon, 2014). Furthermore, Steinmetzger and Rosen (2017) demonstrated delta power increasing in the second half of continuous, intelligible sentences. This serves as key evidence in the role of delta oscillatory activity in speech intelligibility, likely reflecting the benefits of the low-

frequency speech envelope to understanding of speech. Auditory delta activity is further thought to interact with beta motor activity during sensorimotor integration in the left precentral gyrus through phase-amplitude coupling (Morillon *et al.,* 2019), likely reflective of a network for speech production, though this remains unclear (Morillon *et al.*, 2015). Whilst not directly related to speech processing, beta-gamma activity (20 – 35 Hz) projecting from the ventromedial prefrontal cortex has been associated with unexpected learning outcomes from interactions in our environment (Marco-Pallarés *et al.*, 2015). Thus, beta-gamma coupling may arise in response to sudden, positive sounds processed in the environment and when learning language. Furthermore, theta-gamma coupling appears to play a role in speech comprehension during initial listening (Lizarazu *et al.* 2019). These examples of gamma coupling in auditory processing may be reflective of additional gamma input from short-term selective attentional processing in the prefrontal cortex (see Kaiser & Lutzenberger, 2005). Finally, looking at alpha-band activity, it has been suggested that alpha oscillations actively regulate and monitor long-range dependence across the brain, allowing for inhibition of current and future activity (Becker *et al.* 2018). As outlined by Guan *et al.* (2023), long-range dependence conceptualises how current-in-time oscillatory activity may influence the course of activity at a future state, both in the same or a differing brain region. This could be related to attenuation of continuous noise in the environment during selective attention of a target speaker (Daly & Pitt, 2021). Moreover, there are differences seen in peak resting-state alpha frequencies between visual and auditory cortices, with 'visual alpha' peaking typically at the higher alpha ranges (10 – 13 Hz) and 'auditory alpha' in the lower alpha ranges (around 8.5 Hz) (Capilla *et al.* 2022).

Taken together, it does appear that all oscillatory frequency bands play some role in the processing of speech. Primarily, speech processing and comprehension seems driven by delta- and theta-band activity, with hierarchical coupling to higher frequencies occurring when speech processing involves working memory (Hjortkjær *et al.*, 2020), attention (Kaiser &

Lutzenberger, 2005), or other higher-level auditory processes (see Meyer, 2018). Delta and theta range oscillations are informative, as they exhibit slow moving frequency changes much like the speech-envelope expresses the slow fluctuating changes in amplitude of the speech signal over time. As previously discussed, the speech-envelope plays a critical role in speech intelligibility and comprehension (Shannon *et al*., 1995; Ahissar *et al*., 2001; Apoux & Healy, 2013). The exact way in which the speech-envelope is processed to reach speech understanding is not entirely clear; phase-locking of neuronal activation to changes in the speech-envelope, and delta/theta neural entrainment may provide further insight (Peelle & Davis, 2012). Specifically, the recent emergence in auditory neuroscience of investigations into cortical speech-envelope tracking in the brain will next be discussed.

1.2.2 Cortical Speech-Envelope Tracking

One such way the auditory system may be utilising the speech-envelope is through phase-locking in the auditory cortex. Phase-locking – or neural tracking – is the process in which electrical activity in the auditory pathway (particularly cortical regions) 'lock' into patterns that closely match the stimuli in question (Heil & Peterson, 2015; Plack, 2018). A nonspeech auditory example of this is through consistent musical beats, such as a drumline, to which the auditory system can match its neural activity to the low-frequency beats of the drum (Zuk *et al*., 2021). In the case of speech, phase-locking would represent the brain's ability to match its oscillatory activity to temporal information from the speech perceived. The auditory cortex has been shown to phase-lock to the speech envelope of attended streams, identified through the low-frequency neural activity that correlates with said envelopes (Ding & Simon, 2014; see also Ding *et al.*, 2014; Issa *et al.,* 2024). This specific process is known as cortical speech-envelope tracking - also referred to as neural speech tracking. Neural speech tracking may be important for segmenting and quickly processing multiple, smaller units of speech, allowing for easier recognition (Giraud & Poeppel, 2012),

though the causal role of this tracking remains unclear (Kösem *et al.*, 2023). This cortical tracking of speech-envelopes is present in infancy too, with evidence from Barajas *et al.* (2021) indicating tracking is present from birth for both familiar and unfamiliar languages (see also: Menn *et al.*, 2022; Jessen *et al.*, 2021). Barajas *et al.* further determined that changes to neural tracking occur at 6 months of age, with the ability to track unfamiliar languages reducing as a primary language develops. This is particularly interesting, as it matches perceptual phoneme boundary development at a similar age, wherein infants lose the ability to perceive phonemes that are not present in their primary language (Trehub, 1976; Werker & Tees, 1984). Furthermore, neural tracking remains robust in older adults (aged 65-80 years: Kurthen *et al*., 2021), specifically for delta band oscillations (McHaney *et al*., 2021).

Speech envelope tracking exists in the low-frequency oscillatory ranges on most accounts, including the delta band (0.5-4 Hz; Etard & Reichenbach, 2019; Ding & Simon, 2013; Ding & Simon, 2014), theta band (4-8 Hz; Etard & Reichenbach, 2019), and low alpha band (8-15 Hz; Dimitrijevic *et al*., 2017), though can also occur in the gamma band (Kubanek *et al.,* 2013). It has been suggested that phonemes are represented through tracking of the gamma band (above 30 Hz), syllables in the theta band (4 – 8 Hz), and intonational phrase boundaries in the delta band (0.5 – 4 Hz) (Meyer, 2018). The intonational phrase boundaries (ITBs) denote the separation of chunks of speech (phrases, sentences, or parts of sentences) that may be spoken within or between breath gaps. Delta band processing of ITBs potentially reflects how low-frequency speech envelope tracking may be used to segment speech into chunks for easier processing. Evidence from invasive electrocorticography (ECoG) studies sheds further light on the auditory cortex's tracking of speech envelopes (Kubanek *et al*., 2013). ECoG is an invasive neuroscientific method of recording electrical activity directly from cortical surfaces on the brain (as well as deeper subcortical regions), providing a more

robust and direct measure of cortical activity than other non-invasive methods like EEG (Ball *et al.*, 2009; see also: Kanth & Ray, 2020). Due to the invasive nature of ECoG, however, participant pools typically consist of epileptic surgery patients, limiting the generalisability of such measures (Schomer & Da Silver, 2012). Kubanek *et al.* (2013) found that gamma bands were highly correlated to speech-envelope information in the belt regions of the auditory cortex. Of note, these early belt regions also tracked the envelope of musical melody and not solely speech. Likewise, higher areas of the auditory cortex, such as the superior temporal sulcus and Broca's area, tracked the envelope of speech through gamma oscillations. This was only apparent for speech envelopes, suggesting these areas are responsible in part for the linguistic analysis of speech. There also appears to be a dominance of low-frequency envelope tracking in the right hemisphere over the left, though the implications of this distinction are yet to be discerned (Abrams *et al.,* 2008). As a note of distinction, Pefkou and colleagues (2017) identified that lower theta frequencies between 4 and 8 Hz were sensitive to the syllabic rate of the speech heard but reflected no sensitivity to the understanding of said speech, whereas the beta frequency range (14 – 21 Hz) reflected the inverse. Other studies have corroborated findings on theta frequency contributions to envelope tracking but have discerned delta frequencies (0.5 and 4 Hz) to be essential during tracking for speech comprehension (Vanthornhout *et al*., 2018). More clarification in the literature on the purpose of cortical speech tracking to each frequency band may be needed.

These discrepancies may be due to differences in research methods, such as ECoG, magnetoencephalography (MEG), and electroencephalography (EEG). As an invasive methodology, ECoG may be better tuned to detect gamma frequency tracking in auditory regions. Higher frequencies may be tracked in the auditory regions that non-invasive techniques like EEG and MEG are not able to detect or highlight because of poor signal-to-noise quality (Todaro, *et al.*, 2019). Obleser *et al.* (2012) also argued against the envelope

being the solely tracked source of speech information, suggesting that the auditory cortex phase-locks and tracks across the low and high frequency bands - not just slow, temporal fluctuations. It may be that other features of speech are equally as important to speech intelligibility as the envelope information, such as spectral content or the temporal fine structure. This is further corroborated by Dimitrijevic and colleagues (2017), who show that alpha and beta oscillations above 15 Hz are a strong predictor of speech intelligibility, despite other work suggesting that the frequencies below 15 Hz are more critical for speech understanding (Pefkou et al., 2017). More points of contention arise when looking at neural tracking earlier in the auditory pathway, such as in the brainstem. For example, previous research into the auditory brainstem response (ABR: see Kraus, 2011) - and the resulting frequency-following response (FFR: see Krizman & Kraus, 2019) - looked at high-frequency responses to auditory stimuli (such as phonemes or sinewave sounds), finding encoding in the auditory nerve with sinusoidal (therefore non-oscillatory) modulations at 88 and 39 Hz (Herdman *et al*. 2002). Here, however, it is important to note that brainstem response studies are limited in their capacity to investigate phase-locking at higher frequencies, as phase-locking is not seen at high frequencies past 1000 Hz (Palmer & Russell, 1986; Verschooten & Joris, 2014). Another point of discrepancy can come from univariate (as in, through single-channel analysis) versus multivariate (through many channels) methods of investigating tracking accuracy. Investigating univariately with event-related potentials (ERPs), Aiken and Picton (2008) found that posterior auditory cortex regions follow the speech envelope. These ERPs were present when changes in the speech envelope occurred, reflecting the brain's representation of changes in the envelope through cortical speech-envelope tracking. They found that vertical dipole source waveforms significantly correlated with speech envelopes. Here, work was univariate and applied across multiple channels in dipole-source analyses. Thus, whilst it is possible to evidence neural tracking during listening using ERPs, it is

difficult to say whether these univariate measures convey a full picture of tracking across the auditory system. This is especially important as we know many other regions play a role in our understanding of speech, such as posterior cortical regions utilised in semantic processing (Dick *et al*., 2009). Therefore, multivariate approaches to investigating cortical speech tracking may be preferable as to cover the wide range of neural networks engaged in speech perception.

1.3 Investigating Neural Tracking Accuracy

Cortical speech-envelope tracking accuracy – a correlational measure of how well-represented a speech-envelope is in the auditory cortex via phase-locking – can be obtained through many functions, including temporal response functions (Crosse *et al.*, 2016). Forward-modelled temporal response functions (TRFs) are univariate, in that channels of neural recordings can only be related to a stimulus feature independently. As outlined by Crosse *et al.*, more complex stimuli in our environment, particularly speech, is decoded through multiple frequency channels in the cochlea, and - as discussed – many afferent and efferent pathways throughout the auditory pathway. To reduce this decoding to a singular point of reference in neural recordings would not be ecologically valid. Multivariate TRFs (or mTRFs), on the other hand, can be computed to derive tracking accuracy from complex speech listening. Through backwards modelling, these mTRFs can be used in stimulus reconstruction, reconstructing a stimulus feature like the speech envelope using neural data across any number of recorded channels at once during continuous speech presentation. This provides a more valid assumption of the neural representation of the speech envelope across all auditory networks in the brain that have been captured by neural recordings, both invasively and non-invasively. Stimulus reconstruction makes use of a decoder model, which computes the reconstructed envelope of speech across designated time lags and weights each channel of neural data when mapping them back onto to the original speech envelope. The

decoder is trained on a subset of the neural data and tested on the rest. Cross-validation of training and test subsets are typically done using a 'leave-one-trial-out' method. The weight distribution is calculated such that channels deemed more useful in the tracking of speech during the task are given more value (or weight) to the reconstruction of the envelope. This reconstructed envelope is then correlated (typically using Pearson's R) with the original speech envelope, with a higher positive correlation relating to higher accuracy of speech envelope tracking in the recorded brain regions.

Both MEG and EEG are sufficient methods of collecting neural data for speech envelope reconstruction (Destoky *et al.,* 2019). However, MEG requires significantly fewer data to train decoding models and correlation estimates of clear speech are higher in MEG than in EEG data. Whilst MEG may be more beneficial, EEG is still adequately robust (Biesmans *et al.*, 2016; Crosse *et al.,* 2016). More work by Mirkovic and colleagues (2015) shows that neural data from EEG can still sufficiently train decoders for reconstructing speech envelopes when the data is reduced to as little as 24 channels, and when the decoder is trained with 15-minutes of the subject-independent neural input. As further support for decoding models with non-invasive recordings, Anumanchipalli *et al.* (2019) found that stimulus reconstruction methods are robust enough to allow for further synthesis of original speech. Using neural data from superior temporal gyrus, inferior frontal gyrus, and ventral sensorimotor regions responsible for lip, mouth, and tongue movements, they synthesised speech from reconstructed envelope and spectrogram information with success, even from silently mimed speech. This highlights the robustness of reconstruction methods as a base for synthesising speech from neural data alone and has been further expanded in recent literature (see: Wairagkar *et al.*, 2023; Luo *et al.*, 2023; Chen *et al.*, 2024).

The beneficial frequency band parameters for a decoder model, as per recommendations Crosse *et al.* (2016), are between 0.5 and 15 Hz, covering delta, theta, and

alpha oscillatory bands. Further narrowing of the frequency range used in the decoder can help to gauge any effects there may be of specific bands on the value of reconstruction accuracy, though using non-optimised parameters for a given neural dataset may result in erroneous inflation of correlational values if representation of the original envelope is misrepresented at the chosen frequencies. There are also methodological choices that differ slightly between studies that may be influencing why some studies report more robust activity in the theta band than others. For example, Bourguignon and colleagues (2020) found that cortical tracking is influenced not only by external speech but also by speech produced by us. Study designs where participants repeat sentences aloud as a measure of intelligibility versus lexical decision tasks may produce different results in neural tracking accuracy as a result. Interestingly, in Bourguignon's study, there was an enhancement of tracking in the theta band when reading aloud versus listening to others. This may suggest that speech prediction and top-down modulation utilising the sensorimotor and speech production regions of the brain play a role in our ability to accurately track the speech envelope.

Whilst understanding speech processing in clear environments is beneficial, it is also important to consider processing when speech is in difficult listening conditions. These situations affect speech processing in our everyday lives, sometimes to detrimental effect (Dubbelboer & Houtgast, 2007; Venetjoki *et al*., 2007). Understanding how speech processing and perception is altered in these conditions, and how both speech intelligibility and tracking accuracy are lowered as a result, may assist in providing insight for ways to improve communication and speech perception. This will be further discussed in the context of multisensory integration and learning.

1.4 Difficult Listening Conditions

Difficult conditions are commonplace in everyday life, such as background noise, multiple speakers talking at once, or the attended speech being quiet. This degradation of the

original speech envelope in these conditions has been evidenced in some studies to lower neural speech-envelope tracking accuracy (An *et al*., 2023). Moreover, Vanthornhout and colleagues (2018) demonstrated a high correlation between measures of accuracy of speech envelope tracking in the auditory cortex and the understanding of speech. Attention and cognitive processes can alleviate this difficulty, by putting more listening effort and cognitive resources into understanding the speech signal (Wingfield, 2016; White & Langdon, 2021). As more noise is introduced, we require more cognitive effort to understand speech. Thus, a decline in our cognitive functioning – such as a natural decline with age – undoubtedly affects our speech perception as well (Gosselin & Gagne, 2011). This was evidenced in ageing studies by Saija and colleagues (2014), who show that older adults have a decline in their ability to understand speech in noise compared to younger adults, but this decline can be restored by slowing down speech or introducing top-down restoration of speech as a form of assisting with the cognitive effort needed to decode it. Investigating how our brains come to understand and process speech on a neural level has become increasingly more beneficial to the development of methods for improving speech comprehension in difficult listening environments.

An interesting distinction in impaired processing is that hearing-impaired listeners demonstrate a reduced ability to use the temporal fine structure to benefit speech understanding in noise (Moon & Hong, 2014), but not a reduction in the ability to use additional speech-envelope information compared to normal-hearing participants (Lorenzi *et al.*, 2006). Given that the neural accuracy of speech-envelope tracking is assumed to be linked to speech intelligibility, such that greater tracking accuracy associates with greater speech intelligibility (Kong *et al*., 2015; Vanthornhout *et al.*, 2018), one would assume that intelligibility and tracking are linked. This intrinsic link is supported by speech-in-noise studies, where decreases in tracking accuracy occur as more noise is added to a speech signal

causing reductions in intelligibility (Vanthornhout *et al.,* 2018; An *et al*., 2023). However, if hearing impairment does not affect envelope-tracking yet does decrease intelligibility in noise, this assumed intrinsic link must be brought into question. Indeed, in normal-hearing populations, difficult listening conditions also diminish speech intelligibility (Dubbelboer & Houtgast, 2007), yet evidence suggests that this is not linked to tracking accuracy (Kösem *et al*., 2023). It may be that tracking accuracy plays an alternative role in speech perception, which is not related to the segmentation of units of speech for better understanding during processing and may instead be related to attentional decoding or speech prediction (see: Geirnaert *et al.*, 2021; Geirnaert *et al*., 2024; Straetmans, 2022).

It is also important to briefly consider attention and its impact on cortical tracking. When attention is involved in speech processing, such as in the common cocktail party effect, we see an increase in early alpha activity between 8 and 12 Hz, indicative of attentional processes, as well as further activity in the theta band (4-8 Hz) when attending to a continuous stream of speech (Kerlin *et al.*, 2010). In this study, sentences were structured in a way that only the last word was changed, such as: ‘His other friend was tired’ and ‘His other friend was fast’. It is then likely that participants attended to the last word only and so may not be an accurate representation of frequency activity for full sentences, but rather the understanding that the last word had changed and the subsequent processing of that word. Brodbeck *et al.* (2018) conducted an MEG study, showing similar findings to early and late responses to attended and unattended speech. Here, later peaks in temporal response functions at around 150 ms reflected attended speech were found. Interestingly, they found that these lexical processing peaks were absent from the unattended speech. This would support the notion that the speech envelope is tracked to help locate and segment speech for processing, but the primary lexical processing is reserved only for the attended stream. As a final point on the effect of correlation, Vanthornhout *et al.* (2019) tested to see if there was a

difference in neural tracking correlations in attentional tasks (where participants were asked to listen to speech whilst being asked questions) versus passive listening tasks (where participants were asked to ignore speech and watch a silent, unrelated movie). They found that in noise conditions with lower signal-to-noise ratios, the neural tracking accuracy was higher in the attention conditions, whereas in clear speech there was little difference. With these findings, they speculated that the passive listening task gave rise to stable, yet low, tracking accuracies. The active attentional task gave rise to higher tracking accuracies, but likely for shorter periods when attention was kept high, thus showing less stable accuracy values across the task.

On the other hand, there is an exception to this seen with healthy ageing, as older adults show better neural tracking of the speech envelope than younger adults (Decruy *et al.*, 2019), with a greater correlation between neural tracking and speech intelligibility. This is hypothesised to be related to inefficient cognitive functioning and an increase in listening and cognitive effort required as we age (Gosselin & Gagne, 2011). This also calls into question the role of neural tracking, as it is unexpected for the tracking accuracy of older adults to be more robust than younger adults, given age-related decline in hearing and speech processing (Lin, 2024). There is the notion that the stimulus intensity used in the experiment may have a large effect on tracking accuracy for methods like stimulus reconstruction. In these cases, differences in intensity used to train a decoder versus reconstruct data on a participant-by-participant basis can cause results to be skewed. Verschueren *et al.* (2021) highlighted this, showing that if the same stimulus intensity is used to train the decoder as is used in experimentation, then the reconstruction accuracy remains robust. Therefore, when working with any age in the population, it is important to match participants based on hearing ability more closely. This is because if one participant has a slightly different level of hearing than another, their perception of the intensity of the stimulus would differ. This intensity problem

can also be avoided by training a separate decoder for each participant using a long speech stimulus – such as a story chapter – before the task and having each reconstruction be based on the participant's trained decoder and own neural testing data. With a younger population, this exception is not currently present in the literature. Enhancing neural speech envelope tracking in a younger population should, therefore, have great benefit to speech recognition, especially in situations where recognition is not at ceiling level, such as in noise. On the other hand, an alternative explanation can be drawn from results by Woodfield and Akeroyd (2010), who found that there was little to no difference between older and younger adults' ability to segment speech. As the neural tracking of older adults retains its robustness with age, it may be that the neural tracking enhancement is not reflecting enhanced speech intelligibility at all. Instead, the assistance with the segmentation of speech may be separate from speech intelligibility performance.

To investigate tracking in younger, normal-hearing populations further, we can look at cases of multisensory integration, such as visual lipreading, where an increase in both intelligibility of speech and cortical speech tracking occurs because of added sensory input relevant to speech heard.

1.5 Multisensory Integration in Speech Perception

Multisensory integration in the brain occurs in many regions, with the auditory system seeing benefit in processing through sensorimotor, visual, and tactile systems. To be discussed are the potential ways in which audio-visual and audio-tactile information can be integrated to improve speech intelligibility and cortical speech-envelope tracking accuracy.

1.5.1 Audio-visual Speech Integration

Audio-visual speech integration typically uses lipreading to provide the auditory cortex with non-auditory speech-relevant cues to assist with speech perception (Sumby &

Pollack, 1954; Bernstein *et al.*, 2004; Maier *et al*., 2011). This is thought to be executed by reading visually distinct movements made by the lips, categorised as visual visemes (Fisher, 1968; Massaro *et al.*, 2012; Bernstein, 2018). These visemes represent a narrow range of phonemes spoken that share the same distinguishable visual features. This allows us not only to enhance our ability to attend to a single stream of speech in a noisy setting – by focusing on the lips of the attended speaker – but to also better understand the speech perceived in noise (Sumby & Pollack, 1954; Schwartz *et al*., 2004). The inverse can also occur, wherein incongruent lip movements influence our ability to discriminate between speech sounds. An example of this is the McGurk Effect (McGurk & MacDonald, 1976), where presenting the speech phoneme 'Ba' with visual lip movements associated with 'Ga' leads to perceptions of the sound 'Da' instead. Neural evidence exists showing the benefit of visual input. Golumbic and colleagues (2013) conducted a MEG study, finding that lipreading enhanced the speech envelope tracking of participants in the auditory cortex. It was greatly beneficial for differentiating between multiple speakers and phase-locking neural activity to the attended speaker alone. This provided an enhancement not only in the accuracy of the cortical tracking to the original speech envelope but also in speech intelligibility. This research has been further backed up by studies utilising neural reconstruction methods of analysis, too (Haider *et al.,* 2024; Crosse *et al.*, 2015). In background noise, visual information is recruited further (Yuan *et al*., 2021). Regions involved in audio-visual integration are widespread across many cortical and subcortical locations, with the most influential regions for speech intelligibility being superior temporal gyrus, right occipital gyrus, and the right thalamus (Gao *et al.*, 2023). Evidence also suggests that both the dorsal and ventral pathways of the visual system are active during audio-visual speech integration (Kaposvári *et al.*, 2015).

When speaking with others, we typically see lip movements before we hear the spoken words (Chandrasekaran *et al*., 2009) as a form of natural stimulus onset asynchrony

(SOA). SOA is when two different modalities of information in cross-modal stimuli are presented at different onsets. The window of integration is the term given to the period in which visual information can lead or lag speech sounds before the visual information is no longer perceived as part of the same stimulus (Stein & Meredith, 1993). If the lip movements are desynchronised from the speech sounds within a specific period, then we still perceive the lip movements and the speech we hear to be congruent. If the SOA is large enough that the auditory and visual information do not fall within the same window of integration, we may perceive the two modalities as separate, and therefore not process the visual information as helpful extra information to discern and comprehend the speech. For speech signals, syllables have a window with an upper limit of about 240 ms and short words of about 300 ms (Navarra *et al*., 2005). Although it is important to note that this window of integration can be highly stimulus-dependent, and ranges in the literature between 150 and 800 ms (Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al.,* 2017), and even differs between age groups (Ren *et al.,* 2017).

Alongside lipreading, there also exists a visual-motor network for integration in speech perception too. Here, the motor system involved with the articulation of our lips is actively involved in the assistance of speech perception (Bruderer *et al*., 2015; Tiippana *et al.*, 2015). This has also been demonstrated with somatosensory stimulation during an audio-visual speech discrimination task, which when congruent improved participant performance and activated parts of the temporal gyrus and the right occipital region (S. Ito *et al.*, 2021). These two networks have been distinguished in fMRI studies (Okada & Hickok, 2009) as utilising different brain regions. Whether the visual-motor network itself also phase-locks to temporal speech envelope information however is not as clear in the literature. As a point of contention, blind individuals show better neural tracking in auditory regions than sighted individuals in a study by Hertrich *et al.* (2013). This was for ultra-fast speech, suggesting

heightened neural sensitivity to speech in the absence of visual information processing. Despite its benefits to speech perception, visual input is not always present. With the increased use of facemask wearing post-COVID-19 pandemic, speech intelligibility has lessened because of an absence of visual input from readable visemes (Brown *et al.*, 2021; Yi *et al.,* 2021; Smiljanic *et al.*, 2021). The impact of COVID-19 may require a reassessment of our understanding of audio-visual integration in speech perception and it's benefits to speech intelligibility, which will be investigated in Chapter 2. In a similar manner, neural speech tracking in audio-visual speech perception will be investigated in Chapter 4 in relation to the potential purpose of tracking in speech prediction. Furthermore, there are many cases where visual information is inaccessible, such as speaking from a distance, through communication devices that utilise auditory streams only like phone calls, and for those with visual impairments due to age, injury, or even blindness.

Investigating other multisensory networks in the brain may provide insight into other beneficial approaches to speech perception in noise. One such sense that will be focused on is the sense of touch through numerous means of tactile stimulation.

1.5.2 Audio-tactile Speech Integration

Whilst visual information can be defined distinctly, tactile information can be provided through several different means, such as vibrations, pressure variations on the skin, spatial differentiation via tactile input moving across different parts of the body, and even non-direct sensations such as air puffs (Gick *et al.*, 2010). Even so, many avenues of tactile input have been observed to assist with speech processing (Rizza *et al.*, 2018; Khoo *et al.,* 2013; Maereg *et al*., 2017; Dementyev *et al*., 2021) using on-body applications. Some posit that this is because such inputs inform our speech perception systems of when to listen rather than provide any distinct information about the speech heard (Tjan *et al.*, 2013). Tjan and colleagues (2013) found that this was the case for both visual and tactile input, suggesting

that audio-tactile and audio-visual integration in speech perception have a similar purpose for assisting with speech processing. In early work, Brooks and Frost (1983) aimed to test if tactile sensations could be used to train participants to learn words. To this end, they developed a tactile vocoder. This vocoder took live-feed audio information from the researcher and produced unique tactile stimulation corresponding to the speech. They then provided tactile stimulation alone and taught participants specific set lists of lexical items through it. Participants managed to learn large chunks of words after extensive training. This shows that tactile information is sufficient for learning speech, so long as extensive training is provided. These vocoders were further applied alongside visual lipreading and found to enhance tactile-visual lipreading ability when participants were trained over time (Bernstein *et al.*, 1991). However, simply learning vocabulary through the set, tactile vocoded signals is not an efficient method of using other senses to assist with our speech perception, as it may not help with understanding speech in noise, nor understanding words that have not yet been learnt. In a similar example, the Tadoma method for deaf-blind individuals also shows how tactile information can be used to represent speech in the absence of visual and auditory streams (Reed *et al*., 1985; Chomsky, 1986). In this method, deaf-blind individuals place their hands on a speaker's face, with their fingers touching the speaker's throat, cheek, lips, and chin. This provides them with a tactile representation of speech through the movement of the lips, vocal cords, and even from the expelling of air. This method requires extensive training, however, and only really sees benefit in deaf-blind populations. A similar, adapted version of the Tadoma method was used by Sato *et al.* (2010) to further investigate the use of touch with sound. Their findings showed that audio-tactile integration occurred when participants placed their hands on the mouth and face of a speaker whilst listening to speech both embedded in noise and not. When the talker's mouth movements did not match the speech spoken, intelligibility was reduced again, suggesting that the tactile information played a role in the

interpretation of speech in noise. This is an important step in showing the benefit of the tactile sense in speech perception. Furthermore, the study showed that this effect was beneficial for both blind participants and sighted individuals. However, as discussed before, this is one of many ways the tactile system could potentially be utilised with speech. The need to feel the mouth of the speaker may not be a suitable method of acquiring this information in everyday life and represents a motion more akin to the Tadoma method of multisensory speech. Other methods of tactile transcription, such as vibrational or spatial across the skin, may be more suitable.

One recent method of alleviating problems with noise in cochlear implants has been the use of a form of tactile stimulation called electro-haptic stimulation (EHS). EHS techniques have been investigated to improve speech in noise processing (Fletcher *et al.*, 2019; Fletcher *et al.*, 2020; Fletcher, 2021), wherein vibrations that correspond to electrical frequencies of the cochlear implant have been used to provide missing auditory information to the skin. This has been shown to work well with short training programmes, improving speech recognition in multi-talker scenarios (Fletcher *et al.*, 2018) and spatial awareness of speakers (Fletcher *et al.*, 2020). More importantly, it has demonstrated that tactile stimulation can carry speech signal information, and this can be utilised by participants after short-training regimes. It is, however, only in the context of electro-haptic stimulation for aiding cochlear implant usage, which is a different hearing environment to non-implant users.

Another key aspect of current multisensory integration investigations is the length of the window of integration. In audio-visual speech, visual information of the lips moving can follow the speech spoken by 250 ms and precede the speech by approximately 70 ms before it no longer provides any useful benefit to speech comprehension, though this window is highly variable and stimulus dependent (Navarra *et al*., 2005; Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al.,* 2017). Gick *et al.* (2010) determined a temporal window of

audio-tactile integration in speech perception to be up to 200 ms of asynchrony for the air puff to follow the audio and 50 ms for preceding the audio. Whilst a narrower window of integration than seen in the audio-visual stream, this shows that, to some capacity, there is an inherent ability for the brain to utilise tactile streams of information for speech integration. Moreover, in a study by Riecke and colleagues (2019), speech envelopes were extracted from sentences as a form of degradation to speech stimuli. The remaining fine structure was presented to participants. As expected, cortical tracking diminished with the removal of speech envelope information. However, when this information was provided to the participant in the form of tactile stimulation to the fingertips, there was an enhancement of speech envelope-tracking seen. This follows similar audio-visual envelope tracking trends in previous literature and highlights an avenue for the use of tactile sensations to improve speech processing in the absence of – or in combination with – visual lipreading.

Interestingly, unlike evidence from audio-visual integration, Riecke and colleagues did not identify any increase in speech recognition or intelligibility in audio-tactile trials, despite a significant increase in cortical tracking being present. This goes against previous notions that speech envelope tracking is critical for enhancing our understanding of speech heard. There are two accounts for this pointed out in the paper. Firstly, that tactile stimulation simply is not a sufficient sense for providing speech information. Secondly, that training or exposure to tactile sensations that relate directly to speech envelopes is needed to see improvements in speech recognition. This second comment is of particular interest. In the case of audio-visual input, most people have been exposed to lipreading and viseme information for speech for most of their lives and arguably were naturally trained with such stimuli from very early childhood. There is evidence that infants as young as 5 months old demonstrate a neural capability to track speech envelopes, and that this is enhanced with congruent visual lipreading (Tan *et al*., 2022). Again, further evidence of audio-visual

integration of speech being present from a young age, essentially trained and exposed to throughout most of our lives.

It may be that short-term training with tactile stimulations will provide the missing link of intelligibility increase alongside further enhancements to cortical speech envelope tracking. Though, on the other hand, it may be more complex than simple exposure or training, as highlighted by studies showing that visual-tactile input improves speech with no previous exposure (Fowler & Dekle, 1991). Turning to recent work investigating cortical speech tracking in multiple languages, Reetzke *et al.* (2021) found that non-native speakers demonstrated a level of significant neural tracking of English speech. Native English speakers, however, demonstrated higher tracking than non-native speakers. This could reflect limited tracking to lower levels of speech processing. For example, there could be elements of phoneme or syllabic tracking with shared sounds of the native and foreign language that maintain high correlations, whereas word and sentence segmentation are lost in non-native speakers and not robustly tracked, lowering the overall correlation between cortical activity and original tracked speech. At the very least, without some semblance of training or exposure to aspects of speech, low-level neural tracking correlates could indicate their purpose relating to attention to speech as opposed to lexical processing.

Training to audio-tactile speech-in-noise may be key to providing tactile benefit to speech intelligibility. The potential benefits of short-term training on audio-tactile speech will be investigated in Chapter 3.

1.6 Perceptual Training in Speech-in-noise

Before making assumptions and hypotheses based on training with tactile information to improve speech tracking and intelligibility, it is important to review findings from perceptual training research. Perceptual training has long been studied to better understand

adaptive brain plasticity and methods by which we can improve our perception of our environment, and the respective systems used to decode it (Rosenzweig & Bennett, 1996; Ahissar, 2001). In particular, the benefits of training with speech-in-noise both behavioural and neural are of relevance. Song and colleagues (2011) investigated speech-in-noise training with naturalistic, multi-talker babble scenarios. Here, participants were either trained over 20 short sessions or simply tested regularly in their ability to discriminate speech from noise. Before and after training, baseline measures were taken, and brainstem responses were recorded during the task. They found a significant increase in speech discrimination performance in trained versus non-trained participants. Impressively, these increases from training were retained for six months, highlighting the effectiveness of simple, short-term auditory training regimes. Moreover, investigations using ERPs have narrowed the length of training to as low as two days, with 12 hours between sessions for consolidation (Atienza *et al.*, 2002). This consolidation likely plays an important role in learning in speech processing, stabilising the benefits from training at a faster rate than when sleep consolidation is not accessible post-training (Drouin *et al.*, 2023). Speculatively, this may be in part related to the oscillatory delta activity which furthers speech intelligibility through tracking of the speech-envelope in the auditory cortex (Etard & Reichenbach, 2019; Ding & Simon, 2014) and plays a role in deep-sleep brain states (Nir *et al.*, 2011; Lechat *et al.*, 2022). Perceptual training has seen other forms of success in the auditory domain too, such as for the improvement of speech and cognitive processes in children with auditory disorders (Kumar *et al.*, 2021), and with ADHD (Mishra *et al*., 2016). Though, on the other hand, it is important to note that there is disagreement in the perceptual learning literature regarding seen training effects, with some believing that trained performance in perceptual learning research does not adequately evidence far transfer (applied training benefits to a dissimilar task) and thus should be taken

with care (Don *et al.*, 2023; Fransen, 2024), whilst others claim evidence of far transfer (Gao *et al.*, 2020).

Further, in the context of multisensory perceptual training, it has been observed that short-term audio-visual training has improvements to speech discrimination over training with auditory stimuli only, with retention of these benefits for at least one month (Moradi *et al.,* 2017). The benefits from this multimodal training group were seen in just a single session (though the length of this training session is undefined), suggesting that, whilst multiple short sessions are required for single-modality training procedures, the added benefit of the visual medium through lipreading was great enough to reduce training times drastically and promote improvement just after a day. This rapid learning would align with research showing that the lateral geniculate nucleus, a thalamic region likely crucial for visual learning (Yu *et al.*, 2016; Weyand, 2015), engages in rapid plasticity (Moore *et al.* 2011). Together, alongside the medial geniculate body of the thalamic centre is being utilised for auditory learning, this repeated projection through the thalamic pathways for integration may be enhancing perceptual learning at a greater rate than when learning is unimodal. In further speculation, it may also be possible to use other forms of multisensory integration for auditory speech perception in training programmes given this benefit, such as tactile speech envelope stimulation. There has been research investigating tactile training with pure tone sounds for profoundly deaf individuals (Gonzalez-Garrido *et al*., 2017). Whilst only using simplistic pure-tone sounds, this still highlights the brain's ability to adapt to relevant tactile input and use it accordingly. There is the concern that this training is with too simplistic a set of stimuli. For example, being able to discriminate pure tones using tactile stimulation might not reflect your ability to discriminate more complex speech or even just another frequency of a pure tone. This concern, however, may not be valid, as recent research has found that training speech-in-noise benefits speech perception regardless of task specificity (Gao *et al*. 2020).

Therefore, it can be entirely reasonable to infer from perceptual training studies that specific, enhanced speech discrimination ability seen in trained groups may transfer to other speech discrimination contexts.

1.6.1 Bottom-up vs Top-down Training

Whilst investigating general perceptual learning from a unimodal and multimodal perspective is useful for understanding if tactile training may be beneficial, it is also important to reflect upon different types of training that utilise both bottom-up and top-down processes. It is known that speech perception involves a combination of dynamic bottom-up and top-down processing (Zekveld *et al.*, 2006; Diekhof *et al*., 2009). As discussed, there are many ascending and descending pathways in the auditory processing stream, with sensory processing primarily taking place along the ascending route through the cochlear nucleus, inferior and superior colliculi, medial geniculate body of the thalamus, and through to the primary and nonprimary regions of the auditory cortex (Plack, 2018). Top-down processing, on the other hand, is more complex, with descending corticofugal projections from the auditory cortex to every lower subcortical region, including corticothalamic and corticollicular (Asilador & Llano, 2021; Souffi *et al.*, 2021; Oberle *et al.*, 2022; Ford *et al.*, 2024). Whilst training may be possible on a sensory level with audio-tactile speech, it might be more effective to present training that utilises both top-down and bottom-up processes to capture the wider speech processing network in the brain. Indeed, both bottom-up and top-down focused training paradigms can successfully lead to improvements in speech intelligibility (Gohari *et al.*, 2023). As detailed by Gohari *et al.* (2023), a top-down training paradigm might include memory-based training (Ingvalson *et al.*, 2015; Schneiders *et al.*, 2012) or speech-in-noise training (Fletcher *et al.*, 2020; Ciesla *et al*., 2022; O'Hanlon *et al.*, in prep., 2025) to make use of top-down selective attention and contextual speech cues. A bottom-up training paradigm on the other hand might include temporal integration (Zerr *et*

al., 2019) or phonemic training (Schumann *et al*., 2015) to make use of sensory speech processing alone. This form of training would likely rely on the auditory system's natural or pre-trained ability to integrate multimodal speech-relevant information, as top-down modulation would not be a focus of the paradigm to learn integration beyond the training stimulus set. As audio-visual integration is developed at an early age (Soto-Faraco *et al*., 2012), audio-visual integration systems may be sufficiently developed for successful bottom-up training, whereas audio-tactile training would likely require further learning input. This discrepancy between bottom-up and top-down effectiveness for audio-tactile training will be explored in Chapter 4.

1.7 Summary and Thesis Outline

In summary, the auditory system processes complex speech information through a variety of afferent and efferent pathways. Neural speech tracking in the brain can provide a means of understanding how components of speech, such as the speech envelope, are represented in the brain during processing. With difficult listening conditions, speech become more difficult to understand, with both neural tracking accuracy and intelligibility decreasing. Multisensory integration can assist with speech perception in these difficult listening conditions, though may need facilitation with training to encourage benefit, like with audio-tactile integration.

This thesis aims to further our understanding of multisensory integration in speech perception, and how audio-visual and audio-tactile speech can enhance neural speech tracking and speech intelligibility. The first aim is to reassess the benefits of audio-visual speech. This will be investigated in Chapter 2, which examines the behavioural benefits of audio-visual speech when stimuli are selected from different viseme categories. Chapter 5 will also provide further audio-visual speech understanding from a neural perspective. The second aim of this thesis is to investigate if short-term training provides audio-tactile benefit

to speech intelligibility, as well as further enhancements in neural tracking accuracy. This will be investigated by both Chapters 3 and 4. Further to this, these chapters will also contribute to the aim of understanding the differences in applying top-down versus bottom-up training for audio-tactile speech processing. Finally, this thesis aims to further our understanding of the role of neural tracking accuracy in speech processing. Chapter 5 will primarily investigate this through a secondary data analysis of intracranial electrocorticography data, examining the potential role of tracking in the prediction of oncoming speech through audio-visual integration. To conclude, Chapter 6 will provide general discussion to this thesis, along with speculation for future research and final conclusions.

References

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience*, *28*(15), 3958-3965.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367-13372.

Ahissar, M. (2001). Perceptual training: a tool for both modifying the brain and exploring it. *Proceedings of the National Academy of Sciences, 98*(21), 11842-11843.

Ahveninen, J., Jääskeläinen, I. P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., ... & Belliveau, J. W. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. *Proceedings of the National Academy of Sciences, 103*(39), 14608-14613.

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*(2), 139-157.

An, H., Lee, J., Suh, M. W., & Lim, Y. (2023). Neural correlation of speech envelope tracking for background noise in normal hearing. *Frontiers in Neuroscience, 17*, 1268591.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493-498.

Apoux, F., & Healy, E. W. (2013). *A glimpsing account of the role of temporal fine structure information in speech recognition*. In Basic Aspects of Hearing (pp. 119-126). *Springer, New York, NY*.

Asilador, A., & Llano, D. A. (2021). Top-down inference in the auditory system: potential roles for corticofugal projections. *Frontiers in Neural Circuits, 14*, 615259.

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory, 9*(3), 138-150.

Ball, T., Kern, M., Mutschler, I., Aertsen, A., & Schulze-Bonhage, A. (2009). Signal quality of simultaneously recorded invasive and non-invasive EEG. *NeuroImage, 46*(3), 708-716.

Balmer, T. S., & Trussell, L. O. (2021). Trigeminal contributions to the dorsal cochlear nucleus in mouse. *Frontiers in Neuroscience, 15*, 715954.

Balmer, T. S., & Trussell, L. O. (2022). Descending axonal projections from the inferior colliculus target nearly all excitatory and inhibitory cell types of the dorsal cochlear nucleus. *Journal of Neuroscience, 42*(16), 3381-3393.

Barajas, M. C. O., Guevara, R., & Gervain, J. (2021). The origins and development of speech envelope tracking during the first months of life. *Developmental Cognitive Neuroscience*, *48*, 100915.

Bartlett, E. L. (2012). the Medial geniculate body. *Translational Perspectives in Auditory Neuroscience: Normal Aspects of Hearing, 1*, 207.

Bean, N. L., Smyre, S. A., Stein, B. E., & Rowland, B. A. (2023). Noise-rearing precludes the behavioral benefits of multisensory integration. *Cerebral Cortex, 33*(4), 948-958.

Becker, R., Van De Ville, D., & Kleinschmidt, A. (2018). Alpha oscillations reduce temporal long-range dependence in spontaneous human brain activity. *Journal of Neuroscience, 38*(3), 755-764.

Bernstein, L. E. (2018). Response errors in females' and males' sentence lipreading necessitate structurally different models for predicting lipreading accuracy. *Language Learning, 68*, 127-158.

Bernstein, L. E., Auer Jr, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1-4), 5-18.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, *90*(6), 2971-2984.

Beste, C., Münchau, A., & Frings, C. (2023). Towards a systematization of brain oscillatory activity in actions. *Communications Biology, 6*(1), 137.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25*(5), 402-412.

Borjigan, A., & Bharadwaj, H. M. (2019). Investigating the role of temporal fine structure in everyday hearing. *The Journal of the Acoustical Society of America, 145*(3), 1872-1873.

Boucher, V. J., Gilbert, A. C., & Jemel, B. (2019). The role of low-frequency neural oscillations in speech processing: revisiting delta entrainment. *Journal of Cognitive Neuroscience, 31*(8), 1205-1215.

Bourguignon, M., Molinaro, N., Lizarazu, M., Taulu, S., Jousmäki, V., Lallier, M., ... & De Tiege, X. (2020). Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud. *NeuroImage*, *216*, 116788.

Braun, M. (2000). Inferior colliculus as candidate for pitch extraction: multiple support from statistics of bilateral spontaneous otoacoustic emissions. *Hearing Research, 145*(1-2), 130-140.

Brewer, A. A., & Barton, B. (2016). Maps of the auditory cortex. *Annual Review of Neuroscience, 39*(1), 385-407.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Transformation from auditory to linguistic representations across auditory cortex is rapid and attention dependent for continuous speech. *BioRxiv*, 326785.

Brooks, P. L., & Frost, B. J. (1983). Evaluation of a tactile vocoder for word recognition. *The Journal of the Acoustical Society of America*, *74*(1), 34-39.

Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications, 6*(1), 49.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences, 112*(44), 13531-13536.

Cabral, J., Castaldo, F., Vohryzek, J., Litvak, V., Bick, C., Lambiotte, R., ... & Deco, G. (2022). Metastable oscillatory modes emerge from synchronization in the brain spacetime connectome. *Communications Physics, 5*(1), 184.

Capilla, A., Arana, L., García-Huéscar, M., Melcón, M., Gross, J., & Campo, P. (2022). The natural frequencies of the resting human brain: An MEG-based atlas. *NeuroImage, 258*, 119373.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436.

Chen, X., Wang, R., Khalilian-Gourtani, A., Yu, L., Dugan, P., Friedman, D., ... & Flinker, A. (2024). A neural speech decoding framework leveraging deep learning and speech synthesis. *Nat Mach Intell, 6*, 467–480.

Chomsky, C. (1986). Analytic study of the Tadoma method: Language abilities of three deaf-blind subjects. *Journal of Speech, Language, and Hearing Research, 29*(3), 332-347.

Cieśla, K., Wolak, T., Lorens, A., Mentzel, M., Skarżyński, H., & Amedi, A. (2022). Effects of training and using an audio-tactile sensory substitution device on speech-in-noise understanding. *Scientific Reports, 12*(1), 3206.

Cole, S. R., & Voytek, B. (2017). Brain oscillations and the importance of waveform shape. *Trends in Cognitive Sciences, 21*(2), 137-149.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in Integrative Neuroscience, 4*, 1316.

Crone, N. E., Korzeniewska, A., & Franaszczuk, P. J. (2011). Cortical gamma responses: searching high and low. *International Journal of Psychophysiology, 79*(1), 9-15.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195-14204.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Daly, H. R., & Pitt, M. A. (2021). Distractor probability influences suppression in auditory selective attention. *Cognition, 216*, 104849.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience, 37*(27), 6539-6557.

Decruy, L., Vanthornhout, J., & Francart, T. (2019). Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of Neurophysiology*, *122*(2), 601-615.

Dementyev, A., Getreuer, P., Kanevsky, D., Slaney, M., & Lyon, R. F. (2021, October). VHP: Vibrotactile Haptics Platform for On-body Applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (pp. 598-612).

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., ... & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage*, *184*, 201-213.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology, 25*(19), 2457-2465.

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping, 30*(11), 3509-3526.

Diekhof, E. K., Biedermann, F., Ruebsamen, R., & Gruber, O. (2009). Top-down and bottom-up modulation of brain structures involved in auditory discrimination. *Brain Research, 1297*, 118-123.

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2017). Cortical alpha oscillations predict speech intelligibility. *Frontiers in Human Neuroscience*, *11*, 88.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728-5735.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience, 8*, 311.

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage, 88*, 41-46.

Don, H. J., Goldwater, M. B., & Livesey, E. J. (2023). Cognition of relational discovery: why it matters for effective far transfer and effective education?. *Frontiers in Psychology, 14*, 957517.

Driscoll, M. E., & Tadi, P. (2020). *Neuroanatomy, inferior colliculus*. StatPearls Publishing, Treasure Island (FL).

Drouin, J. R., Zysk, V. A., Myers, E. B., & Theodore, R. M. (2023). Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech. *Journal of Speech, Language, and Hearing Research, 66*(2), 720-734.

Dubbelboer, F., & Houtgast, T. (2007). A detailed study on the effects of noise on speech intelligibility. *The Journal of the Acoustical Society of America, 122*(5), 2865-2871.

Esghaei, M., Treue, S., & Vidyasagar, T. R. (2022). Dynamic coupling of oscillatory neural activity and its roles in visual attention. *Trends in Neurosciences, 45*(4), 323-335.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience, 39*(29), 5750-5759.

Farhadi, A. (2023). *Modeling the Medial Olivocochlear Efferent in the Descending Auditory Pathway With a Dynamic Gain Control Feedback System*. University of Rochester.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*(4), 796-804.

Fletcher, M. D., & Verschuur, C. A. (2021). Electro-haptic stimulation: A new approach for improving cochlear-implant listening. *Frontiers in Neuroscience, 15*, 581414.

Fletcher, M. D., Hadeedi, A., Goehring, T., & Mills, S. R. (2019). Electro-haptic enhancement of speech-in-noise performance in cochlear implant users. *Scientific Reports, 9*(1), 11428.

Fletcher, M. D., Mills, S. R., & Goehring, T. (2018). Vibro-tactile enhancement of speech intelligibility in multi-talker noise for simulated cochlear implant listening. *Trends in Hearing, 22*, 2331216518797838.

Fletcher, M. D., Song, H., & Perry, S. W. (2020). Electro-haptic stimulation enhances speech recognition in spatially separated noise for cochlear implant users. *Scientific Reports, 10*(1), 12723.

Ford, A. N., Czarny, J. E., Rogalla, M. M., Quass, G. L., & Apostolides, P. F. (2024). Auditory corticofugal neurons transmit auditory and non-auditory information during behavior. *Journal of Neuroscience, 44*(7).

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 816.

Fransen, J. (2024). There is No Supporting Evidence for a Far Transfer of General Perceptual or Cognitive Training to Sports Performance. *Sports Medicine*, 1-8.

Gafoor, S. A., & Uppunda, A. K. (2024). Role of the medial olivocochlear efferent auditory system in speech perception in noise: a systematic review and meta-analyses. *International Journal of Audiology, 63*(8), 561-569.

Gao, C., Green, J. J., Yang, X., Oh, S., Kim, J., & Shinkareva, S. V. (2023). Audiovisual integration in the human brain: a coordinate-based meta-analysis. *Cerebral Cortex, 33*(9), 5574-5584.

Gao, X., Yan, T., Huang, T., Li, X., & Zhang, Y. X. (2020). Speech in noise perception improved by training fine auditory discrimination: far and applicable transfer of perceptual learning. *Scientific Reports, 10*(1), 1-12.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B. T., ... & Bertrand, A. (2021). EEG-based auditory attention decoding: Towards neuro-steered hearing devices. *arXiv preprint*. Accessed at: https://arxiv.org/abs/2008.04569

Geirnaert, S., Zink, R., Francart, T., & Bertrand, A. (2024). Fast, accurate, unsupervised, and time-adaptive EEG-based auditory attention decoding for neuro-steered hearing

devices. *In Brain-Computer Interface Research: A State-of-the-Art Summary 11* (pp. 29-40). Cham: Springer Nature Switzerland.

Gick, B., Ikegami, Y., & Derrick, D. (2010). The temporal window of audio-tactile integration in speech perception. *The Journal of the Acoustical Society of America*, *128*(5), EL342-EL346.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511-517.

Gohari, N., Dastgerdi, Z. H., Rouhbakhsh, N., Afshar, S., & Mobini, R. (2023). Training programs for improving speech perception in noise: A review. *Journal of Audiology & Otology, 27*(1), 1.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. *Journal of Neuroscience*, *33*(4), 1417-1426.

González-Garrido, A. A., Ruiz-Stovel, V. D., Gómez-Velázquez, F. R., Vélez-Pérez, H., Romo-Vázquez, R., Salido-Ruiz, R. A., ... & Campos, L. R. (2017). Vibrotactile discrimination training affects brain connectivity in profoundly deaf individuals. *Frontiers in Human Neuroscience, 11*, 28.

Gosselin, P. A., & Gagne, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944-958.

Gourévitch, B., Martin, C., Postal, O., & Eggermont, J. J. (2020). Oscillations in the auditory system and their possible role. *Neuroscience & Biobehavioral Reviews, 113*, 507-528.

Guan, S., Jiang, R., Chen, D. Y., Michael, A., Meng, C., & Biswal, B. (2023). Multifractal long-range dependence pattern of functional magnetic resonance imaging in the human brain at rest. *Cerebral Cortex, 33*(24), 11594-11608.

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. *Journal of Cognitive Neuroscience, 36*(1), 128-142.

Hall, D. A., Hart, H. C., & Johnsrude, I. S. (2003). Relationships between human auditory cortical structure and function. *Audiology and Neurotology, 8*(1), 1-18.

Hall, D., & Barker, D. (2012). Coding of basic acoustical and perceptual components of sound in human auditory cortex. In*: The human auditory cortex* (pp. 165-197). New York, NY: Springer New York.

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell, 184*(18), 4626-4639.

Heil, P., & Peterson, A. J. (2015). Basic response properties of auditory nerve fibers: a review. *Cell and Tissue Research*, *361*(1), 129-158.

Heinz, M. G., & Swaminathan, J. (2009). Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *Journal of the Association for Research in Otolaryngology*, *10*(3), 407-423.

Herdman, A. T., Lins, O., Van Roon, P., Stapells, D. R., Scherg, M., & Picton, T. W. (2002). Intracerebral sources of human auditory steady-state responses. *Brain Topography*, *15*(2), 69-86.

Hertrich, I., Dietrich, S., & Ackermann, H. (2013). Tracking the speech signal–Time-locked MEG signals during perception of ultra-fast and moderately fast speech in blind and in sighted listeners. *Brain and Language*, *124*(1), 9-21.

Hjortkjær, J., Märcher-Rørsted, J., Fuglsang, S. A., & Dau, T. (2020). Cortical oscillations and entrainment in speech processing during working memory load. *European Journal of Neuroscience, 51*(5), 1279-1289.

Hockley, A., Wu, C., & Shore, S. E. (2022). Olivocochlear projections contribute to superior intensity coding in cochlear nucleus small cells. *The Journal of Physiology, 600*(1), 61-73.

Hopkins, K., & Moore, B. C. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America, 127*(3), 1595-1608.

Hu, F., & Dan, Y. (2022). An inferior-superior colliculus circuit controls auditory cue-directed visual spatial attention. *Neuron, 110*(1), 109-119.

Ingvalson, E. M., Dhar, S., Wong, P., & Liu, H. (2015). Working memory training to improve speech perception in noise across languages. *The Journal of the Acoustical Society of America, 137*(6), 3477-3486.

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., & Lizarazu, M. (2024). On the Speech Envelope in the Cortical Tracking of Speech. *NeuroImage*, 120675.

Ito, S., Si, Y., Litke, A. M., & Feldheim, D. A. (2021). Nonlinear visuoauditory integration in the mouse superior colliculus. *PLoS Computational Biology, 17*(11), e1009181.

Ito, T., Ohashi, H., & Gracco, V. L. (2021). Somatosensory contribution to audio-visual speech processing. *Cortex, 143*, 195-204.

Jedrzejczak, W. W., Milner, R., Ganc, M., Pilka, E., & Skarzynski, H. (2020). No change in medial olivocochlear efferent activity during an auditory or visual task: Dual evidence from otoacoustic emissions and event-related potentials. *Brain Sciences, 10*(11), 894.

Jennings, S. G. (2021). The role of the medial olivocochlear reflex in psychophysical masking and intensity resolution in humans: a review. *Journal of Neurophysiology, 125*(6), 2279-2308.

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in Human Neuroscience, 4*, 186.

Jensen, O., Spaak, E., & Zumer, J. M. (2019). Human brain oscillations: From physiological mechanisms to analysis and cognition. *Magnetoencephalography: From Signals to Dynamic Cortical Networks*, 471-517.

Jessen, S., Obleser, J., & Tune, S. (2021). Neural tracking in infants–An analytical tool for multisensory social processing in development. *Developmental Cognitive Neuroscience, 52*, 101034.

Kaiser, J., & Lutzenberger, W. (2005). Cortical oscillatory activity and the dynamics of auditory memory processing. *Reviews in the Neurosciences, 16*(3), 239-254.

Kanth, S. T., & Ray, S. (2020). Electrocorticogram (ECoG) is highly informative in primate visual cortex. *Journal of Neuroscience, 40*(12), 2430-2444.

Kaposvári, P., Csete, G., Bognár, A., Csibri, P., Tóth, E., Szabó, N., ... & Kincses, Z. T. (2015). Audio–visual integration through the parallel visual pathways. *Brain Research, 1624*, 71-77.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *Journal of Neuroscience*, *30*(2), 620-628.

Khoo, W. L., Knapp, J., Palmer, F., Ro, T., & Zhu, Z. (2013). Designing and testing wearable range-vibrotactile devices. *Journal of Assistive Technologies*.

Kikuchi, S., Ishizaka, Y., Otsuka, S., & Nakagawa, S. (2023, November). Effects of orienting attention to a specific frequency on medial olivocochlear reflex-A study of dependence on target frequencies. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 268*(2), 5996-6000.

King, A. J., & Schnupp, J. W. (2007). The auditory cortex. *Current Biology, 17*(7), R236-R239.

King, A. J., Teki, S., & Willmore, B. D. (2018). Recent advances in understanding the auditory cortex. *F1000Research, 7*.

Kong, Y. Y., Somarowthu, A., & Ding, N. (2015). Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. *Journal of the Association for Research in Otolaryngology, 16*, 783-796.

Kösem, A., Dai, B., McQueen, J. M., & Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage, 272*, 120040.

Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A., & Hari, R. (2020). Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage, 219*, 116936.

Kraus, N. (2011). Listening in on the listening brain. *Physics Today, 64*(6), 40-45.

Krizman, J., & Kraus, N. (2019). Analyzing the FFR: A tutorial for decoding the richness of auditory function. *Hearing Research, 382*, 107779.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PLoS One*, *8*(1), e53398.

Kuenzel, T. (2019). Modulatory influences on time-coding neurons in the ventral cochlear nucleus. *Hearing Research, 384*, 107824.

Kumar, P., Singh, N. K., & Hussain, R. O. (2021). Effect of speech in noise training in the auditory and cognitive skills in children with auditory processing disorders. International *Journal of Pediatric Otorhinolaryngology, 146*, 110735.

Kunchur, M. N. (2023). The human auditory system and audio. *Applied Acoustics, 211*, 109507.

Kurthen, I., Galbier, J., Jagoda, L., Neuschwander, P., Giroud, N., & Meyer, M. (2021). Selective attention modulates neural envelope tracking of informationally masked speech in healthy older adults. *Human Brain Mapping, 42*(10), 3042-3057.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology, 94*(3), 1904-1911.

Lauer, A. M., Jimenez, S. V., & Delano, P. H. (2022). Olivocochlear efferent effects on perception and behavior. *Hearing Research, 419*, 108207.

Lechat, B., Hansen, K. L., Melaku, Y. A., Vakulin, A., Micic, G., Adams, R. J., ... & Zajamsek, B. (2022). A novel electroencephalogram-derived measure of disrupted delta wave activity during sleep predicts all-cause mortality risk. *Annals of the American Thoracic Society, 19*(4), 649-658.

Lin, F. R. (2024). Age-related hearing loss. *New England Journal of Medicine, 390*(16), 1505-1512.

Litovsky, R. Y., Fligor, B. J., & Tramo, M. J. (2002). Functional role of the human inferior colliculus in binaural hearing. *Hearing Research, 165*(1-2), 177-188.

Liu, M., Dai, J., Zhou, M., Liu, J., Ge, X., Wang, N., & Zhang, J. (2022). Mini-review: The neural circuits of the non-lemniscal inferior colliculus. *Neuroscience Letters, 776*, 136567.

Liu, X., Huang, H., Snutch, T. P., Cao, P., Wang, L., & Wang, F. (2022). The superior colliculus: cell types, connectivity, and behavior. *Neuroscience Bulletin, 38*(12), 1519-1540.

Lizarazu, M., Lallier, M., & Molinaro, N. (2019). Phase− amplitude coupling between theta and gamma oscillations adapts to speech rate. *Annals of the New York Academy of Sciences, 1453*(1), 140-152.

Lombardi, F., Herrmann, H. J., Parrino, L., Plenz, D., Scarpetta, S., Vaudano, A. E., ... & Shriki, O. (2023). Beyond pulsed inhibition: Alpha oscillations modulate attenuation and amplification of neural activity in the awake resting state. *Cell Reports, 42*(10).

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, 103*(49), 18866-18869.

Luo, B., Li, J., Liu, J., Li, F., Gu, M., Xiao, H., ... & Xiao, Z. (2022). Frequency-dependent plasticity in the temporal association cortex originates from the primary auditory cortex, and is modified by the secondary auditory cortex and the medial geniculate body. *Journal of Neuroscience, 42*(26), 5254-5267.

Luo, S., Rabbani, Q., & Crone, N. E. (2023). Brain-computer interface: applications to speech decoding and synthesis to augment communication. *Neurotherapeutics, 19*(1), 263-273.

Maereg, A. T., Nagar, A., Reid, D., & Secco, E. L. (2017). Wearable vibrotactile haptic device for stiffness discrimination during virtual interactions. *Frontiers in Robotics and AI*, *4*, 42.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 245.

Marco-Pallarés, J., Münte, T. F., & Rodríguez-Fornells, A. (2015). The role of high-frequency oscillatory activity in reward processing and learning. *Neuroscience & Biobehavioral Reviews, 49*, 1-7.

Massaro, D. W., Cohen, M. M., Tabain, M., & Beskow, J. (2012). Animated speech: Research progress and applications In Clark RB, Perrier J, P, & Vatikiotis-Bateson E (Eds.), Audiovisual Speech Processing (pp. 246–272). *Cambridge: Cambridge University*.

May, B. J. (2000). Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hearing Research, 148*(1-2), 74-87.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

McHaney, J. R., Gnanateja, G. N., Smayda, K. E., Zinszer, B. D., & Chandrasekaran, B. (2021). Cortical tracking of speech in delta band relates to individual differences in speech in noise comprehension in older adults. *Ear and Hearing, 42*(2), 343-354.

Meng, Q., & Schneider, K. A. (2022). A specialized channel for encoding auditory transients in the magnocellular division of the human medial geniculate nucleus. *Neuroreport, 33*(15), 663-668.

Meng, Q., & Schneider, K. A. (2022). The magnocellular division of the human medial geniculate nucleus preferentially responds to auditory transients. *bioRxiv*, 2022-06.

Menn, K. H., Michel, C., Meyer, L., Hoehl, S., & Männel, C. (2022). Natural infant-directed speech facilitates neural tracking of prosody. *NeuroImage, 251*, 118991.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609-2621.

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, *12*(4), 046007.

Mishra, J., Sagar, R., Joseph, A. A., Gazzaley, A., & Merzenich, M. M. (2016). Training sensory signal-to-noise resolution in children with ADHD in a global mental health setting. *Translational Psychiatry, 6*(4), e781-e781.

Moerel, M., De Martino, F., & Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience, 8*, 225.

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important?. *Korean Journal of Audiology, 18*(1), 1.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology, 9*(4), 399-406.

Moore, B. D., Kiley, C. W., Sun, C., & Usrey, W. M. (2011). Rapid plasticity of visual responses in the adult lateral geniculate nucleus. *Neuron, 71*(5), 812-819.

Moradi, S., Wahlin, A., Hällgren, M., Rönnberg, J., & Lidestam, B. (2017). The efficacy of short-term gated audiovisual speech training for improving auditory sentence identification in noise in elderly hearing aid users. *Frontiers in Psychology, 8*, 368.

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience & Biobehavioral Reviews, 107*, 136-142.

Morillon, B., Hackett, T. A., Kajikawa, Y., & Schroeder, C. E. (2015). Predictive motor control of sensory dynamics in auditory active sensing. *Current Opinion in Neurobiology, 31*, 230-238.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507.

Nir, Y., Staba, R. J., Andrillon, T., Vyazovskiy, V. V., Cirelli, C., Fried, I., & Tononi, G. (2011). Regional slow waves and spindles in human sleep. *Neuron, 70*(1), 153-169.

Oberle, H. M., Ford, A. N., Dileepkumar, D., Czarny, J., & Apostolides, P. F. (2022). Synaptic mechanisms of top-down control in the non-lemniscal inferior colliculus. *Elife, 10*, e72730.

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural oscillations in speech: don't be enslaved by the envelope. *Frontiers in Human Neuroscience*, *6*, 250.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience, 5*(11), 1203-1209.

Ojima, H., & Rouiller, E. M. (2010). Auditory cortical projections to the medial geniculate body. In *The auditory cortex* (pp. 171-188). Boston, MA: Springer US.

Okada, K., & Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data. *Neuroscience Letters*, *452*(3), 219-223.

Palmer, A. R., & Kuwada, S. (2005). Binaural and spatial coding in the inferior colliculus. In *The inferior colliculus* (pp. 377-410). New York, NY: Springer New York.

Palmer, A. R., & Russell, I. J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research, 24*(1), 1-15.

Papadatou-Pastou, M. (2011). Handedness and language lateralization: Why are we right-handed and left-brained. *Hellenic Journal of Psychology, 8*(2), 248-265.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*, 320.

Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A. L. (2017). Θ-Band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech. *Journal of Neuroscience*, *37*(33), 7930-7938.

Pickles, J. O. (2015). Auditory pathways: anatomy and physiology. *Handbook of Clinical Neurology, 129*, 3-25.

Plack, C. J. (2018). *The sense of hearing*. Routledge.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience, 32*(41), 14125-14131.

Potdevin, D., Adibpour, P., Garric, C., Somogyi, E., Dehaene-Lambertz, G., Rämä, P., ... & Fagard, J. (2023). Brain Lateralization for Language, Vocabulary Development and Handedness at 18 Months. *Symmetry, 15*(5), 989.

Rebhan, M., & Leibold, C. (2021). A phenomenological spiking model for octopus cells in the posterior–ventral cochlear nucleus. *Biological Cybernetics, 115*(4), 331-341.

Reed, C. M., Rabinowitz, W. M., Durlach, N. I., Braida, L. D., Conway-Fithian, S., & Schultz, M. C. (1985). Research on the Tadoma method of speech communication. The *Journal of the Acoustical society of America, 77*(1), 247-257.

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

Ren, Y., Yang, W., Nakahashi, K., Takahashi, S., & Wu, J. (2017). Audiovisual integration delayed by stimulus onset asynchrony between auditory and visual stimuli in older adults. *Perception, 46*(2), 205-218.

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*, 116134.

Rizza, A., Terekhov, A. V., Montone, G., Olivetti-Belardinelli, M., & O'Regan, J. K. (2018). Why early tactile speech aids may have failed: No perceptual integration of tactile and auditory signals. *Frontiers in Psychology, 9*, 767.

Rosenzweig, M. R., & Bennett, E. L. (1996). Psychobiology of plasticity: effects of training and experience on brain and behavior. *Behavioural Brain Research, 78*(1), 57-65.

Rubio, M. E. (2018). Microcircuits of the ventral cochlear nucleus. *The Mammalian Auditory Pathways: Synaptic Organization and Microcircuits*, 41-71.

Saija, J. D., Akyürek, E. G., Andringa, T. C., & Başkent, D. (2014). Perceptual restoration of degraded speech is preserved with advancing age. *Journal of the Association for Research in Otolaryngology*, *15*(1), 139-148.

Sainburg, R. L. (2014). Convergent models of handedness and brain lateralization. *Frontiers in Psychology, 5*, 1092.

Santurette, S., & Dau, T. (2011). The role of temporal fine structure information for the low pitch of high-frequency complex tones. *The Journal of the Acoustical Society of America, 129*(1), 282-292.

Sato, M., Cavé, C., Ménard, L., & Brasseur, A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, *48*(12), 3683-3686.

Schneiders, J. A., Opitz, B., Tang, H., Deng, Y., Xie, C., Li, H., & Mecklinger, A. (2012). The impact of auditory working memory training on the fronto-parietal working memory network. *Frontiers in Human Neuroscience, 6*, 173.

Schomer, D. L., & Da Silva, F. L. (2012). *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*(3), 106-113.

Schumann, A., Serman, M., Gefeller, O., & Hoppe, U. (2015). Computer-based auditory phoneme discrimination training improves speech recognition in noise in experienced adult cochlear implant listeners. *International Journal of Audiology, 54*(3), 190-198.

Schwartz, J. L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10*(7), e1003743.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), B69-B78.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304.

Shore, S. E. (2005). Multisensory integration in the dorsal cochlear nucleus: unit responses to acoustic and trigeminal ganglion stimulation. *European Journal of Neuroscience, 21*(12), 3334-3348.

Smiljanic, R., Keerstock, S., Meemann, K., & Ransom, S. M. (2021). Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *The Journal of the Acoustical Society of America, 149*(6), 4013-4023.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87-90.

Song, J. H., Skoe, E., Banai, K., & Kraus, N. (2012). Training to improve hearing speech in noise: biological mechanisms. *Cerebral Cortex, 22*(5), 1180-1190.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J. F., & Lewkowicz, D. J. (2012). The development of audiovisual speech perception. *Multisensory Development*, 207-228.

Souffi, S., Nodal, F. R., Bajo, V. M., & Edeline, J. M. (2021). When and how does the auditory cortex influence subcortical auditory structures? New insights about the roles of descending cortical projections. *Frontiers in Neuroscience, 15*, 690223.

Stein, B. E., & Meredith, M. A. (1993). *The Merging of the Senses*. MIT Press.

Steinmetzger, K., & Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia, 95*, 173-181.

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2022). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering, 18*(6), 066054.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Tan, S. J., Kalashnikova, M., Di Liberto, G. M., Crosse, M. J., & Burnham, D. (2022). Seeing a talking face matters: The relationship between cortical tracking of continuous auditory-visual speech and gaze behaviour in infants, children and adults. *NeuroImage*, *256*, 119217.

Teng, X., Cogan, G. B., & Poeppel, D. (2019). Speech fine structure contains critical temporal cues to support speech segmentation. *NeuroImage, 202*, 116152.

Tiippana, K., Möttönen, R., & Schwartz, J. L. (2015). Multisensory and sensorimotor interactions in speech perception. *Frontiers in Psychology, 6*, 458.

Tjan, B. S., Chao, E., & Bernstein, L. E. (2014). A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. *European Journal of Neuroscience*, *39*(8), 1323-1331.

Todaro, C., Marzetti, L., Valdés Sosa, P. A., Valdés-Hernandez, P. A., & Pizzella, V. (2019). Mapping brain activity with electrocorticography: resolution properties and robustness of inverse solutions. *Brain Topography, 32*, 583-598.

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, 466-472.

Tripathi, A. (2022). Analysis of EEG frequency bands for Envisioned Speech Recognition. *arXiv preprint, arXiv*:2203.15250.

Trussell, L. O., & Oertel, D. (2018). Microcircuits of the dorsal cochlear nucleus. *The Mammalian Auditory Pathways: Synaptic Organization and Microcircuits*, 73-99.

Tzourio-Mazoyer, N., Zago, L., Cochet, H., & Crivello, F. (2020). Development of handedness, anatomical and functional brain lateralization. In *Handbook of clinical neurology* (Vol. 173, pp. 99-105). Elsevier.

Vanthornhout, J., Decruy, L., & Francart, T. (2019). Effect of task and attention on neural tracking of speech. *Frontiers in Neuroscience*, *13*, 977.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, *19*(2), 181-191.

Venetjoki, N., Kaarlela-Tuomaala, A., Keskinen, E., & Hongisto, V. (2006). The effect of speech and speech intelligibility on task performance. *Ergonomics, 49*(11), 1068-1091.

Verschooten, E., & Joris, P. X. (2014). Estimation of neural phase locking from stimulus-evoked potentials. *Journal of the Association for Research in Otolaryngology, 15*(5), 767-787.

Verschueren, E., Vanthornhout, J., & Francart, T. (2021). The effect of stimulus intensity on neural envelope tracking. *Hearing Research*, *403*, 108175.

Von Stein, A., & Sarnthein, J. (2000). Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. *International Journal of Psychophysiology, 38*(3), 301-313.

Wairagkar, M., Hochberg, L. R., Brandman, D. M., & Stavisky, S. D. (2023, April). Synthesizing speech by decoding intracortical neural activity from dorsal motor cortex. In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 1-4). IEEE.

Warnecke, M., Peng, Z. E., & Litovsky, R. Y. (2020). The impact of temporal fine structure and signal envelope on auditory motion perception. *PLoS One, 15*(8), e0238125.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49-63.

Weyand, T. G. (2016). The multifunctional lateral geniculate nucleus. *Reviews in the Neurosciences, 27*(2), 135-157.

White, B. E., & Langdon, C. (2021). The cortical organization of listening effort: New insight from functional near-infrared spectroscopy. *NeuroImage, 240*, 118324.

Winer, J. A. (1984). The human medial geniculate body. *Hearing Research, 15*(3), 225-247.

Winer, J. A., Diehl, J. J., & Larue, D. T. (2001). Projections of auditory cortex to the medial geniculate body of the cat. *Journal of Comparative Neurology, 430*(1), 27-55.

Wingfield, A. (2016). Evolution of models of working memory and cognitive resources. *Ear and Hearing, 37*, 35S-43S.

Woodfield, A., & Akeroyd, M. A. (2010). The role of segmentation difficulties in speech-in-speech understanding in older and hearing-impaired adults. *The Journal of the Acoustical Society of America*, *128*(1), EL26-EL31.

Yi, H., Pingsterhaus, A., & Song, W. (2021). Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Frontiers in Psychology*, 12.

Yu, L., Hu, J., Shi, C., Zhou, L., Tian, M., Zhang, J., & Xu, J. (2021). The causal role of auditory cortex in auditory working memory. *Elife, 10*, e64457.

Yu, Q., Zhang, P., Qiu, J., & Fang, F. (2016). Perceptual learning of contrast detection in the human lateral geniculate nucleus. *Current Biology, 26*(23), 3176-3182.

Yuan, Y., Meyers, K., Borges, K., Lleo, Y., Fiorentino, K. A., & Oh, Y. (2021). Effects of visual speech envelope on audiovisual speech perception in multitalker listening environments. *Journal of Speech, Language, and Hearing Research, 64*(7), 2845-2853.

Zachlod, D., Rüttgers, B., Bludau, S., Mohlberg, H., Langner, R., Zilles, K., & Amunts, K. (2020). Four new cytoarchitectonic areas surrounding the primary and early auditory cortex in human brains. *Cortex, 128*, 1-21.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11*(10), 946-953.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage, 32*(4), 1826-1836.

Zerr, M., Freihorst, C., Schütz, H., Sinke, C., Müller, A., Bleich, S., ... & Szycik, G. R. (2019). Brief sensory training narrows the temporal binding window and enhances long-term multimodal speech perception. *Frontiers in Psychology, 10*, 2489.

Zuk, N. J., Murphy, J. W., Reilly, R. B., & Lalor, E. C. (2021). Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLoS Computational Biology, 17*(9), e1009358.

Chapter 2

2 Reassessing the Benefits of Audio-Visual Integration to Speech Perception and Intelligibility

Linking Statement:

This chapter investigates the contributions of audio-visual integration to speech perception and intelligibility when viseme categories – group boundaries for phonemes that share visual distinctiveness through lipreading – are considered. The chapter also investigates the window of integration for these visually distinct phonemic stimuli. To be discussed is a reassessment of the benefits of audio-visual speech.

Author Note: *This paper was accepted for publication in the Journal of Speech, Language, and Hearing Research in September 2024 and published in January 2025. It was produced in collaboration with Dr. Helen Nuttall, and Prof. Christopher Plack as co-authors.*

Statement of Authorship

Chapter 2 (Paper One):	Reassessing the Benefits of Audio-Visual Integration to Speech Perception and Intelligibility
Authors:	Brandon O'Hanlon, Christopher J Plack, and Helen E Nuttall
Publication status:	**Published** Publication has been accepted Publication has been submitted Unpublished/Unsubmitted but in manuscript style
Reference:	O'Hanlon, B., Plack, C. J., & Nuttall, H. E. (2025). Reassessing the Benefits of Audiovisual Integration to Speech Perception and Intelligibility. *Journal of Speech, Language, and Hearing Research*, *68*(1), 26-39. (https://doi.org/10.23641/asha.27641064)
Student/Principle Author:	Brandon Lee O'Hanlon
Contribution:	Theoretical conceptualization; study design; data collection; statistical data analysis; manuscript development; manuscript revisions based on supervisor and peer-reviewed feedback.

Principle Author Signature: Date: 14.04.2025

Through signing this statement, the Co-authors agree that:

(a) The student's contribution to the above papers is correct

(b) The student can incorporate this paper within the thesis

(c) The contribution of all co-authors for each paper equals 100% minus the contribution of the student.

Co-author Name:	Christopher J Plack
Co-author Contributions:	Co-supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 25/04/2025

Co-author Name:	Helen E Nuttall
Co-author Contributions:	Primary supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 28/04/2025

2.1 Abstract

Purpose: In difficult listening conditions, the visual system assists with speech perception through lipreading. Stimulus onset asynchrony (SOA) is used to investigate the interaction between the two modalities in speech perception. Previous estimates of audiovisual benefit and SOA integration period differ widely. A limitation of previous research is a lack of consideration of visemes - categories of phonemes defined by similar lip movements when produced by a speaker - to ensure that selected phonemes are visually distinct. This study aimed to reassess the benefits of audiovisual lipreading to speech perception when different viseme categories are selected as stimuli and presented in noise. The study also aimed to investigate the effects of SOA on these stimuli.

Method: Sixty participants were tested online and presented with audio-only and audiovisual stimuli containing the speaker's lip movements. The speech was presented either with or without noise and had six different SOAs (0, 200, 216.6, 233.3, 250, and 266.6 ms). Participants discriminated between speech syllables with button presses.

Results: The benefit of visual information was weaker than that in previous studies. There was a significant increase in reaction times as SOA was introduced, but no significant effects of SOA on accuracy. Furthermore, exploratory analyses suggest that the effect was not equal across viseme categories: 'Ba' was more difficult to recognise than 'Ka' in noise.

Conclusion: In summary, the findings suggest that the contributions of audiovisual integration to speech processing are weaker when considering visemes but are not sufficient to identify a full integration period.

Keywords: audiovisual speech, speech perception, multisensory integration, visemes, vision

2.2 Reassessing the Benefits of Audio-Visual Integration to Speech Perception and Intelligibility

Intelligible speech is built up from speech phonemes. Phonemes are small linguistic units - such as the /b/ phoneme that begins 'boy' in the English language - and play a large role in the identification of speech (Ewen & Van der Hulst, 2001; Bowers *et al.*, 2016). Processing of speech can be made more difficult with the introduction of noise in the environment, which reduces the ability to discriminate successfully between phonemes (Summerfield, 1992). In many cases, information from the visual sense that is relevant to the speech – such as from lipreading – can be integrated into speech processing systems to improve comprehension. In background noise, this assisting sense is recruited further (Yuan *et al*., 2021). Viewing the lip movements when an individual is speaking can help to improve the intelligibility of speech-in-noise versus when the lips are not visible (Sumby & Pollack, 1954; Maier *et al.*, 2011). The inverse can also occur, wherein incongruent lip movements influence our ability to discriminate between speech sounds. An example of this is the McGurk Effect (McGurk & MacDonald, 1976), where presenting the auditory phoneme 'Ba' with visual lip movements associated with 'Ga' leads to perceptions of the sound 'Da' instead. A more recent example comes from face mask-wearing due to the COVID-19 pandemic. Brown *et al.* (2021) found that if the speaker wore a facemask that either fully or partially covered lip movements, performance on speech discrimination tasks decreased dramatically. These data indicate that the visual and auditory systems interact to influence how we perceive speech.

However, estimates of audiovisual benefit vary widely in the literature, likely due to stimulus-dependent effects (Ma *et al.*, 2009), in that, how the stimuli are created for lab experimentation drastically affects how participants respond to speech discrimination tasks. For example, whilst it is important for research on audio-visual processing to consider how

auditorily distinctive sounds are, visual distinctiveness is equally important. A way to examine the effect of visual distinctiveness is to select phonemes from separate viseme categories for testing. A viseme category is a group of phonemes from the English language that share the lip movements and visual information portrayed by each phoneme when spoken (Massaro *et al.*, 2012). Fisher (1968) identified five viseme categories based purely on visual distinguishability for English phonemes. Examples of phonemes that belong to the same viseme category are /b/, /p/, and /m/ which in syllable form can correspond to 'Ba', 'Pa', and 'Ma'. If two speech tokens share the same viseme, then it is impossible to discern which was spoken through lip-reading alone (Van Engen *et al*., 2022) and are only distinguishable through sound. This means that any measure of audiovisual benefit derived from discriminating within a viseme category will be lessened. It is therefore important to select stimuli from separate viseme categories when investigating how auditory and visual systems work together during speech syllable discrimination.

When speaking with others, we typically see lip movements before we hear the spoken words (Chandrasekaran *et al*., 2009) as a form of natural stimulus onset asynchrony (SOA). SOA is when two different modalities of information in cross-modal stimuli are presented at different onsets. The window of integration is the term given to the period in which visual information can lead or lag speech sounds before the visual information is no longer perceived as part of the same stimulus (Stein & Meredith, 1993). If the lip movements are desynchronised from the speech sounds within a specific period, then we still perceive the lip movements and the speech we hear to be congruent. If the SOA is large enough that the auditory and visual information do not fall within the same window of integration, we may perceive the two modalities as separate, and therefore not process the visual information as helpful extra information to discern and comprehend the speech. For speech signals, syllables have a window with an upper limit of about 240 ms and short words of about 300 ms

(Navarra *et al*., 2005). Although it is important to note that this window of integration can be highly stimulus-dependent, and ranges in the literature between 150 and 800 ms (Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al.,* 2017), and even differs between age groups (Ren *et al.,* 2017). This mixed range in the literature could also be due to a mismatch with reported display refresh rates (typically 60 Hz), video framerates (typically 30 – 60 frames per second) and levels of SOA used in research if reported at all (Ren *et al.*, 2017). For example, in 10 ms increments, a 60 Hz monitor can't display separate visual streams of information that refresh every 10 ms, as it is only capable of doing so every 16.6 ms, assuming the video plays at a full 60 frames per second as well.

The present study aimed to reassess the benefits of visual information to speech-in-noise perception using stimuli with visual distinctiveness. We also aimed to determine the effect of SOA on audiovisual speech perception. We tested the following hypotheses:

(i) purely audio speech discrimination accuracy will be decreased when speech is presented in noise compared to without noise.

(ii) reaction time to correctly discriminated purely audio speech will be increased when speech is presented in noise compared to without noise.

(iii) speech-in-noise discrimination accuracy will be increased when speech is presented with congruent visual information of the speaker's lip movements (audiovisual stimuli) compared to when no visual information is present (purely audio stimuli).

(iv) reaction time to correctly discriminated speech-in-noise will be decreased when speech is presented with congruent visual information of the speaker's lip movements (audiovisual stimuli) compared to when no visual information is present (purely audio stimuli).

(v) as the visual information precedes the auditory information by larger SOAs (0 - 266 ms), speech-in-noise discrimination accuracy will decrease.

(vi) as the visual information precedes the auditory information by larger SOAs, reaction time to correctly discriminated speech-in-noise will increase.

Further exploratory analysis also investigated the window of integration for these audiovisual stimuli, as well as differences in visual benefit between each syllable used.

2.3 Methods

2.3.1 Design

To address hypotheses (i) and (ii), a single within factor (noise type: speech without noise, and speech-in-noise) design was used for purely audio trials with no stimulus asynchrony. For hypotheses (iii) and (iv), a single within factor (stimulus type: audiovisual, and purely audio) design was used for speech-in-noise trials with no stimulus asynchrony. Finally, for hypotheses (v) and (vi), a single within factor (SOA; 0, 200, 216.6, 233.3, 250, and 266.6 ms) design was used for audiovisual, speech-in-noise trials only. In total, participants took part in all 14 unique conditions (see Table 1), and both the accuracy of speech discrimination and reaction time in the discrimination task were recorded.

Ethical approval was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (approval reference: FST-2022-2122-RECR-2, project ID: 2122). The study was pre-registered on AsPredicted.org before commencing data collection. The pre-registration can be found at https://aspredicted.org/aq98a.pdf. All deviations from this pre-registration are listed in the section below. The collected data have been archived on the Open Science Framework (OSF: https://osf.io/kcbzs).

2.3.2 Deviations from pre-registration

In the original study pre-registration, there were three set hypotheses listed:

- There will be a decrease in the accuracy of speech discrimination (measured by correct responses in trials) or an increase in response times in the auditory-only condition when the speech is in noise compared to speech without noise.
- When visual information is present (audiovisual), the accuracy of speech discrimination and response times for each trial will not be as obstructed in

speech-in-noise conditions compared to audio-only conditions (when no visual information is present).

- As the visual information precedes the auditory information by larger margins (200 ms, 216 ms, 233 ms, 250 ms, 266 ms), the accuracy of speech discrimination in the speech-in-noise conditions will decrease - or response times will increase - in audiovisual conditions compared to when the audiovisual information is congruent (0 ms).

These were changed to the hypotheses listed in the introduction by splitting the dependent measures into separate hypotheses and improving readability. This was done to make interpretations of results more clearly defined when referring to the hypotheses. To accompany this, the models used to test the hypothesis were also adjusted, giving six separate models of analysis - one for each hypothesis - instead of four. Generalised linear mixed-effects models (GLMER) were used to test all six hypotheses, instead of the mixture of GLMER models for accuracy data and LMER models for reaction time data that was listed in the pre-registration. This was done as GLMER models are more appropriate than LMER models for reaction time data, which is generally positively skewed (Lo & Andrews, 2015). These GLMER models were preferable still over repeated measures generalised linear models for considering random effects that may be present on a participant-by-participant basis. Finally, in our sample size calculation using data simulation (see section ‘Sample size calculation’), it was determined that 60 participants were needed to sufficiently power the study. In the pre-registration, we then added a further 10% after a priori calculations (another 6 participants) to make a sample size estimate of 66. Due to the availability of resources, this extra 10% was not collected, leaving the sample of the study at the original number of 60 participants.

2.3.3 Participants

All data were collected online, with 81 participants recruited for the study. Of these, a total of 60 participants completed the study (mean age = 25.66 years, 28 male, 30 female, two non-binary). The other 21 participants completed the eligibility questionnaire but were either not eligible or did not proceed to the study task and provide study data. Participants were recruited via online advertisements or through Prolific and were compensated for their time. All participants were monolingual, native speakers of British English to control any potential speech perception differences across languages and in bilingualism and multilingualism (Lotfi *et al.*, 2019). Participants reported no hearing disorders and had either normal or corrected-to-normal vision. Only those between the ages of 18 and 35 were tested, as the window of integration for audiovisual information increases significantly with age, which can make speech discrimination more difficult (Ganesh et al., 2018; Sekiyama *et al.*, 2014). Participants reported no developmental disorders, such as dyslexia, or history of developmental disorders. This was important as the window of integration for audiovisual stimuli is wider in individuals with learning difficulties such as autism spectrum disorder and developmental dyslexia (Smith & Bennetto, 2007; Megnin-Viggars & Goswami, 2013; Michalek et al., 2014; Noel et al., 2018). All participants were right-handed. Finally, participants had no musical expertise, as previous research suggests that individuals with continuous experience as musicians can detect smaller SOAs, even for speech syllables (Lee & Noppeney, 2014; Sorati & Behne, 2019). Musical expertise was defined as training with a single musical instrument or voice for more than 7 years (Varnet *et al.*, 2015; Lee *et al.*, 2020) and for at least 3.5 hours a week (Lee & Noppeney, 2014). Participants were screened for the experiment using Qualtrics (see section ‘Procedure’).

2.3.4 Sample size calculation

Before testing, data simulation was conducted using R 4.2.2 (R Core Team, 2022) for power and sample size analysis. *Lme4* (v1.1-27.1; Bates *et al*., 2015), *afex* (v1.0-1; Singmann *et al.*, 2024) and *simr* (v1.0.5; Peter *et al.*, 2019) were the core packages utilised in this process. Firstly, means and standard deviations of accuracy were gathered from studies that used syllable or bi-syllable phonetic speech tokens to investigate visual integration in speech perception. These studies typically used either multiple signal-to-noise ratios (SNRs; between -12 and -18 dB: Altieri *et al.*, 2014; Grant & Seitz, 1998; Sekiyama *et al.*, 2014) or individualised ratios (Ten Oever *et al.* 2013). For those studies that used multiple speech-to-noise ratios, we took data from – or closest to – -16 dB SNR. -16 dB was selected for our speech-shaped noise as this was the average SNR at which there was a notable difference between perceiving speech with or without visual aid (Bernstein *et al.*, 2004). An average estimated mean and standard deviation were then calculated for each condition. A dataset was produced using the *rtruncnorm* function (*truncnorm* package; v1.0-8; Mersmann *et al*., 2018) - to randomly generate data for each condition that had a mean and standard deviation close to the ones calculated. This was repeated for each speech token ('Ba', 'Fa', and 'Ka') and all trials of each condition, providing a full dataset of expected results.

The dataset was then analysed using our planned experimental analyses (see below) to generate predicted results. Simulations were repeated 1000 times. An aggregation of power was then calculated. If the power was insufficient (below .80 at an alpha level of .05), the sample size of the dataset was manually adjusted, and the data simulation was conducted again. This was done until a minimal sample size with sufficient power was found. A total of 60 participants were calculated to be needed for sufficient power. The code for data simulation is available on OSF (https://osf.io/kcbzs).

2.3.5 Materials

The experiment was created using PsychoPy 3's builder tools (v2021.2.3; Peirce *et al.*, 2019) and hosted online through Pavlovia. A consent form and a screening form were created and hosted on Qualtrics (Qualtrics, 2005). Three single-syllable speech tokens were used: 'Ba', 'Fa', and 'Ka'. These were chosen as they belong to three distinct viseme categories, did not rely on distinguishing any tongue movements that would have been obscured from sight (such as labiodental phonemes), and could be easily distinguished without visual aid when not in noise. These speech tokens were spoken by a native British English-speaking male speaker and were recorded using personal home equipment. An external USB 3.0 condenser microphone was used to record audio (HyperX Quadcast with default windshield, set to the cardioid position). The initial video footage was recorded at 1920 x 1080 resolution and 60 frames per second using a mobile device (OnePlus 7 Pro). Both devices were connected to a single desktop machine, which recorded the audio and video in tandem using open-source OBS Studio software (Open Broadcaster Software, version 29.1.3). After the initial recording, the speech tokens were edited in length and converted to mp4 files at a resolution of 1280 x 720 and a frame rate of 60 frames per second. As the study would be completed on participants' laptops or desktop systems and using their internet connection, we could not ensure that all participants were using a device with a 1920 x 1080 resolution screen. By reducing the resolution of files to 1280 x 720, all likely participant resolution sizes could be accommodated whilst ensuring that all participants viewed the files at the same resolution. Sixty frames per second was chosen as the frame rate as home device monitors and laptop screens are typically to a standard 60 Hz or higher. By using the lower boundary and not a higher frame rate, we can be sure that all SOAs implemented in the stimuli were visually relayed to the participant. For audiovisual conditions, the video footage contained only the speaker's lower face in view, containing

mouth and lips. This meant that participants were only provided with visual information regarding the lip movements made when speaking, and not any other visual information relevant to other actions the speaker may have made during recordings. For audio-only conditions, the video of the lips was overlayed with a plain black PNG image file. This kept the audio-only stimuli in a consistent video format rather than exporting the file as an mp3. All video files were the same length of 2 s.

Audacity software (Audacity Team, 2021) was then used to rip the audio from the MKV files to be edited as WAV files in Praat software (Boersma & Weenink, 2021) for the creation of speech-shaped noise. First, a sentence using English words – '*His plan meant taking a big risk'* - was recorded to provide a base for the speech-shaped noise. White noise was then produced using Praat's white noise generator. The noise was brought down to an intensity tier, then an amplitude tier. This was then multiplied with the sentence above to create speech-shaped noise (Van Engen *et al.*, 2017). Praat was then used to combine the speech-shaped noise with the speech-in-noise conditions at a speech-to-noise ratio of –16 dB. This was done using a Praat script developed by McCloy (2021). Finally, Audacity was used again to ramp up the start and ramp down the ends of all audio files for every condition. The audio was then stitched back onto the MP4 files.

For the conditions where the stimuli were asynchronous, Lightworks was again used to desynchronise the onset of the audio ahead of the onset of the lip movements using exact frames of the video footage (12, 13, 14, 15, and 16 frames per second) which corresponded with the SOAs of the relevant conditions (audio starting after the visual lip information by 200, 216.6, 233.3, 250, and 266.6 ms). The result was 42 stimuli in MP4 format, representing three speech tokens ('Ba', 'Fa', and 'Ka') for each of the 14 condition levels presented to the participant.

2.3.6 Procedure

Participants were linked to Qualtrics once they had consented to the study. Participants were also reminded at this stage to ensure that they were in a quiet room with no background noise, as well as to load the experiment on either Microsoft Edge, Google Chrome, or Mozilla Firefox internet browsers on a laptop or desktop computer. They were explicitly told not to open the experiment on any other browser, such as Safari, nor a mobile or tablet device as these were incompatible. Participants were also instructed to use headphones for the experiment, rather than to play the stimuli through their device's speakers.

A volume check began, in which a constant pure tone played (440 Hz frequency), and participants were asked to adjust the volume of their device as necessary for a comfortable auditory experience and to ensure that the audio was playing correctly at a sufficient volume level. This tone would play for as long as the participant wished to alter the volume levels of their device. Once complete, the spacebar would be pressed, and the tone stopped. Participants were informed that a video would play either showing no visual information or visual information of lips moving. Meanwhile, speech would be played. Participants were told to listen carefully to the speech sound spoken, and after hearing the sound to press one of three buttons on their keyboards that corresponded with the three available speech tokens. They were instructed to respond to each trial as quickly as possible. They were reminded before and after each trial to press 'z' on their keyboard if they heard 'Ba', 'x' for 'Fa', or 'c' for 'Ka'. If they were unsure, they were told to make a guess.

Participants were given six practice trials before data were collected. This was using the speech without noise, 0 ms, and audiovisual condition stimuli, with two trials for each of the three speech tokens (Ba, Fa, and Ka). A white crosshair would be displayed on the screen for 1000 ms before the trial began to bring attention to the centre of the screen where the

video trials would be displayed. Stimuli were shown for 2500 ms, then the response screen would display. On this screen, the participants were reminded of the buttons to press for each of the three speech sounds. Only the three buttons could be pressed and pressing the buttons whilst the stimuli were still playing would not record a response or stop the trial. A total of 546 trials (not including the practice trials) were completed. The order of the trials and conditions was completely random to avoid any potential order bias. After every 42 trials, a break screen would appear. This screen told the participant to take a short break before continuing with a press of the spacebar. If the participant did not wish to take a break, they were permitted to continue with a spacebar press immediately. There was a total of 12 breaks in the experiment, each with a short attention check question to ensure participants remained attentive to the experiment. Upon completing the study, participants could close the browser tab or window down and all data would remain recorded on the Pavlovia system.

2.3.7 Analysis

Descriptive statistics were first gathered from each condition for both the accuracy ratings and the reaction times. Reaction times were taken from the offset of the stimuli to the participant response. The average accuracy and reaction time of accurately responded trials for each condition and each participant was calculated, with reaction times winsorised over the 95^{th} percentile only. This was done to replace any large, outlying reaction times to trials that may be due to a distraction at home during testing or the participant taking a short break before the break period. The assumptions of linear and generalised linear mixed-effects models were tested, including residual plots to check for linearity, quantile-quantile plots for normality, assessing the levels of multicollinearity between stimulus type, noise, and SOA using variance inflation factors, and ensuring the assumption of homoscedasticity was met. All the above tests were conducted on the dataset and all assumptions were met. As we were

testing six separate hypotheses, the experiment-wise error rate was controlled using the Bonferroni-Holm method (Holm, 1979).

With further regards to stimulus variability, previous studies often employ analyses such as repeated measures analysis of variance (ANOVA) tests which do not consider random effects (Bates *et al.*, 2015). Including random effects is important for ensuring that any effects found in the model are not influenced by differences in participant ability or by the stimuli themselves, as some stimuli may be easier to recognise and comprehend in noise than others. To counter this issue, mixed-effects models can be used that consider the random effects, such as participant number and stimuli number, across intercepts and slopes within the model to provide a more valid interpretation of the integration between visual and auditory systems in speech perception.

Using the *lme4* package (Bates *et al*., 2015), generalized linear mixed-effects regression model (GLMER) analyses were conducted for the accuracy scores to test hypotheses (i), (iii), and (v) and for reaction time scores to test hypotheses (ii), (iv), and (vi). GLMERs were chosen instead of repeated measures generalised linear models such as ANOVA tests because they consider random effects that may be present across all 546 trials on a participant-by-participant basis. GLMER was chosen over LMER for analysis with reaction times as these scores are typically positively skewed. As noted by Lo and Andrews (2015), generalised linear mixed models are more appropriate for skewed datasets in this context. Furthermore, accuracy in a trial is a binary outcome variable that can either be correct (1) or incorrect (0). Therefore, GLMERs were used to ensure that assumptions of categorical dependent variables in mixed-effects models were met. GLMERs were conducted using the *lme4* package still, as this package supported a generalised approach. Due to the generalised nature of the model and package restrictions, no suitable p-values were provided with the GLMER analyses. Instead, significance was interpreted using 99.2% confidence

intervals (CIs), chosen to reflect our lowest criterion of significance in the Bonferroni-Holm approach being $p < .008$ for six comparisons. If the resulting confidence intervals showed insignificance, the next boundary of Bonferroni-Holm ($p < .01$) was checked using 99% confidence intervals. This kept going until either significance was found or no significance was found at a significance level of $p < .05$. Once detected or classed as insignificant, the test was ranked with the other p-values in our analyses as the lowest boundary of significance and Bonferroni-Holm was conducted as normal on our six ranked comparisons.

To test hypothesis (i), a GLMER analysis was conducted using the accuracy of responses on the speech discrimination task as the dependent variable and using noise type (no noise or speech-shaped noise) as the independent variable in the model. As we hypothesised that presenting speech in noise would significantly decrease accuracy compared to without noise, we expected to find a significant effect of noise type from this GLMER analysis. Hypothesis (ii) was the same as the first but looked at reaction times to correctly discriminated speech-in-noise on the same task. A GLMER was used to test this hypothesis, using reaction times as the dependent variable and noise type as the independent variable. Similarly, we expected to find a significant effect of noise type, increasing reaction times.

To test hypothesis (iii), a GLMER analysis was conducted using the accuracy of responses on the speech discrimination task as the dependent variable and using stimulus type (purely audio or audiovisual) as the independent variable in the model. As we hypothesised that presenting audiovisual stimuli in noise would significantly increase accuracy compared to purely audio stimuli in noise, we expected to find a significant effect of stimulus type from this GLMER analysis. Hypothesis (iv) was the same as the third but looked at reaction times to correctly discriminated speech-in-noise on the same task. A GLMER was used to test this hypothesis, using reaction times as the dependent variable and stimulus type as the

independent variable. Similarly, we expected to find a significant effect of stimulus type, decreasing reaction times.

To test hypothesis (v), we conducted a GLMER analysis using accuracy as a dependent variable and SOA as the independent variable. SOA was treated as a categorical variable in this model and the model for hypothesis (vi) below. We expected to find a significant effect of SOA, with accuracy decreasing when more asynchrony was introduced to the stimuli. This would reflect that the window of integration for audiovisual speech is important for visual information to be beneficial to understanding speech in noise. Finally, in a similar manner, hypothesis (vi) was tested using a GLMER analysis with reaction times as the dependent variable and with SOA levels as the independent variable in the model. Again, we expected a significant effect of SOA on reaction times, with reaction times increasing with the introduction of asynchrony.

For all six GLMER models listed above, the speech sound token used (Ba, Fa, or Ka), participant age and the participant ID were all included as random effects. No further model selection of these random and fixed effects was undergone, as we wanted a conservative model that included a full random effects structure to account for the expected larger individual differences of an online experiment. All model equations and structures can be found in the supplementary materials (Table 2).

Furthermore, we also conducted exploratory analyses to assess the effect of noise on speech discrimination accuracy between the three visually distinct, chosen phonemes (‘Ba’, ‘Fa’, and ‘Ka’). To do this, a GLMER analysis was conducted using accuracy as the dependent variable and speech token as the independent variable. Purely audio trials in noise were used for this analysis. Furthermore, we also conducted pairwise comparisons within the GLMER models used to test hypotheses (v) and (vi) as another exploratory analysis,

comparing between each level of our SOA independent variable. We expect that not all the SOA interactions will show significance. As we expected the benefits of visual stimuli to only be present during the window of integration, there would only be a significant decrease in accuracy and an increase in reaction times at SOAs outside this window. Therefore, this exploratory analysis can be used to better understand the window of integration for our stimuli. All exploratory analyses will use an inference criterion of $p < .008$ as this was the strictest threshold for significance included in our Bonferroni-Holm correction.

2.4 Results

2.4.1 Descriptive statistics

The means and standard deviations of the accuracy of responses and reaction times of responses can be seen in Table 1. Descriptive statistics were also calculated for each speech token (Ba, Fa, and Ka). Figure 1 shows the mean reaction times and mean accuracy rates for both audio-only and audiovisual stimuli when no SOA is considered (0 ms SOA), whilst Figure 2 shows these data for all SOAs when audiovisual stimuli are used for speech-in-noise conditions. Figure 3 shows the mean reaction times and accuracy rates for all SOAs when audiovisual stimuli are presented without noise. Furthermore, Figure 4 shows accuracy rates and reaction times in purely audio and audiovisual stimuli in noise between each of the three speech tokens. Violin plots were used for all figures to highlight the distribution of accuracies and reaction times across participants for each condition, as individual differences were large in this dataset likely due to online experimentation.

2.4.2 Effect of noise on speech perception

The first planned GLMER analysis was conducted to test hypothesis (i). There was a significant effect of noise type (with or without noise), showing a decrease in accuracy in speech-in-noise discrimination when noise was introduced versus clear speech (β = -.29, t = -12.95, 99.2% *CI* = [-.35, -.23], $p < .008$). This model supports hypothesis (i), as we expected to find that the introduction of noise to speech would decrease performance. For testing hypothesis (ii), the planned GLMER analysis was conducted. There was a significant effect of noise type on reaction times (β = .06, t = 3.10, 99.2% CI = [.01, .11], $p < .008$). This model supports hypothesis (ii), as we expected to find that introducing noise would increase reaction times to correctly discriminated speech.

2.4.3 Effect of congruent, distinguishable visual information on speech perception

Our next planned GLMER analysis was conducted to test hypothesis (iii). There was a significant effect of stimulus type (purely audio or audiovisual), as there was an increase in accuracy in speech-in-noise discrimination when stimulus type was audiovisual versus purely audio ($\beta = .26$, $t = 11.36$, 99.2% $CI = [.20, .32]$, $p < .008$). This model supports hypothesis (iii), as we expected to find that introducing relevant visual information would improve speech perception in noise. For testing hypothesis (iv), the planned GLMER analysis was conducted. There was a significant effect of stimulus type on reaction times ($\beta = -.08$, $t = -4.15$, 99.2% CI = [-.13, -.03], $p < .008$). This model supports hypothesis (iv), as we expected to find that introducing relevant visual information would decrease reaction times and improve speech perception in noise.

2.4.4 Effect of stimulus onset asynchrony on audiovisual speech perception

When testing hypothesis (v), the planned GLMER analysis was done for data across all SOA levels for audiovisual speech-in-noise stimuli only. There was no significant effect of SOA on accuracy at any interval, even at a 95% confidence interval, showing no support for hypothesis (v). Finally, our planned GLMER analysis was run to test hypothesis (vi). There was a significant main effect of SOA ($\beta = .04$, $t = 3.31$, $p < .008$) on reaction times, indicating reaction times increased with SOA. This supports hypothesis (vi).

2.4.5 Exploratory analyses

As a further, exploratory analysis, a GLMER model was used to investigate phoneme differences in speech-in-noise discrimination. Looking at pairwise comparisons, there was a significant difference between accuracy rates of the ‘Ba’ and ‘Fa’ tokens ($\beta = -.17$, $t = -5.03$, $p < .008$), ‘Ba’ and ‘Ka’ tokens ($\beta = -.53$, $t = -15.50$, $p < .001$), and ‘Fa’ and ‘Ka’ tokens ($\beta = -.36$, $t = -10.47$, $p < .008$) for purely audio stimuli. For audiovisual stimuli, however, there

was no significant change in accuracy rate between the three tokens. A GLMER model for reaction times showed similar patterns, although only ‘Ba’ and ‘Ka’ were significantly different for purely audio stimuli, with ‘Ba’ having increased reaction times in comparison to ‘Ka’ ($\beta = .14$, $t = 4.21$, $p < .008$).

Finally, to explore differences between SOA intervals to see if a window of integration could be determined, pairwise comparisons were made on the GLMER analyses used to test hypothesis (vi). Pairwise comparisons were not made on the GLMER used to test hypothesis (v) as no significant effect of SOA on accuracy was observed. Pairwise comparisons made on the GLMER to test hypothesis (vi) indicated that reaction times were significantly reduced compared to 0 ms at 250 ($\beta = -.05$, $t = -3.94$, $p = .001$) and 266.6 ms ($\beta = -.05$, $t = -3.88$, $p = .002$). However, no other comparisons between levels of SOA were significantly different. Whilst this implies that a minimal end of the window of integration could lie above 233.3 ms (as SOAs between 233.3 and 250 ms were not tested), no accurate window of integration can be determined from the data.

2.5 Discussion

This study aimed to reassess the contribution of audiovisual integration to speech perception in noise when stimuli belonged to different viseme categories. As speech perception can differ wildly with stimuli sets, it was important to first reassess the detriment of noise on speech discrimination, as well as the benefits of speech-relevant visual integration. The study incorporated the visual distinguishability of each speech phoneme used in the speech discrimination task by selecting phonemes from separate viseme categories. Furthermore, the study also aimed to examine the effects of stimulus onset asynchrony (SOA) on audiovisual speech perception. This may assist in determining a window of integration for these stimuli, which was explored in further analyses.

2.5.1 Reassessing the detriment of noise on speech perception

GLMERs were used to investigate the influence of the predictor variables on accuracy ratings on the speech discrimination task. The first model, using noise type as the predictor, supported our first hypothesis, showing that there was a decrease in accuracy for purely audio stimuli when the speech was presented in noise compared to without noise. Additionally, the introduction of noise to the speech signal increased reaction times significantly. These results support our second hypothesis. As both the accuracy and reaction time to trials with noise differed significantly from those without, it can be said that the detriment of noise on speech perception was present with our created stimuli and chosen SNR ratio of -16 dB using speech-shaped white noise.

2.5.2 Reassessing the contribution of audiovisual information on speech processing in noise

There was a significant increase in accuracy when relevant, congruent visual information was present with the stimuli versus purely audio stimuli in noise. This supports

hypothesis (iii) and confirms previous findings regarding the contribution of audiovisual information to speech-in-noise processing. However, it should be noted that whilst the effect is prominent, it is not as great as previous literature findings which used a similar speech-to-noise ratio (Van de Rijt *et al.*, 2019). Here, the effectiveness of audiovisual enhancement of speech recognition was assessed with SNR ratios as low as -21 dB SNR, where the introduction of relevant visual cues provided an increase in accuracy of up to 50% for some stimuli, with greater enhancements for words like 'Pieter'. Even at -16 dB SNR, Van de Rijt et al.'s data suggests that greater audiovisual enhancement should have been seen, though reaction time data was not reported in the study.

This could also be explained using results from our exploratory analysis. When the speech was in noise and the stimuli contained auditory information only, the token 'Ba' displayed much lower mean accuracy scores than the other tokens. This suggests that there are specific differences in the acoustic properties of the tokens used that are influencing the perception of speech-in-noise. In previous literature, 'Ba' and other tokens within the same viseme (such as 'Pa') are frequently used, which could suggest why results in previous literature show a larger speech discrimination effect in noise. It is therefore important for future research to determine if there are differences in speech perception between other viseme categories that were not used in this study (Fisher, 1968). In our LMER model for hypothesis (iii), the token used was loaded as a random factor. This variance between tokens was removed from the variance found in fixed effects in the outputs of the model. This mixed effect modelling also considered participant differences and age, unlike previous literature that did not investigate speech discrimination effects using more complex models (Bernstein *et al.*, 2004; Sekiyama *et al.*, 2014). As the tokens appear to be largely variant, this could further account for the weaker overall patterns of change seen between fixed effects.

Next, there was a significant decrease in reaction times when audiovisual stimuli were used over purely audio, supporting hypothesis (iv). Interestingly, there was a decrease in reaction time in audiovisual conditions with noise over without noise as well. When processing multisensory stimuli that are not beneficial to us, reaction times likely increase due to extra unnecessary processing (Brown & Strand, 2019). In this case, the audiovisual information is only beneficial to us in noise. Therefore, in this model where no comparisons to clear speech are made, reaction times significantly decrease with the introduction of noise as the extra processing of visual information becomes beneficial. Comparatively, when audiovisual information is present without noise, reaction times increase as the added visual information is no longer beneficial to speech recognition as it is already clear to understand.

2.5.3 Investigating the effects of stimulus onset asynchrony on the speech processing benefits of audiovisual information

Our GLMER model testing hypothesis (v) uncovered no meaningful change in accuracy between any SOA value. In previous research, the maximal window of integration was around 250 to 260 ms for syllables (Dixon & Spitz, 1980). Here, SOAs up to 266.6 ms did not affect speech discrimination accuracy, implying that the stimuli were still inside the window of integration and that the maximal end of the window lies beyond 266.6 ms. Our final LMER model testing hypothesis (vi) found significant increases in reaction time when an SOA was introduced. Alternatively, this implies that the range of SOAs used does cover the maximal end of the window concerning processing speed, as there was a gradual increase in reaction times as SOA was further increased reducing the benefit of audiovisual information. When looking at exploratory pairwise comparisons between SOA levels, there was a distinct decrease in reaction times at 250 and 266.6 ms compared to no SOA. This implies that the ability to discriminate the speech was made less taxing past 250 ms asynchrony. It could be, based on these findings, that the minimal end of the window of

integration for our created stimuli lies between 233.3 and 250 ms. Given that the stimuli were simple syllables, an alternative interpretation may be that the processing of the auditory and the visual information was completed before integration had finished, although this would not explain the differences in reaction times between the SOA levels. Furthermore, as participants could only respond after the stimuli had played in full, with visual cues preceding the auditory cues, we would expect integration to have occurred if the SOA remained within the window of integration. As these comparisons are exploratory, however, and there is no account of accuracy changing with SOAs, further research would be needed to determine the full window of integration.

2.5.4 Limitations of the study and future directions

One explanation for the audiovisual benefit in our data not being as large as in previous studies could be the lack of ecological validity and the artificial nature of the online experimentation. Speech-shaped white noise was utilised for speech-in-noise conditions. Despite this noise modulating speech, it is still unlike that in a real environment. This may mean that the speech-shaped noise was too distinct from the speech itself, especially considering that we used syllables for recognition rather than words or sentences. Speaker babble or background noise such as light vocal music would be much more akin to that in everyday life, making it perhaps more suitable and valid for investigating audiovisual speech perception when speech is in noise (Krishnamurthy & Hansen, 2009). Furthermore, the stimuli used were single syllable speech tokens, which do not reflect typical communicative speech in a real-world environment. Given their simplicity, other aspects of speech perception, such as prediction of oncoming words in larger sentences, would not be used as a method of speech processing here (Solberg Økland *et al.*, 2019). The overall simplicity and artificial design of these stimuli may be obscuring other benefits of audiovisual integration in speech perception when applied to realistic speech settings. To better reassess audiovisual

integration in speech, further research with more ecologically valid speech stimuli (e.g., full sentences) would be of benefit.

The SNR used for our study was -16 dB. This was selected based on previous research investigating audiovisual syllable perception in noise, for which there was a notable difference between perceiving speech with or without visual aid (Bernstein *et al.*, 2004). However, whilst this may have been true for speech token 'Ba', this did not seem to translate to 'Ka', indicating that different speech viseme categories were affected by speech-shaped noise at the SNR -16 dB. Furthermore, initial data collection for this study was conducted from 2021 to 2022 after multiple lockdowns in the UK due to the COVID-19 pandemic. Many adults in the UK during this time had been socially distancing and wearing facemasks to prevent contamination. These facemasks would obscure the lip and mouth area of the wearer, meaning that social interactions between many people in this period would have lacked visual information to assist with speech perception. In many cases, the facemasks obscured sound, making it more difficult to understand speech and imitating difficult listening conditions (Yi *et al.*, 2021; Smiljanic *et al.*, 2021). It is possible that due to facemask wearing for a year, participants had adapted to listening to speech in difficult conditions without visual aids. Furthermore, only three phonemes from three viseme categories were used in this study. As there was an apparent difference between these phonemes selected, with 'Ba' being more impacted by added noise than 'Ka', future studies may wish to investigate the differences between more viseme categories and the phonemes within them. It may also be beneficial to further apply this to more than single-syllable units of speech. This would provide a broader view of the contributions of visual information to speech processing.

Finally, this experiment did use home equipment to record stimuli as well as the home equipment of participants to play the stimuli through online experimentation. Whilst the

recording equipment was of laboratory standard and the recording procedure rigorous, there will still be discrepancies between these stimuli and other lab-created stimuli which might make replications difficult. Furthermore, the environments that participants were in whilst taking part in the study may be different between participants. We do not have measures of how well participants understood the task, how noisy their environment was during listening, the hardware they used to run the study, and if they followed pre-experiment instructions such as to wear headphones. These are likely to contribute to the large individual differences seen in the dataset. Whilst GLMER models can consider the participant differences, further in-person lab testing with similar methodologies may be needed to fully control these confounds.

2.5.5 Conclusion

A set of purely audio and viseme-controlled audiovisual stimuli was created to investigate the contributions of audiovisual information to speech-in-noise processing. Introducing visual information increased accuracy and decreased reaction times in speech-in-noise conditions relative to audio-only stimuli. When looking at accuracy and reaction times at varying SOA intervals in our audiovisual stimuli, introducing SOAs influenced reaction times, but not accuracy. In the future, more syllables from more viseme categories could be tested to investigate a full range of speech sounds in audio-only and audio-visual contexts, as well as with further SOA intervals to ensure that a window of integration can be determined with accuracy.

2.6 Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC) Training Grant (O'Hanlon, ES/P000665/1), the Manchester Biomedical Research Centre and the National Institute for Health and Care Research (NIHR) (Plack, NIHR203308), and the Biotechnology and Biological Sciences Research Council (BBSRC) New Investigator Grant (Nuttall, BB/S008527/1). We would like to thank all participants who expressed interest and participated in this research. We also thank Kyle Stonehouse for their contributions to study material creation.

2.7 Data Availability Statement

Upon publication, all collected data are available to view online through the Open Science Framework (OSF: https://osf.io/kcbzs), as well as all stimuli used in the experiment code relevant to data analysis.

Tables and Figures

			Accuracy Rate (%)		Reaction Time (ms)	
Speech	Stimuli	SOA (ms)	Mean	Std. Dev	Mean	Std. Dev
Clear	AO	0	96.11	10.24	538	216
Clear	AV	0	96.60	13.55	564	257
Clear	AV	200	95.95	14.09	551	241
Clear	AV	216.6	96.56	13.82	573	232
Clear	AV	233.3	96.58	12.61	575	249
Clear	AV	250	96.54	13.90	568	236
Clear	AV	266.6	96.84	13.23	575	255
Noise	AO	0	67.33	21.91	597	285
Noise	AV	0	93.10	15.21	518	239
Noise	AV	200	92.87	16.08	553	223
Noise	AV	216.6	93.62	15.75	554	218
Noise	AV	233.3	93.35	17.52	562	227
Noise	AV	250	93.11	16.30	570	224
Noise	AV	266.6	93.26	15.01	569	237

Table 1.

Means and Standard Deviations (Std. Dev) of accuracy rates and reaction times for speech with and without noise, audio-only (AO) or audiovisual (AV) stimuli, and different stimulus onset asynchronies (SOAs), with each speech token and participant aggregated into a single mean.

Hypothesis	Model Equation
(i)	Accuracy ~ Noise Type + (1|ID) + (1|Age) + (1|Speech Token)
(ii)	Reaction Time ~ Noise Type + (1|ID) + (1|Age) + (1|Speech Token)
(iii)	Accuracy ~ Stimuli Type + (1|ID) + (1|Age) + (1|Speech Token)
(iv)	Reaction Time ~ Stimuli Type + (1|ID) + (1|Age) + (1|Speech Token)
(v)	Accuracy ~ SOA Level + (1|ID) + (1|Age) + (1|Speech Token)
(vi)	Reaction Time ~ SOA Level + (1|ID) + (1|Age) + (1|Speech Token)

Table 2.

Generalised Linear Mixed-Effects Regression Model (GLMER) equations, including the fixed-effects and random effects structure for each main hypothesis. 'Noise Type' refers to whether the speech was played without noise or with speech-shaped noise. Stimuli Type refers to whether the speech was audio-only or audiovisual. SOA level refers to the level of stimulus onset asynchrony introduced with the speech (0, 200, 216.6, 233.3, 250, or 266.6 ms). ID refers to the participant ID. Age refers to the participant's age at the time of testing. Speech token refers to the single-syllable stimuli used in each trial (Ba, Fa, or Ka).

Figure 1.

Violin plots showing the accuracy rates and reaction times of participants when speech was presented either with or without noise, for both audio-only (AO) and audiovisual (AV) stimuli. Boxplots show the median and interquartile ranges for each condition.

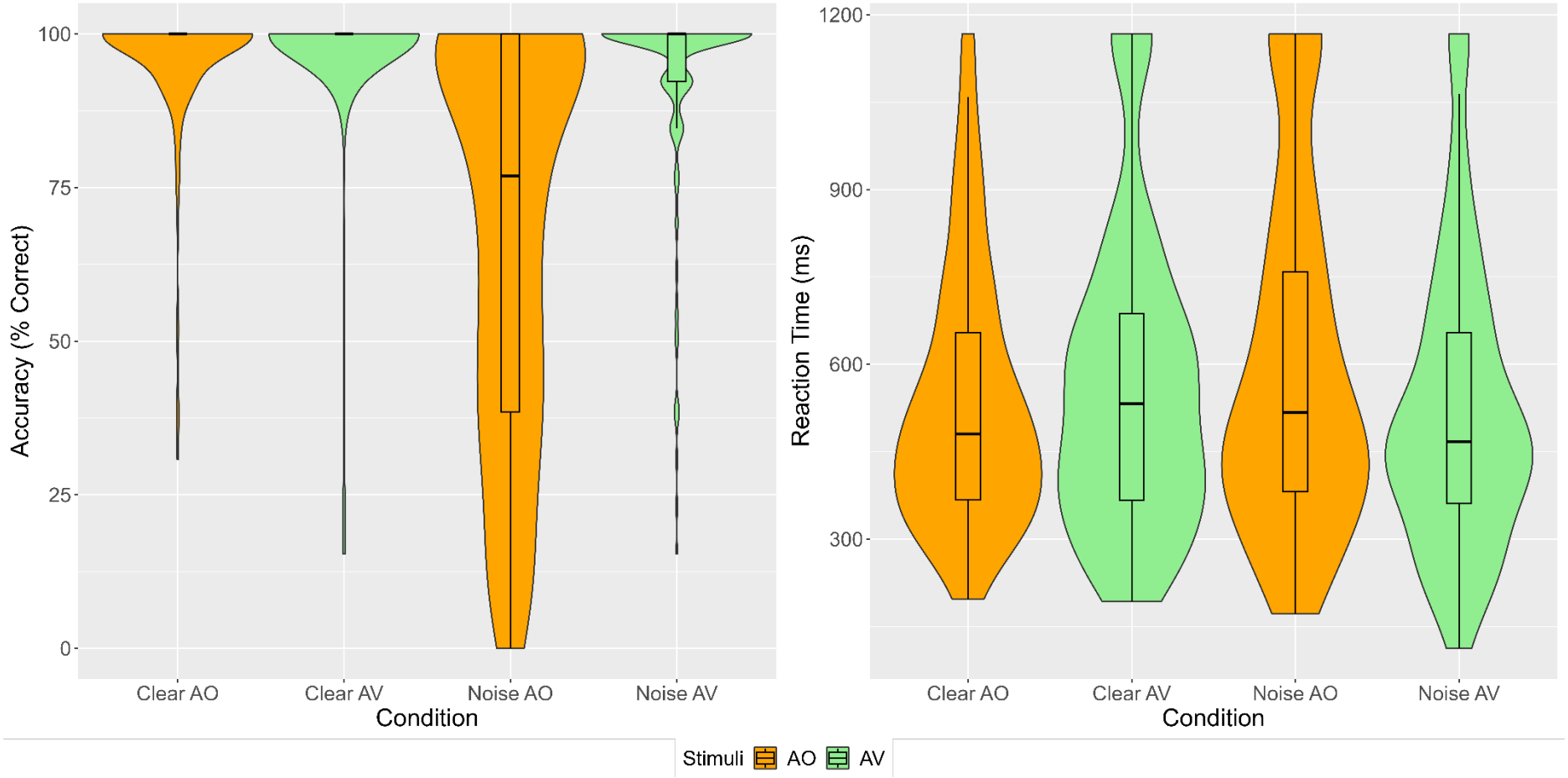


Figure 2.

Violin plots showing the accuracy rates and reaction times of participants when audiovisual stimuli were presented in noise at different SOAs. Boxplots show the median and interquartile ranges for each condition.

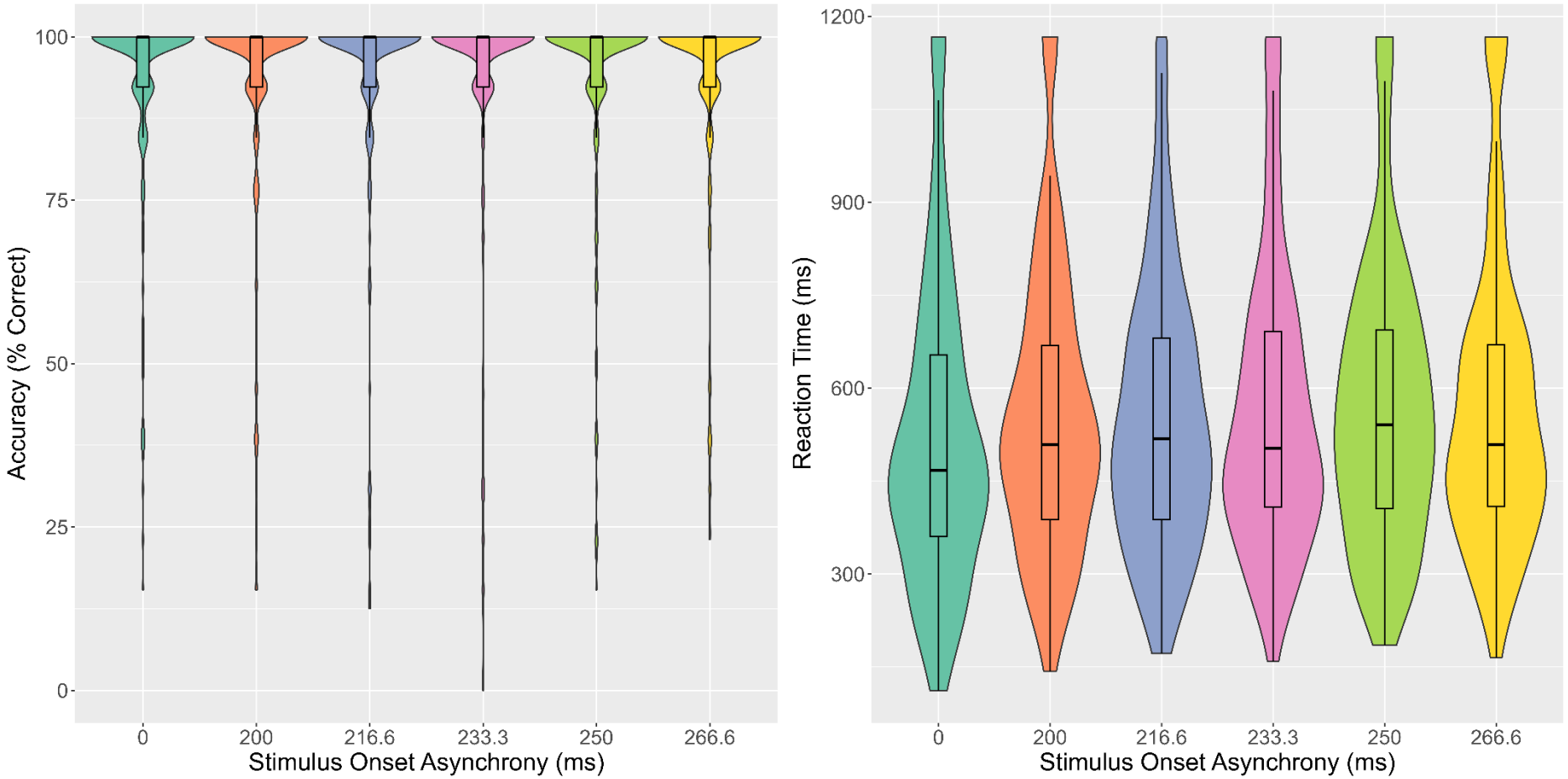


Figure 3.

Violin plots showing the accuracy rates and reaction times of participants when audiovisual stimuli were presented without noise at different SOAs. Boxplots show the median and interquartile ranges for each condition.

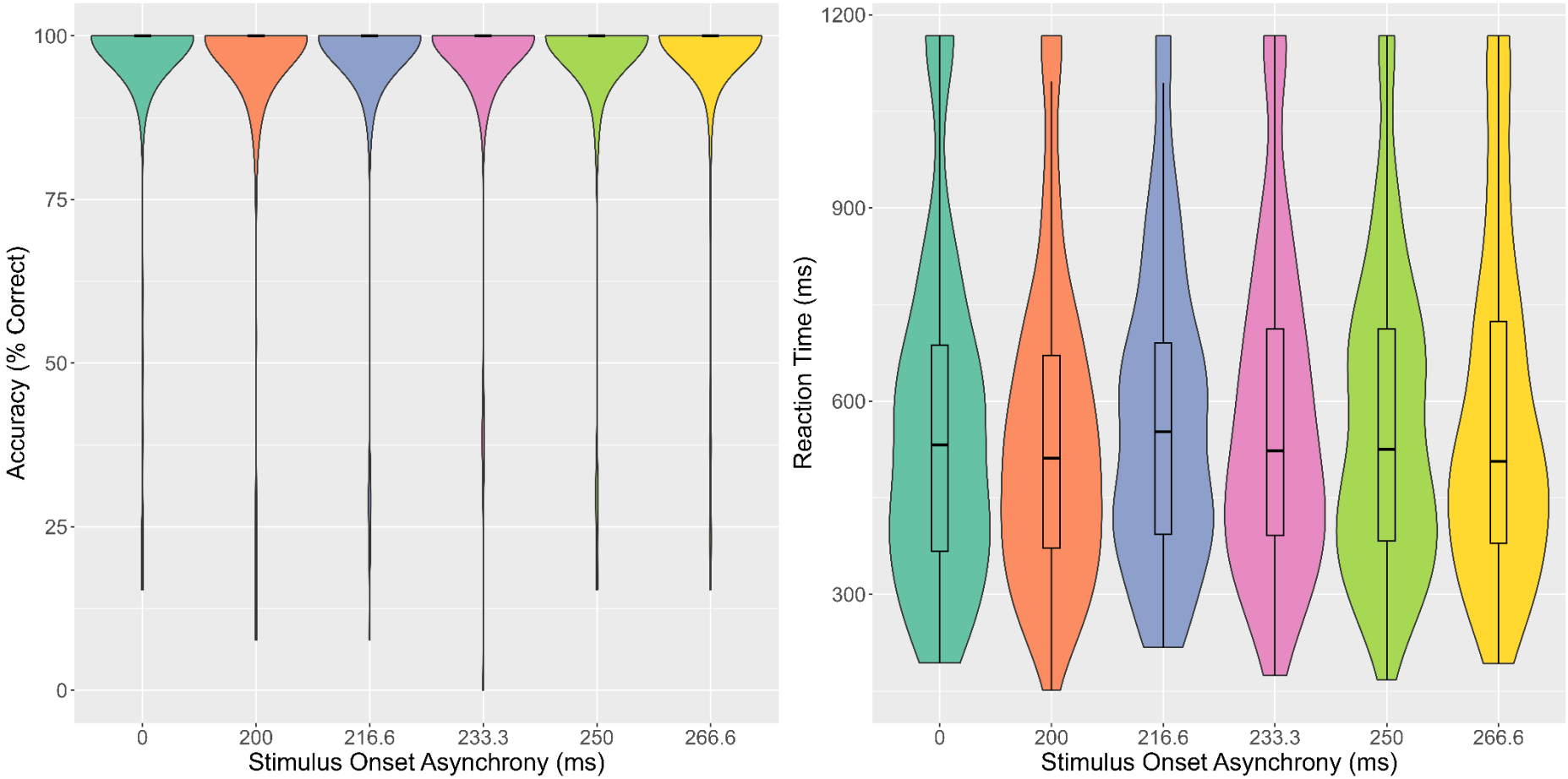
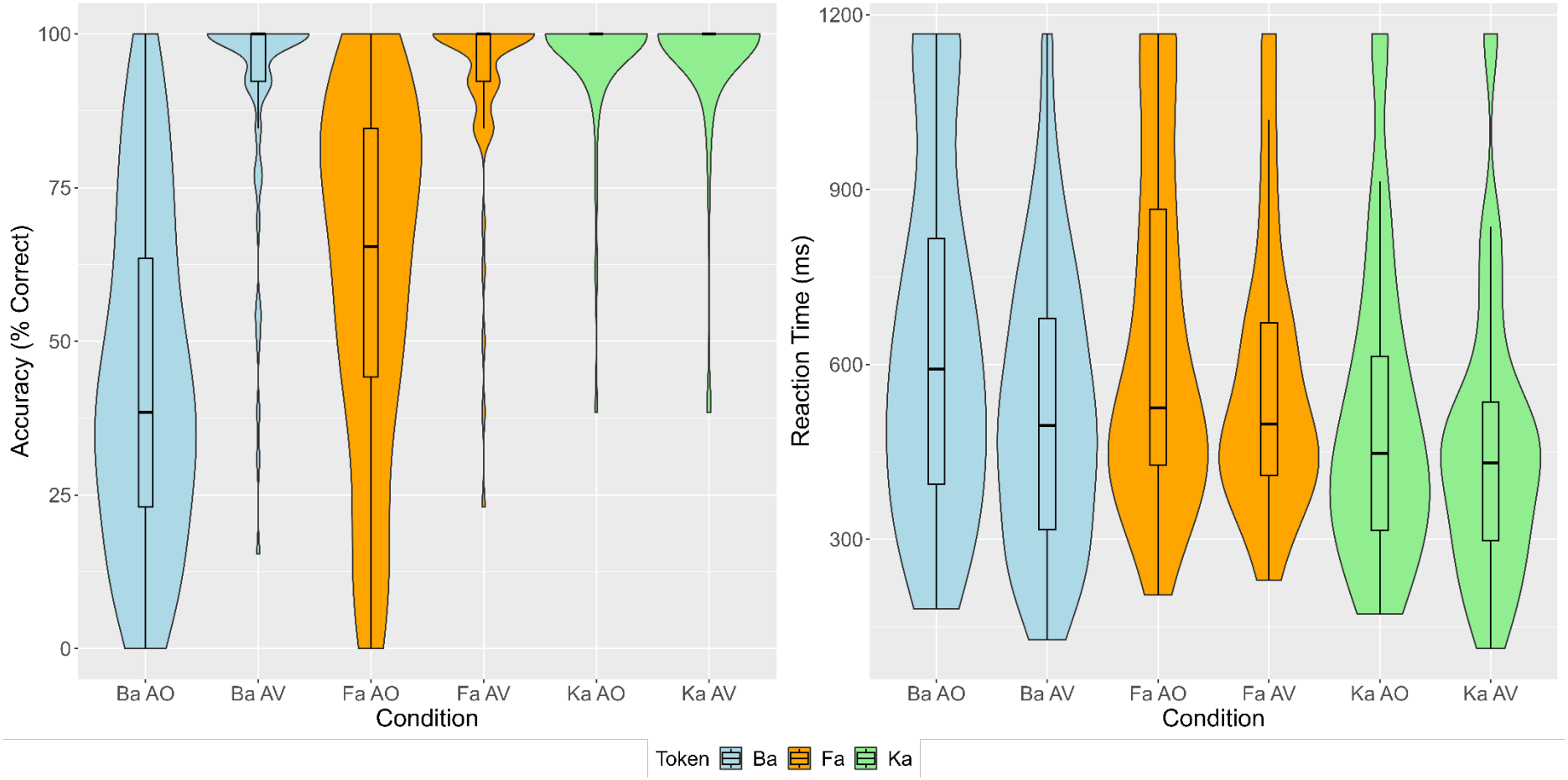


Figure 4.

Violin plots showing the accuracy rates and reaction times of participants when speech tokens were investigated individually in noise for both Audio-Only (AO) and Audiovisual (AV) stimuli. Boxplots show the median and interquartile ranges for each condition.



References

Altieri, N., Townsend, J. T., & Wenger, M. J. (2014). A measure for assessing the effects of audiovisual speech integration. *Behavior Research Methods, 46*(2), 406-415.

Audacity Team. (2021). *Audacity(R): Free Audio Editor and Recorder* [Computer application]. Version 3.0.0 retrieved March 17th, 2021, from https://audacityteam.org/ . *Copyright statement: Audacity® software is copyright © 1999-2021 Audacity Team. The name Audacity® is a registered trademark.*

Bates, D., Maechler, M., Bolker, B., & Walker., S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw, 67*(1), 1-48.

Bernstein, L. E., Auer Jr, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1-4), 5-18.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.53, retrieved August 2021 from http://www.praat.org/

Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language, 87*, 71-83.

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition, 2*(1).

Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications, 6*(1), 1-12.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in Integrative Neuroscience, 4*, 1316.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719-721.

Ewen, C. J., & Van der Hulst, H. (2001). *The phonological structure of words: an introduction*. Cambridge University Press.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*(4), 796-804.

Ganesh, A. C., Berthommier, F., & Schwartz, J. L. (2018). Audiovisual binding for speech perception in noise and in aging. *Language Learning, 68*, 193-220.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*(4), 2438-2450.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Krishnamurthy, N., & Hansen, J. H. (2009). Babble noise: modeling, analysis, and applications. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(7), 1394-1407.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in linear mixed effects models* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/lmerTest/index.html

Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in Psychology, 5*, 868.

Lee, J., Han, J., & Lee, H. (2020). Long-Term Musical Training Alters Auditory Cortical Activity to the Frequency Change. *Frontiers in Human Neuroscience, 14*, 329.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 148545.

Lotfi, Y., Chupani, J., Javanbakht, M., & Bakhshi, E. (2019). Evaluation of speech perception in noise in Kurd-Persian bilinguals. *Auditory and Vestibular Research, 28*(1), 36-41.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494-1502.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One, 4*(3), e4638.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. Journal of Experimental Psychology: *Human Perception and Performance, 37*(1), 245.

Massaro, D. W., Cohen, M. M., Tabain, M., & Beskow, J. (2012). Animated speech: Research progress and applications In Clark RB, Perrier J, P, & Vatikiotis-Bateson E (Eds.), Audiovisual Speech Processing (pp. 246–272). *Cambridge: Cambridge University*.

McCloy, D. (2021). *Praat Script: 'Mix speech with noise'* [Praat script]. LICENSED UNDER THE GNU GENERAL PUBLIC LICENSE v3.0: http://www.gnu.org/licenses/gpl.html

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Megnin-Viggars, O., & Goswami, U. (2013). Audiovisual perception of noise vocoded speech in dyslexic and non-dyslexic adults: the role of low-frequency visual modulations. *Brain and Language, 124*(2), 165-173.

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *Truncated normal distribution* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/truncnorm/

Michalek, A. M., Watson, S. M., Ash, I., Ringleb, S., & Raymer, A. (2014). Effects of noise and audiovisual cues on speech processing in adults with and without ADHD. *International Journal of Audiology, 53*(3), 145-152.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507.

Noel, J. P., Stevenson, R. A., & Wallace, M. T. (2018). Atypical audiovisual temporal function in autism and schizophrenia: similar phenotype, different cause. European *Journal of Neuroscience, 47*(10), 1230-1241.

Open Broadcaster Software. (2024). OBS Studio (Version 29.1.3) [Computer software]. https://obsproject.com/

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Peter, G., Catriona, M., & Phillip, A. (2019). *Power analysis for generalised linear mixed models by simulation* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/simr/index.html

Qualtrics. (2005). *Qualtrics software*, Provo, Utah, USA. Copyright@2021, Current version: 09-21. Retrieved from: https://www.qualtrics.com

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ren, Y., Yang, W., Nakahashi, K., Takahashi, S., & Wu, J. (2017). Audiovisual integration delayed by stimulus onset asynchrony between auditory and visual stimuli in older adults. *Perception, 46*(2), 205-218.

Satterthwaite, F. E. (1941). *Synthesis of variance. Psychometrika, 6*(5), 309–316.

Schwartz, J. L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10*(7), e1003743.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology, 5*, 323.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2024).

Analysis of factorial experiments [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/afex/index.html

Smiljanic, R., Keerstock, S., Meemann, K., & Ransom, S. M. (2021). Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *The Journal of the Acoustical Society of America, 149*(6), 4013-4023.

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry, 48*(8), 813-821.

Solberg Økland, H., Todorović, A., Lüttke, C. S., McQueen, J. M., & De Lange, F. P. (2019). Combined predictive effects of sentential and visual constraints in early audiovisual speech processing. *Scientific Reports, 9*(1), 7870.

Sorati, M., & Behne, D. M. (2019). Musical Expertise Affects Audiovisual Speech Perception: Findings From Event-Related Potentials and Inter-trial Phase Coherence. *Frontiers in Psychology, 10*, 2562.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.

Sumby, W., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*, 212-215.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London. *Series B: Biological Sciences, 335*(1273), 71-78.

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & Van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology, 4*, 331.

Van de Rijt, L. P., Roye, A., Mylanus, E. A., Van Opstal, A. J., & Van Wanrooij, M. M. (2019). The principle of inverse effectiveness in audiovisual speech perception. *Frontiers in Human Neuroscience, 13*, 335.

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America, 152*(6), 3216-3225.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics, 79*, 396-403.

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015). How musical expertise shapes speech perception: Evidence from auditory classification images. *Scientific Reports, 5*(1), 14489.

Yi, H., Pingsterhaus, A., & Song, W. (2021). Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Frontiers in Psychology,* 12.

Yuan, Y., Lleo, Y., Daniel, R., White, A., & Oh, Y. (2021). The impact of temporally coherent visual cues on speech perception in complex auditory environments. *Frontiers in Neuroscience, 15*, 678029.

Chapter 3

3 Effects of Short-Term Audio-Tactile Training on Cortical Speech-Envelope Tracking and Speech Intelligibility

Linking Statement:

Following the previous chapter, we have established that audio-visual speech still benefits our understanding of speech in difficult listening conditions. However, visual lipreading is not always accessible to us in our environment. In these cases, it may be possible to utilise the tactile sense in a similar manner to visual integration. This chapter investigates the contributions of audio-tactile integration to speech perception and processing. Specifically, the chapter discusses how short-term training with speech-relevant tactile stimulation may lead to neural speech tracking enhancement on a cortical level, as well as behavioural benefit through increased speech intelligibility.

Author Note: *This work was produced in collaboration Dr. Helen Nuttall, Prof. Christopher Plack, Prof. Lars Hausfeld, Prof. Lars Riecke, and technician Barrie Usherwood. This paper is currently under review from the wider collaborative team before planned submission to the Journal of Neuroscience for publication.*

Statement of Authorship

Chapter 3 (Paper Two):	Effects of Short-Term Audio-Tactile Training on Cortical Speech-Envelope Tracking and Speech Intelligibility
Authors:	Brandon O'Hanlon, Lars Hausfeld, Lars Riecke, Barrie Usherwood, Christopher J Plack, and Helen E Nuttall
Publication status:	Published Publication has been accepted Publication has been submitted **Unpublished/Unsubmitted but in manuscript style**
Reference:	Not published.
Student/Principle Author:	Brandon Lee O'Hanlon
Contribution:	Theoretical conceptualization; study design; data collection; statistical data analysis; manuscript development; manuscript revisions based on supervisor feedback.

Principle Author Signature: Date: 14.04.2025

Through signing this statement, the Co-authors agree that:

(a) The student's contribution to the above papers is correct

(b) The student can incorporate this paper within the thesis

(c) The contribution of all co-authors for each paper equals 100% minus the contribution of the student.

Co-author Name:	Lars Hausfeld
Co-author Contributions:	Collaborator. Contributed theoretical knowledge and advice on cortical speech-envelope tracking methodology and audio-tactile training design.
Co-author Signature:	Date: 16/04/2025

Co-author Name:	Lars Riecke
Co-author Contributions:	Collaborator. Contributed theoretical knowledge and advice on cortical speech-envelope tracking methodology and audio-tactile training design.
Co-author Signature:	Date: 17/04/2025

Co-author Name:	Barrie Usherwood
Co-author Contributions:	Collaborator. Development and creation of audio-tactile stimulation device.
Co-author Signature:	Date: 16/04/2025

Co-author Name:	Christopher J Plack
Co-author Contributions:	Co-supervisor. Contributed to manuscript draft and final manuscript revisions.
Co-author Signature:	Date: 25/04/2025

Co-author Name:	Helen E Nuttall

Co-author Contributions:	Primary supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 28/04/2025

3.1 Abstract

The auditory cortex tracks speech by synchronising neural activations with fluctuations in the speech envelope. Visual lipreading can improve speech-envelope tracking and intelligibility. Speech-relevant tactile information also improves tracking, but not intelligibility. We hypothesised that this is because we are not exposed to speech-relevant tactile information in our environment. Audio-tactile training may improve intelligibility when tactile information is available, providing crucial sensory aid to understanding speech-in-noise when visual lipreading is inaccessible. Data from 64 young participants (ages: 18-29; 21 males, 42 females, one non-binary) were collected over five EEG sessions. Participants were given a sentence-in-noise recognition task with audio-tactile and audio-only stimuli. They then received training with either tactile information that was congruent with sentences heard (trained group) or incongruent (pseudo-trained group), with feedback after trials. After completing three training sessions, they completed the speech-in-noise recognition task again in a fourth session. Two weeks later, they returned for a follow-up session. Results showed a significant effect of the session (pre- or post-training) on speech intelligibility, but no significant main effect of group or stimulus. At baseline, there was a significant increase in speech-envelope tracking accuracy with audio-tactile stimuli relative to audio-only, suggesting more accurate neural representation of the speech occurred during listening. After training, there was no benefit to congruent training for audio-tactile tracking but there was an enhancement of audio-only tracking with incongruent training. This suggests that speech intelligibility and tracking are not enhanced by short-term audio-tactile training. The unexpected enhancement of non-tactile tracking provides further evidence against the assumed link between tracking and intelligibility.

3.2 Effects of Short-Term Training with Audio-Tactile Stimulation on Cortical Speech-Envelope Tracking and Speech Intelligibility

The speech envelope refers to the slow temporal fluctuations in the overall amplitude of a speech signal. The human auditory cortex can track the speech envelope through phase-locking, wherein neurons fire action potentials in synchrony with the fluctuations in the envelope (Heil & Peterson, 2015; Issa *et al.,* 2024), particularly in the delta and theta ranges (Bröhl & Kayser, 2020; Etard & Reichenbach, 2019). The neural accuracy of speech-envelope tracking is assumed to be linked to speech intelligibility, such that greater tracking accuracy associates with greater speech intelligibility (Kong *et al*., 2015; Vanthornhout *et al.*, 2019). This intrinsic link is supported by speech-in-noise studies, where decreases in tracking accuracy occur as more noise is added to a speech signal causing reductions in intelligibility (An *et al*., 2023). Further evidence of this link between measures comes from audiovisual integration. We are adept at integrating auditory and visual information through lipreading to improve intelligibility (Maier *et al.*, 2011). This visual aid also benefits neural tracking (Golumbic *et al.*, 2013), with recent studies showing that facemask-wearing reduces tracking accuracies for audiovisual speech but not for audio-only speech (Haider *et al.*, 2024).

However, relevant visual information is not always present when listening. In these cases, the sense of touch may be more useful for speech integration. Riecke *et al*. (2019) investigated audio-tactile integration by providing tactile information shaped by the speech envelope alongside the speech stimuli. Speech was degraded using a 30-channel vocoder to reduce the envelope. Results indicated enhanced neural tracking of degraded speech when speech-shaped tactile stimulation was provided. Despite this neural enhancement, no enhancement of speech intelligibility was observed, which questions the link between tracking and intelligibility. Tactile information may be insufficient to enhance intelligibility when speech is degraded, despite enhancing neural representations of the speech.

Alternatively, this might be indicative of neural tracking serving a different purpose in speech perception. For example, tracking may play a role in the prediction of oncoming speech (see Karas *et al.*, 2019 for an example of audio-visual integration assisting with the prediction of oncoming speech) or in attentional decoding (see: Geirnaert *et al.*, 2021; Geirnaert *et al*., 2024; Straetmans, 2022). This would be consistent with further recent evidence showing no link between the two measures (Kösem *et al.*, 2023). Kösem *et al*. (2023) presented participants with two- and four-band vocoded speech before and after training but only trained participants with four-band vocoded speech. They found that audio-only training improved speech intelligibility for four-band vocoded speech and not two-band. However, no changes in neural envelope tracking were observed post-training. In contrast, Riecke *et al.* (2019) showed audio-tactile benefit to neural tracking without training but with no intelligibility benefit. Here, training with audio-tactile speech may provide speech intelligibility improvements akin to Kösem and colleagues, whilst also further enhancing neural tracking accuracy benefits seen from Riecke and colleagues.

Audiovisual speech may show enhancements in tracking and intelligibility because we are adept at integrating lip movements with speech from as early as six months of age (Pons *et al*., 2009) and develop audiovisual integration throughout childhood (Tye-Murray *et al.*, 2014). It may be that speech-relevant tactile stimulation was insufficient to improve intelligibility because we do not have much exposure or training with audio-tactile stimuli. Auditory perceptual learning has been shown to occur in a 24 – 48-hour training window (Atienza *et al.*, 2002). Even in a multisensory context, audio-tactile training has been shown to have comparable benefits to intelligibility as unisensory contexts, though this was not investigated alongside neural speech tracking accuracy (Cieśla *et al*., 2022). Therefore, the aim of this study was to investigate the effects of short-term audio-tactile training on cortical tracking of the speech envelope and speech intelligibility.

We tested the following primary hypotheses:

(i): Neural tracking of the speech envelope of speech-in-noise will be enhanced with the addition of speech-shaped tactile stimulation, compared to neural tracking of the speech envelope of speech-in-noise stimuli with no additional speech-shaped tactile stimulation.

(ii): Neural tracking of the speech envelope of speech-in-noise with speech-shaped tactile stimulation will be further enhanced for participants that have undergone short-term training with audio-tactile stimulation compared to participants given pseudo-training (exposure to irrelevant tactile information).

(iii): Intelligibility of speech-in-noise with speech-shaped tactile stimulation will be enhanced for participants who have undergone short-term training with audio-tactile stimulation compared to participants given pseudo-training (exposure to irrelevant tactile information).

We also tested the exploratory hypothesis:

(iv): There will be no significant difference in intelligibility of speech-in-noise when audio-tactile stimuli are presented in the follow-up session (two weeks after the post-training task of the experiment is complete) compared to the post-training session for the trained group (trained with relevant tactile stimulation).

3.3 Materials and Methods

3.3.1 Participants

Participants were recruited from the Lancaster University campus and the surrounding Lancaster area. Sixty-seven participants passed the eligibility criteria and were recruited for the study (ages: 18-29; $M_{age} = 20.5$; $SD_{age} = 2.43$; 21 males, 45 females, one non-binary). Of these, 64 participants (ages: 18-29; $M_{age} = 20.5$; $SD_{age} = 2.46$; 21 males, 42 females, one non-binary) completed the study. The remaining three participants did not complete all three training sessions of the study, either due to illness or due to scheduling conflicts with later sessions. These partial data were not included in the analyses. Inclusion criteria were that all participants: were between the ages of 18 and 35, were right-handed only, were monolingual native speakers of British English, and had normal hearing. Hearing was measured using pure tone audiometric air conduction testing, bilaterally across 250 to 8000 Hz frequencies and following the British Society of Audiology guidelines (BSA, 2018). All participants had calculated hearing thresholds below 20 dB HL. Exclusion criteria were that no participants: had nerve damage to the fingertips, motor problems in the hands, missing fingers on their hands, or peripheral neuropathy in the hands or fingertips. Participants were screened for eligibility using Qualtrics before the experiment began. Ethical approval was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (approval reference: FST-2022-0766-RECR-4, project ID: 0766).

3.3.2 Sample Size Calculations

Before testing, data simulation was conducted using R 4.2.2 (R Core Team, 2022) for power and sample size analysis using *Lme4* (v1.1-27.1; Bates et al., 2021), *afex* (v1.0-1; Singmann et al., 2021) and *simr* (v1.0.5; Peter *et al.*, 2019). Means and standard deviations were taken from previous neural tracking multisensory studies for simulation (Ricke *et al.*,

2019; Golumbic *et al.*, 2013), providing an estimated large effect size of $R^2 = .88$. Sixty-four participants were deemed sufficient for a power of .80 at an alpha level of .017.

3.3.3 Experimental Design

Participants took part in five electroencephalography (EEG) sessions. There was one session per day, with the first four sessions being on adjacent days either on a Monday to Thursday or Tuesday to Friday pattern. The fifth session was a follow-up session that took place 14-18 days later. Participants were randomised into either the 'Trained' group or the 'Pseudo-Trained' group before testing began. Throughout the pre-training, post-training, and follow-up tasks, participants were presented with a variety of sentences in noise, both audio-only and audio-tactile. The session structure can be seen in Figure 1. Neural and behavioural data were collected in every session. A 2 (Training Group: Training or Pseudo Training) x 2 (Session: Pre-task, or Post-task) x 2 (Stimulation Type: Audio-tactile Stimulation, or Audio-only Stimulation) mixed-factors design was used to evaluate these data and test our three primary hypotheses. For testing exploratory hypothesis (iv), a 2 (Training Group: Training or Pseudo Training) x 2 (Session: Post-task, or Follow-up Task) x 2 (Stimulation Type: Audio-tactile Stimulation, or Audio-only Stimulation) mixed-factors design was used.

3.3.4 Materials

All target sentence and noise audio were wideband, with sampling rates of 44.1 kHz and bandwidths of up to 22.05 kHz. Four-talker babble noise was created first using sentences from the Clarity Speech Corpus (Graetzer *et al.*, 2022), containing a mixture of male and female voices at random. This was done using a MatLab script that randomly selected four sentences from the corpus of various durations, took a random 4000 ms segment from the selected sentences, and then placed all four segments onto the same waveform track. Ten 4000 ms audio files were created using this script, which were used in experimental

trials. The target sentence stimuli used sentences also from the Clarity Speech Corpus, which in turn were selected from the British National Corpus, with 210 selected in total. However, as these needed to be all from the same speaker and of a similar duration, it was decided that the chosen sentences would be re-recorded using the primary researcher as the target speaker. The sentences were recorded along with a 10-minute segment of Alice in Wonderland that was also spoken by the primary researcher. Recordings were made using a HyperX Quadcast external USB microphone, set to a cardioid polar direction. Target sentences were between 3400 and 3650 ms long and always played 250 ms after the noise began. All sentences and noise files were loaded into Audacity together and normalised to 60 dB SPL. The noise files were then presented in the experiment between -10 and 10 dB SNR at 0.5 dB increments, creating 41 possible difficulty levels relative to the 60 dB SPL target sentences.

The SNR selected for the participant was calculated using a speech-in-noise test. This was a custom adaptation of the QuickSIN test (Etymonic) used to personalise the difficulty of the experiment so that participants could correctly recognise approximately 50% of keywords from target sentences, referred to as the Speech Recognition Threshold (SRT). Typically, during QuickSIN, the participant would be required to verbally respond to full sentences and correctly responded keywords would be recorded and used in SNR calculations. However, in this study, participants discriminated between keywords via a button press (see Procedure). This meant that the participant could select correct keywords by chance, resulting in the experiment being set at a lower SNR and potentially lowering response accuracy down towards chance level. Furthermore, the original QuickSIN sentences were spoken by a female speaker, which can be harder to discern in mixed-gender multi-talker noise than a male speaker such as was used for the experiment's target sentences (Larsby *et al.*, 2015). Therefore, the QuickSIN test was adapted so that the method of response and the target

speaker of the test matched the target speaker of the experiment. The same calculation method was used as follows:

- SRT = Starting dB level + (dB step value/2)) - Total Keywords Correct
- Therefore, SRT = 12.5 - Total Keywords Correct

Tactile stimulation was provided to the right index finger of participants in audio-tactile and training conditions using a lab-built tactile device. This device used a hard drive accentuator to generate small horizontal movements to the finger based on slow fluctuations in electric potential, which were set to match slow fluctuations in the speech envelope. The device used a soft ring around the finger to keep it in place during stimulation and was insulated around the area of contact with participants to reduce potential electrical interference with the EEG. Furthermore, it was attached to a handrest to make stimulation more comfortable throughout the experiment. For the creation of audio-tactile stimuli, speech envelopes were extracted from all 210 sentences using Praat (Boersma, & Weenink, 2024) via Hilbert transformation. These extracted envelopes were then loaded back into Audacity to be normalised. For audio-tactile sentences, the sentence and its respective envelope were combined into a single waveform file. This was done by having the left channel of a stereo track be the sentence audio, and the right channel just the sentence envelope. For audio-only sentences, the sentence was placed in the left channel and the right channel was left silent. When played, the left channel was split to a pair of EEG-compatible insert earphones (Etymonic ER3-14A Ear Tips) which played the audio in mono format to both ears, whilst the right channel was split to the tactile stimulation device. These earphones had transducers that were housed in electromagnetic shielding to prevent stimulus artefacts in the data. For training session trials, the trained group received the audio-tactile sentences. The pseudo-trained group however received audio-tactile sentences that had the audio in the left channel matched with incongruent speech envelopes in the right channel.

An eligibility screening and initial consent form were created and hosted using Qualtrics (Qualtrics, 2005). The experiment was designed using PsychoPy Builder (v2022.1.4; Peirce *et al.*, 2019) with custom coding elements. Conditions on the pre-task, post-task, and follow-up were counterbalanced, with some participants experiencing the audio-tactile condition first and others the audio-only condition. Sentences were placed into different conditions on different sessions so that no sentences were repeated across the pre-task, training tasks, and post-task. This placement of sentences was randomised into 10 possible sets which were counterbalanced between training groups. To ensure that the same noise file played alongside the same sentences to avoid potential biases with some sentences being harder to understand in some noise files, all sentences were paired with one of the possible 10 noise files. Therefore, when sentences were randomised into sentence sets, they came attached with the same noise files across those sets for consistency. Sentence and noise files were played to the participant with the same onset during the same PsychoPy routine. A 32-channel BrainProducts EEG kit (Brain Products GmbH, Gilching, Germany) was used in all five sessions of the study to record neural activity whilst the participant listened to stimuli. More channels were not needed for reconstruction validity, as shown by Montoya-Martínez *et al.* (2021). The kit utilised a BrainAmp amplifier and recorded brain activity at a 500 Hz sampling rate. ActiCAP 64 channel standard caps were used with only the standard 32-channel 10%-system electrode locations utilised. Cap sizes used ranged from 54 to 60 cm head circumference. For improving impedance values of electrodes, SuperVisc high viscosity electrolyte gel was used during setup. Acceptable impedance value targets were 5 kΩ or less during setup, which was monitored using BrainVision Recorder software (see BrainVision Analyzer, Brain Products GmbH, Gilching, Germany).

3.3.5 Procedure

Participants were provided with a basic overview of the five sessions of the experiment before and during the study but were blinded regarding their group assignment. Participants completed pre-screening via Qualtrics (Qualtrics, 2005). Once participant eligibility was confirmed, participants were invited to the lab. In the first session, participants were familiarised with the lab setup and equipment, baseline measures were taken, they completed a pre-training task, and they completed their first training session. A pure tone audiometry assessment was conducted using a calibrated audiometer and following BSA guidelines (British Audiology Society, 2018). Participants were then introduced to EEG and the tactile device used for the experiment. They were presented with two randomly selected sentences with tactile stimulation. Participants' individual tactile perception threshold was measured by adjusting the force of the tactile device following a staircase threshold method. Participants informed the researcher if they could feel the movements of the device relevant to random speech sentences that played in tandem and if this tactile stimulation was comfortable. If the intensity needed adjusting, it was done by the researcher. Once a threshold was found, the device was set above threshold by one incremental level. This tactile intensity level was recorded and was kept consistent for each participant throughout all their testing sessions.

3.3.6 Speech-in-Noise Test

Participants then took part in the customised speech-in-noise test. To ensure that any potential distracting effects of the EEG cap were considered during the calculation of the participant's individualised SNR, the EEG cap was set up for the speech-in-noise test despite not recording data. Participants were also asked to place their right index finger into the tactile stimulation device. Again, this device was not switched on during this speech-in-noise test. However, having the index finger of their dominant hand placed on an unusual device

may affect speech discrimination performance in later tasks, and so was considered when calculating the participant's SRT. The researcher provided the verbal instructions for the test alongside written instructions that the participant had to read and click through to proceed to the task. Participants were told that they were to listen for the researcher's voice and to ignore the background noise of other people speaking at the same time. The researcher was present for every session conducted in the study so that their voice as the target sentence was recognisable. The same example sentence was given to all participants: 'The dog ran down the long road'. They were then told that after hearing the sentence in noise, four keywords would pop up on the screen, one in each corner. One of these keywords was the first keyword heard in the target sentence, whilst the other three were semantically or phonetically similar. In the context of the example target sentences, the correct choice was clarified to participants as 'dog', and other options may be 'god' or 'cat'. They were told to press the corners of the number pad to select the corners of the screen. As an example, if 'dog' was in the top right, they would press the top right of the lab keyboard's number pad, which was always '9'. If it was in the bottom left, they were told to press '1'. As explained to the participant, once the selection was made, the four keywords would disappear and four more would show up. One of these would be the second keyword, 'ran'. Participants would keep making selections until all five sets of keywords were presented. After this, the next target sentence would play.

For the speech-in-noise test, there were six trials in each block: the SNRs were always: 10, 5, 0, -5, -10, and -15 dB. Participants were told to guess if they were unsure. There were five blocks in total. The first two blocks acted as practice and familiarisation for the participant. The last three blocks were used to calculate the participant's SRT for the experiment. This was done by summing the number of correctly discriminated keywords from each sentence per block and subtracting that value from 12.5. This was based on the formula from the QuickSIN method (see Materials). As this was done for each of the three

test blocks, the average SRT was calculated between them to provide the final SNR value that was then used for every session with the participant moving forward.

3.3.7 Pre-Training Session

The pre-training task acted as the baseline and consisted of three elements: a short passive listening task using a story excerpt from 'Alice in Wonderland' (see 'Materials'), a block of 30 speech sentence discrimination trials in noise that were audio-only, and a block of 30 sentences in noise that were audio-tactile. The order of the audio-only and audio-tactile blocks was counterbalanced across participants, whilst the story was always presented as the first part of the task. The story segment was presented without noise and was split into two five-minute parts, with a break and a content question after each to ensure participants were paying attention. This was a multiple-choice answer question, such as 'What did the rabbit pull out of its waistcoat?', with four answers to select from using a mouse click. Participants could not proceed without selecting the correct answer. In a case where participants failed to select the correct answer, they listened to the segment again with the same content question. EEG data were recorded throughout this task. Once the story task finished, the researcher would let the participants know whether they would be next listening to audio-tactile or audio-only sentences. This was so that participants were not surprised if the device was active during sentence listening. Participants were reminded at this stage to try not to blink during speech listening and to remain as comfortable as possible. The audio-only and audio-tactile sentence tasks worked the same as the speech-in-noise test, except that the SNR for all sentences was set based on the participant's SRT level. Furthermore, after each sentence trial, the participant was given a voluntary break screen and a progress bar indicating how many sentences were left for the block. These were implemented to ensure that participants had ample opportunity to get comfortable and take a break as they needed, as well as an opportunity for the researcher to adjust any electrodes that may have gone noisy mid-testing.

3.3.8 Training Sessions

After the pre-training task was completed, participants moved on to their first training task for the study. This training task involved another 30 sentences in noise. This time, after all five keywords were responded to, participants were given feedback on their performance. This was in the form of seeing how many keywords they correctly identified in that trial, hearing the target sentence in full without noise and with the same tactile stimulation they received in that trial, and seeing the sentence written out in full whilst listening to it. Secondly, whilst the trained group received tactile stimulation that was relevant to the target speech sentence's envelope, the pseudo-trained group received tactile stimulation that was relevant to the speech envelope of a different sentence. Their feedback also reflected this, replaying the same incorrect tactile stimulation. The training tasks took approximately 20 minutes to complete, though this could vary depending on the amount of break time a participant may have taken between trials. For the second and third sessions, participants were set up in EEG again and completed their next training tasks. There were three training tasks in total and the sessions were set up the same as the session one's training task.

3.3.9 Post-Training and Follow-Up Sessions

For the fourth session, the participants only took part in a post-training task. This mimicked the pre-training task in that they were given the same 'Alice in Wonderland' story segment to listen to in two halves, with different content questions to the first session, and then presented with the audio-tactile and audio-only sentence blocks in a counterbalanced order. Finally, for the fifth session, participants came back two weeks later to again complete the story listening task and both audio-tactile and audio-only sentence blocks. Whilst this was booked to be always on the Friday two weeks after the end of the fourth session, sometimes the participant was unable to make the session. In these cases, the follow-up session occurred on the following day or – at the latest – the following Monday. At the end of the fifth session,

participants were debriefed on the true aims of the study and paid to compensate them for their time. In all, the study provided neural and behavioural data across all five sessions.

3.4 Statistical Analyses

3.4.1 Variables

There were three independent variables. The training group variable refers to whether the participant was placed in the trained group or the pseudo-trained group. The session variable refers to which of the five sessions that data were collected during. In the case of the analysis listed below, these were the pre-training task and the post-training task data. For exploratory analyses and data figures, we also looked at the follow-up task and all training task data. Finally, the stimulation type variable referred to whether the condition was with audio-only sentences in noise or audio-tactile sentences in noise. Stimulation type and session were within-subject factors, with participants taking part in every session and with both audio-only and audio-tactile conditions. The training group was a between-subjects factor, with 32 participants placed in the trained group and 32 in the pseudo-trained group. Finally, the retention session variable was a within-subject factor used in the exploratory models for testing hypothesis (iv) to differentiate between the post-training session and the two-week follow up sessions specifically. For dependent variables, both speech intelligibility (SI) and cortical speech-envelope tracking accuracy (Rz) were measured. SI was defined as the percentage of correctly discriminated keywords in a sentence trial, which was averaged over all 30 sentence trials in a condition. To calculate Rz, the multivariate temporal response function toolbox (mTRF, Crosse *et al.*, 2016) was used. This process involved utilising a decoder function in the mTRF toolbox to reconstruct an estimation of the target sentence speech envelope based on the inputs of collected neural data and then correlate this estimated envelope with the original stimulus envelope. This correlation was used as the measure of Rz.

3.4.2 Missing Data

One participant was unable to attend the post-training session (pseudo-trained group). As they still returned for the two-week follow-up session, their missing data was estimated. The group mean change from pre- to post-training for the audio-tactile and audio-only conditions was added to the participant's baseline (pre-training) scores, providing an estimation of the post-training effect relevant to their individual baseline.

3.4.3 Pre-processing and Decoding

EEG data was pre-processed using EEGLab (Delorme & Makeig, 2004) in MatLab. Initially, the data were recorded at a sampling rate of 500 Hz and online filtered across the frequency ranges of 0.1 and 44 Hz to keep file sizes efficient. Data were recorded throughout each condition, with a new recording file being made per condition. Using EEGLab, the data were first resampled to 100 Hz and filtered using a Finite Impulse Response (FIR) filter with a low pass at 1 Hz, before independent components analysis (ICA) was run. The spheres and weight matrices outputted by the ICA were saved to be used for a future decomposition. This method of early ICA was selected as our target frequency range of 0.5 – 15 Hz included delta below 1 Hz, which is susceptible to slow-drift distortion with extended infomax ICA (Pontifex *et al.*, 2017). The raw data were then reloaded back into EEGLab and resampled to 100 Hz again. The data were filtered to our target range next using a FIR filter, with a low pass at 15 Hz and a high pass at 0.5 Hz. Next, the data were re-referenced using the average. The previously decomposed ICA weights and spheres that were determined from the first loading of the data were then placed on this second iteration. ICLabel (Pion-Tonachini *et al.,* 2019) was used to automatically flag components for muscle, eye, heart, line-based, and channel-based noise, with boundaries for all set at 85%. These flagged components were then removed before finally the stimulus presentation periods were extracted using the onset and offset of each sentence file played. To remove the onset of event-related potentials, the first

second of each sentence trial was removed. The result was a three-second epoch per trial. For the 'Alice in Wonderland' story segments, the same pre-processing steps were used. However, the first two seconds of each five-minute segment were removed instead, resulting in two 298-second epochs.

3.4.4 Speech-envelope Tracking Accuracy (Rz)

Rz was obtained using the stimulus reconstruction method via the multivariate Temporal Response Function toolbox in MatLab (see Crosse *et al.,*2016). This method of reconstruction uses a backwards approach with a decoder for the neural data. For cross-validation, the method of 'leave-one-trial-out' was chosen (see Riecke *et al.*, 2019). As we were using a low SNR that matched participants' individual SRTs, outputs of the reconstruction method were expected to be lower for sentence trials than in previous literature. Furthermore, due to the quicker sentence duration, each sentence trial could not provide enough EEG data alone for valid reconstruction. The required amount of EEG data for valid envelope reconstruction is not entirely clear in the literature, with some referencing 60 seconds as sufficient for 87.5% accuracy (Biesmans, *et al.*, 2016). A comparative look between EEG and Magnetoencephalography (MEG) suggests that EEG requires as much as three times the duration of MEG for valid reconstruction, coming to approximately 120 seconds (Destoky, *et al.*, 2019). It is essential to provide as much EEG data to the decoder as possible, with a minimum duration of somewhere between 60 and 120 seconds in mind. This meant that for all 30 sentences in a condition, we would need every epoch available in that condition to be combined for a more reliable reconstruction. By stitching together epochs, however, we run the risk of training the decoder on the 'seams' of the individual epochs, which may provide inefficient decoder parameters when it comes to calculating the final speech-envelope tracking accuracy value. To alleviate this issue and the issue of low SNR during sentence listening, the two five-minute story segments in clear speech were used to

train the decoder first and output optimal parameters for the reconstruction of the 30 stitched-together sentences. This provided an optimal regularisation parameter (lambda) and number of 'folds' or 'segments' (nf), which were then applied as the parameters for sentence reconstruction. Reconstruction outputs were averaged across all leave-one-trial-out validations to provide a final Rz value for each session's conditions.

3.4.5 Models for Analysing Neural Data

For testing hypotheses (i) and (ii), linear mixed-effects models (LMERs) were used taking Rz as the dependent variable. For our first hypothesis, we expect that tracking accuracy would see enhancement with audio-tactile speech versus audio-only speech, regardless of group, when looking at baseline data only. The ID of the participants and the sentence list assigned to them were loaded as random factors. The LMER model was as follows:

Rz ~ Group + Stimulation Type + Group*Stimulation Type + (1|ID) + (1|sentence))

To accept this hypothesis, we would expect to see a significant main effect of stimulation type (audio-tactile or audio-only). For the second hypothesis, we expect to find that Rz scores will increase post-training with audio-tactile speech for the trained group, but not for the pseudo-trained group. This would be looking at both pre- and post-training data. The ID of the participants and the sentence list assigned to them were loaded as random factors. The LMER model was as follows:

Rz ~ Group + Stimulation Type + Session + Group*Stimulation Type + Group*Session + Stimulation Type*Session + Group*Stimulation Type*Session + (1|ID) + (1|sentence))

To accept this hypothesis, we would expect to see a significant interaction effect between the three independent variables. The model would then be split by the group variable

and the two-way interaction between session and stimulation type would be assessed for both the trained group and the pseudo-trained group. If again significant, a pairwise comparison test should then signify that this interaction is significant for the trained group receiving training benefits to tracking with audio-tactile versus audio-only speech.

3.4.6 Models for Analysing Behavioural Data

For testing hypothesis (iii), a generalised LMER model (GLMER) was used as our accuracy scores were bound based on choice-selection in the speech discrimination task. We expect that speech intelligibility will increase post-training with audiotactile speech for the trained group, but not for the pseudo-trained group. This would be looking at both pre- and post-training data. The ID of the participants and the sentence list assigned to them were loaded as random factors. The GLMER model was as follows:

SI ~ Group + Stimulation Type + Session + Group*Stimulation Type + Group*Session + Stimulation Type*Session + Group*Stimulation Type*Session + (1|ID) + (1|sentence))

To accept this hypothesis, we would expect to see a significant interaction effect between the three independent variables. The model would then be split by the group variable and the two-way interaction between session and stimulation type would be assessed for both the trained group and the pseudo-trained group. If again significant, a pairwise comparison test should then signify that this interaction is significant for the trained group receiving training benefits to intelligibility with audio-tactile versus audio-only speech.

3.4.7 Models for Exploratory Analyses

Hypothesis (iv) is a preregistered exploratory analysis. This was done because we were interested in understanding if any benefits of audio-tactile training to speech intelligibility were retained after a short period of two weeks. However, due to study

restrictions, we could not increase our sample size to accommodate further levels in our main hypothesis models. Furthermore, there may not be a potential benefit to training that could be retained, thus it did not feel appropriate to increase resources to test this as a main hypothesis. In this regard, any results drawn from testing this hypothesis and other exploratory analyses are not sufficiently powered and therefore should not be conclusive without further testing. The following models were used on the post-training and follow-up session data:

LMER: Rz ~ Group + Stimulation Type + Retention Session + Group*Stimulation Type + Group*Session + Stimulation Type*Session + Group*Stimulation Type*Session + (1|ID) + (1|sentence))

GLMER: SI ~ Group + Stimulation Type + Retention Session + Group*Stimulation Type + Group*Session + Stimulation Type*Session + Group*Stimulation Type*Session + (1|ID) + (1|sentence))

We expect that any benefits of audio-tactile training on speech intelligibility will be retained after this short two-week period. This will be reflected with no significant effect of the group, stimulation type, or retention session variables on neural tracking accuracy and on speech intelligibility, and no significant interactions between variables. This will offer insight into potential short-term retention of any neural benefits to audio-tactile training. Furthermore, an additional exploratory correlational analysis was conducted between the behavioural and neural dependent variables across all sessions through Pearson's R. This provided further insight into the assumed link between the two dependent variables across all sessions.

3.4.8 Pre-registration and Deviations from Pre-registration

The study was pre-registered on the Open Science Framework (OSF) before data collection began. Further details on the data simulation methodology and the preregistration

itself can be found at: https://osf.io/9fehp. In the pre-registration, the sample size was calculated as 64 participants plus six more to account for attrition. It was deemed not necessary to continue testing to 70 full datasets as a sufficiently powered sample size was achieved with little attrition. Furthermore, the inference criteria were originally listed as $p < .05$ for determining significance. However, as we will be testing three main hypotheses, we will be using the Bonferroni-Holm method to reduce family-wise error rates (banded as: $p < .017$, $p < .025$, and $p < .05$). Additionally, a separate GLMER model for testing hypothesis (i) was added. This was to separate this hypothesis test from post-training variables in the main GLMER model proposed for hypothesis (ii) (see Statistical Analyses).

3.5 Results

3.5.1 Effect of Tactile Stimulation on Speech-Envelope Tracking Accuracy

Figure 2 shows the mean neural tracking accuracy of audio-tactile (mean Rz = .21) and audio-only (mean Rz = .19) sentences before training, collapsed across the trained and pseudo-trained groups. The linear mixed-effects model (LMER) used to test hypothesis (i) indicated a significant main effect of stimuli type in session 1 before training, with audio-tactile speech increasing speech-envelope tracking accuracy (β = .06, t = 2.92, 95% *CI* = [.02, .10], p = .004), remaining significant after Bonferroni-Holm correction (p < .017) and supporting hypothesis (i).

3.5.2 Effect of Short-Term Training with Tactile Stimulation on Speech-Envelope Tracking Accuracy

Figure 3 shows the difference in mean neural tracking accuracies between the post- and pre-training tasks, for both audio-tactile and audio-only sentences, in the trained group (audio-tactile mean = +.05, audio-only mean = +.01) and the pseudo-trained group (audio-tactile mean = +.01, audio-only mean = +.09). The LMER analysis for testing hypothesis (ii) indicated a significant main effect of session (pre-training or post-training; β = .09, t = 3.68, 95% *CI* = [.04, .13], p < .001), showing an enhancement of neural tracking accuracy with training across all conditions and groups. There was also a significant effect of stimulation type (audio-tactile or audio-only; β = .06, t = 2.46, 95% *CI* = [.01, .10], p = .015), again showing further enhancement of tracking accuracy with the introduction of speech-relevant tactile stimulation. The main effect of group was not significant (trained or pseudo-trained; β = .04, t = 1.71, 95% *CI* = [-.006, .09], p = .09). Importantly for testing hypothesis (ii), there was a significant three-way interaction between session, group, and stimulation type (β = .11,

t = 2.31, 95% CI = [.02, .20], p = .022), which remained significant after Bonferroni-Holm correction (p < .025). This provides support for hypothesis (ii).

However, when splitting this three-way interaction by group to determine the direction of this significant interaction, there was a significant two-way interaction between session and stimulation type for the pseudo-trained group (β = -.07, t = -2.23, 95% CI = [-.13, -.01], p = .028) but not for the trained group (β = .04, t = 1.05, 95% CI = [-.03, .10], p = .30). Further pairwise comparisons of the significant interaction for the pseudo-trained group show a significant increase post-training in tracking for audio-only speech (β = .09, z = 3.82, 95% CI = [.03, .15], p < .001). These results go against predictions that the trained group would see significant increases in tracking accuracy post-training for audio-tactile speech. Therefore, hypothesis (ii) cannot be accepted.

3.5.3 Effect of Short-Term Training with Tactile Stimulation on Speech Intelligibility

Figure 4 shows the difference in mean speech intelligibility scores between the post- and pre-training tasks, for both audio-tactile and audio-only sentences, in the trained group (audio-tactile mean = +5%, audio-only mean = +5%) and the pseudo-trained group (audio-tactile mean = +6%, audio-only mean = +7%). The GLMER analysis for testing hypothesis (iii) indicated a significant main effect of session (pre-training or post-training; β = .07, t = 6.30, 95% CI = [.05, .09], p < .017). There was no significant main effect of group (trained or pseudo-trained; β = -.002, t = -.07, 95% CI = [-.05, .05], p > .05), or stimulation type (audio-tactile or audio-only; β = -.01, t = -.43, 95% CI = [-.04, .02], p > .05). The three-way interaction between session, group, and stimulation type was also not significant (β = .03, t = 1.53, 95% CI = [-.01, .07], p > .05). Hence, these data do not provide support for hypothesis (iii).

3.5.4 Exploratory Analyses

Figure 5 shows the difference in mean neural tracking accuracies between the follow-up and post-training tasks, for both audio-tactile and audio-only sentences, in the trained group (audio-tactile mean = +.02, audio-only mean = +.001) and the pseudo-trained group (audio-tactile mean = -.003, audio-only mean = -.02). Figure 6 shows the difference in mean speech intelligibility between the follow-up and post-training tasks, for both audio-tactile and audio-only sentences, in the trained group (audio-tactile mean = -.8%, audio-only mean = -.3%) and the pseudo-trained group (audio-tactile mean = +.8%, audio-only mean = -1%). The planned exploratory LMER analysis for testing hypothesis (iv) was conducted. There was no significant effect of session (β = -.02, t = -.99, 95% CI = [-.06, .02], p > .05), group (β = -.04, t = -1.74, 95% CI = [-.08, .004], p > .05), stimulation type (β = -.01, t = -.70, 95% CI = [-.06, .03], p > .05), or any significant three-way interaction (β = -.002, t = -.05, 95% CI = [-.08, .08], p > .05) on speech-envelope tracking accuracy. For testing speech intelligibility, the GLMER analysis was conducted. There was no significant effect of session (β = -.01, t = -.94, 95% CI = [-.04, .01], p > .05), group (β = -.03, t = -.86, 95% CI = [-.04, .01], p > .05), stimulation type (β = -.02, t = -1.25, 95% CI = [-.09, .03], p > .05), or any significant three-way interaction (β = -.03, t = -.98, 95% CI = [-.08, .03], p > .05) on speech intelligibility.

Therefore, the data support our exploratory hypothesis (iv), in that the general training effects across all conditions and groups for both neural tracking accuracy and speech intelligibility were retained two weeks later, along with the enhancement of tracking in the pseudo-trained group for audio-only sentences. For the final exploratory analysis, Pearson's R correlational analyses between intelligibility and tracking accuracies found no significant relation between the two outputs (r = -.05, 95% CI = [-.17, .07], p > .05), further highlighting a discrepancy between the supposed intrinsic link and the role of neural tracking in speech perception.

3.6 Discussion

This experiment investigated the effects of short-term audio-tactile training on cortical speech-envelope tracking accuracy and speech intelligibility. Replicating previous research (Riecke *et al.,* 2019), we found that, at baseline, audio-tactile speech was associated with greater cortical speech-envelope tracking accuracy relative to audio-only speech, whilst speech intelligibility remained the same. This finding supports our first experimental hypothesis. After short-term audio-tactile training, speech-envelope tracking was not enhanced relative to the pseudo-trained group. This finding does not provide support for our second experimental hypothesis. Finally, there was no enhancement of speech intelligibility following audio-tactile training relative to the pseudo-trained group, showing no support for our third hypothesis. In summary, we observed initial neural benefits of audio-tactile integration to speech-envelope tracking, but these audio-tactile benefits were not enhanced with short-term training. Furthermore, there was no evidence for behavioural benefits of audio-tactile speech to speech intelligibility. These findings contrast with previous research that has reported an intrinsic link between neural speech tracking and intelligibility (Vanthornhout *et al.,* 2018; An *et al*., 2023) and align with recent findings that question the link between tracking and intelligibility (Kösem *et al.*, 2023). These findings raise questions about the role of speech-envelope tracking in speech perception, and our understanding of the neural processing of speech signals in our environment.

Exploratory analyses tested the short-term retention of training effects, as well as relations between the neural and behavioural data, and assessments of listening effort through parietal alpha power. There were no significant differences in tracking between the post-training and follow-up tasks, indicating that enhancements in tracking observed post-training for the pseudo-trained group for audio-only sentences were retained even after a two-week long period, along with general training effects across both groups. For the correlational

analyses, no significant relationship between the neural and behavioural data were found. This is in line with results from our main hypotheses, providing preliminary exploratory evidence against an intrinsic link between tracking and intelligibility, although future research powered for such explorations is needed to confirm this finding.

These findings indicate that audio-tactile speech can provide similar neural benefits to speech processing as seen with audiovisual speech (Crosse *et al.*, 2015). However, these benefits were lost post-training where training with non-congruent tactile stimulation enhanced audio-only tracking and inhibited audio-tactile tracking in the pseudo-trained group. As there was no significant benefit of congruent training to the tracking of either stimuli condition in the trained group, one possibility is that this form of training is not sufficient to catalyse neural benefits. Moreover, this work further supports recent evidence that there is no intrinsic link between neural speech-envelope tracking and speech intelligibility (Kösem *et al.*, 2023). This puts into question the role of neural speech-envelope tracking in speech perception. Tracking enhancements may be a precursor to intelligibility enhancements, where intelligibility enhancements may occur with further, more long-term training (see Tawfik *et al.*, 2015). Specifically, longer-term training might be required to observe an audio-tactile benefit to speech-in-noise intelligibility. Alternatively, or possibly in addition, a different training approach may also be required. The selected training task was a bottom-up process, in which participants were asked to identify key words in speech-in-noise, directing attention to the auditory modality. This type of training did not require that participants explicitly attend to the tactile domain, only that tactile stimulation was present during training to provide further speech-relevant information to the trained group. For some participants in the trained group, the tactile stimulation could have been ignored during training, leading to ineffective learning of audio-tactile integration. Likewise, this could explain why the pseudo-trained group demonstrated greater audio-only tracking

improvements. In this case, the pseudo-trained group would be more likely to ignore the tactile stimuli due to it not being useful, thus directing more cognitive resources towards the auditory stimuli. Future research into audio-tactile speech tracking should use top-down training paradigms to ensure that participants are focusing on the tactile stimulation. Such paradigms may provide the missing link between speech tracking and speech intelligibility.

If future research can increase the behavioural benefit from audio-tactile training, this will create new opportunities for speech-relevant tactile stimulation in real-world listening environments. Devices with readily available tactile components, such as mobile phones, smart watches, or hearing aids, could be utilised to boost understanding of speech. This may be of particular benefit to current developments in neuro-steered hearing technologies, which aim to dynamically adjust hearing algorithms in real-time using neural measures such as tracking accuracy (see: Geirnaert *et al.*, 2021; Geirnaert *et al*., 2024; Straetmans, 2022). If a user experiences a decrease in tracking during listening, speech-relevant tactile stimulation could be presented to them to aid these neuro-steered aids in enhancing accurate tracking and intelligibility and providing more accurate updates to hearing aid algorithms in real-time. It also brings to light the importance of ensuring that intelligibility is hindered or enhanced in real-time based on neural tracking measurements alone. It is imperative that the role of tracking in speech perception be fully understood before it is relied upon as a measure of speech understanding in technology. Furthermore, speech tracking in the audiovisual field would also require a reassessment, as neural tracking benefits there may not be relevant to increases in speech intelligibility also seen with audiovisual speech.

3.6.1 Conclusion

In conclusion, the study replicated previous research findings on the immediate benefits of audio-tactile speech to speech-envelope tracking and its disassociation with speech intelligibility. The introduction of short-term training did not enhance tracking with

audio-tactile speech for those trained congruently but did with audio-only speech for those trained incongruently. This unexpected short-term training effect did not translate to behavioural benefit to speech intelligibility either. This work provides insight into the use of audio-tactile speech to benefit neural representations of speech prior to training. This baseline neural tracking enhancement indicates that tactile stimulation may have a benefit to real-world listening, especially in difficult listening conditions, through touch-based devices such as mobile phones, smart watches, and hearing aids. Future research should investigate other possible training paradigms to improve speech intelligibility benefits to audio-tactile speech.

Tables and Figures

Figure 1.

Flowchart showing the experimental structure, including all tasks participants completed and the timeline of completion.

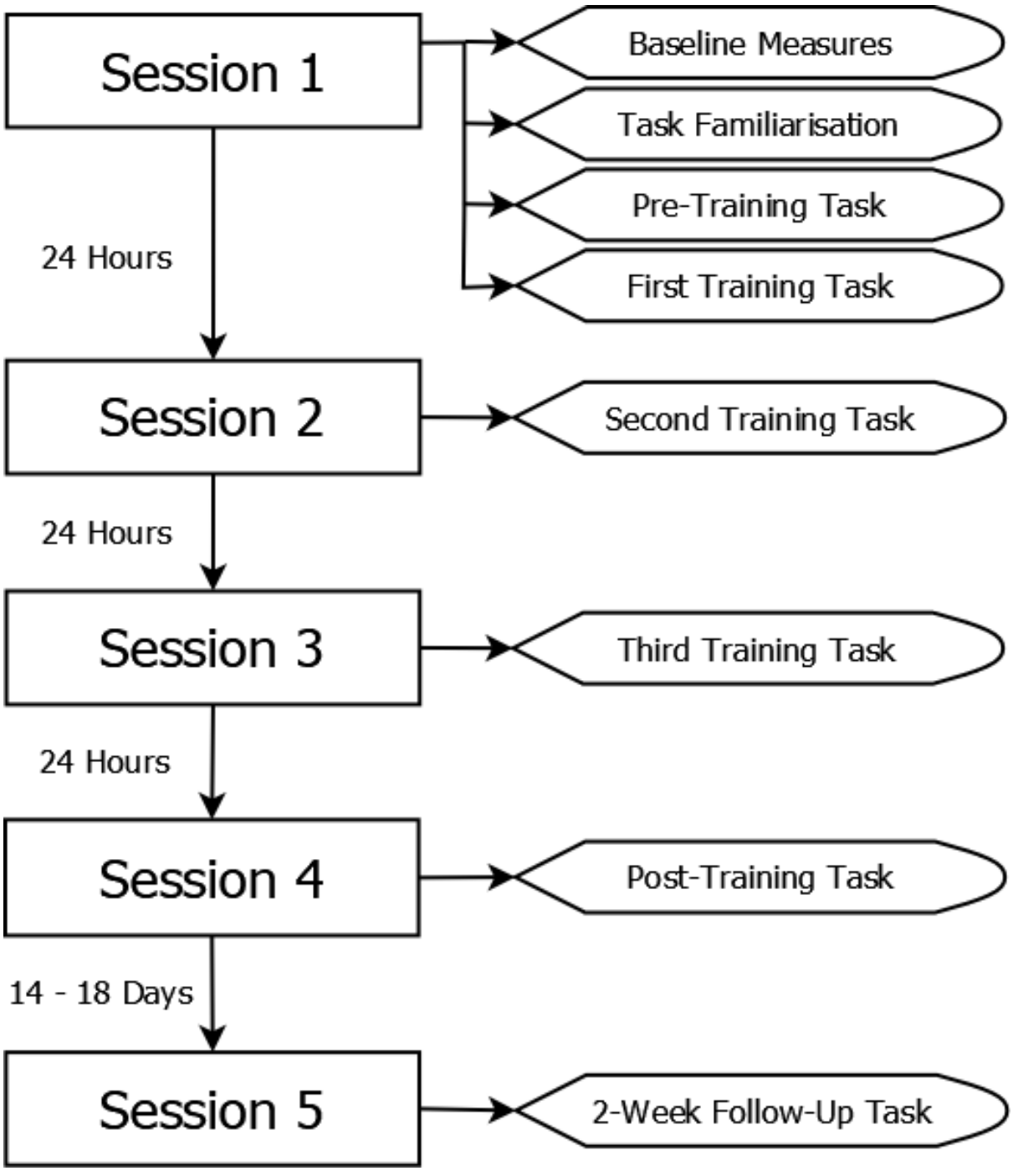


Figure 2.

Boxplots showing the median and interquartile ranges for tracking accuracy (Rz) of all participants in the pre-training task, for both audio-only and audio-tactile speech.

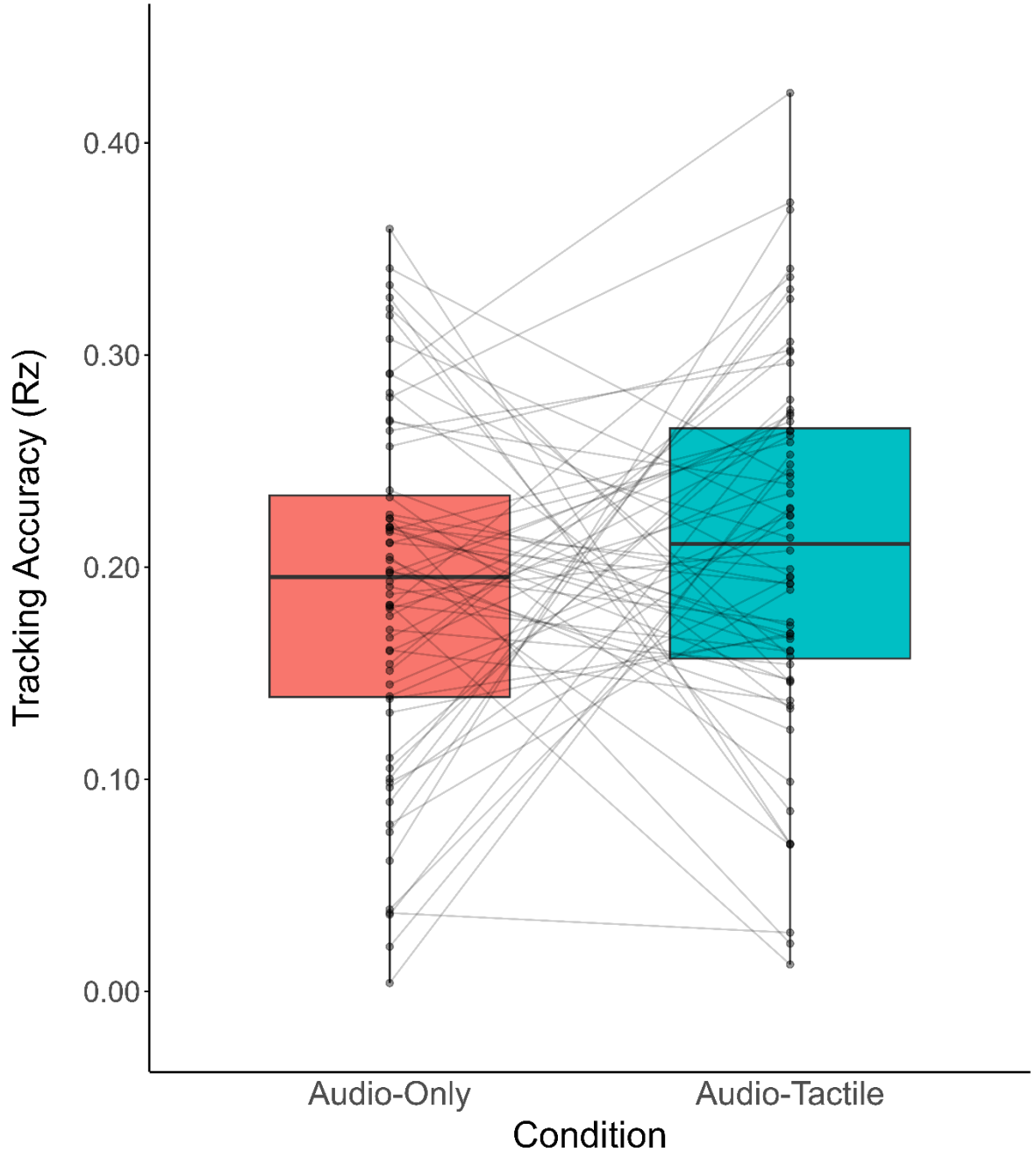


Figure 3.

Boxplots showing the median and interquartile ranges for the difference in tracking accuracy from post-training to baseline (post- minus pre-training, dRz) of participants in both the trained and the pseudo-trained groups, for both audio-only and audio-tactile speech.

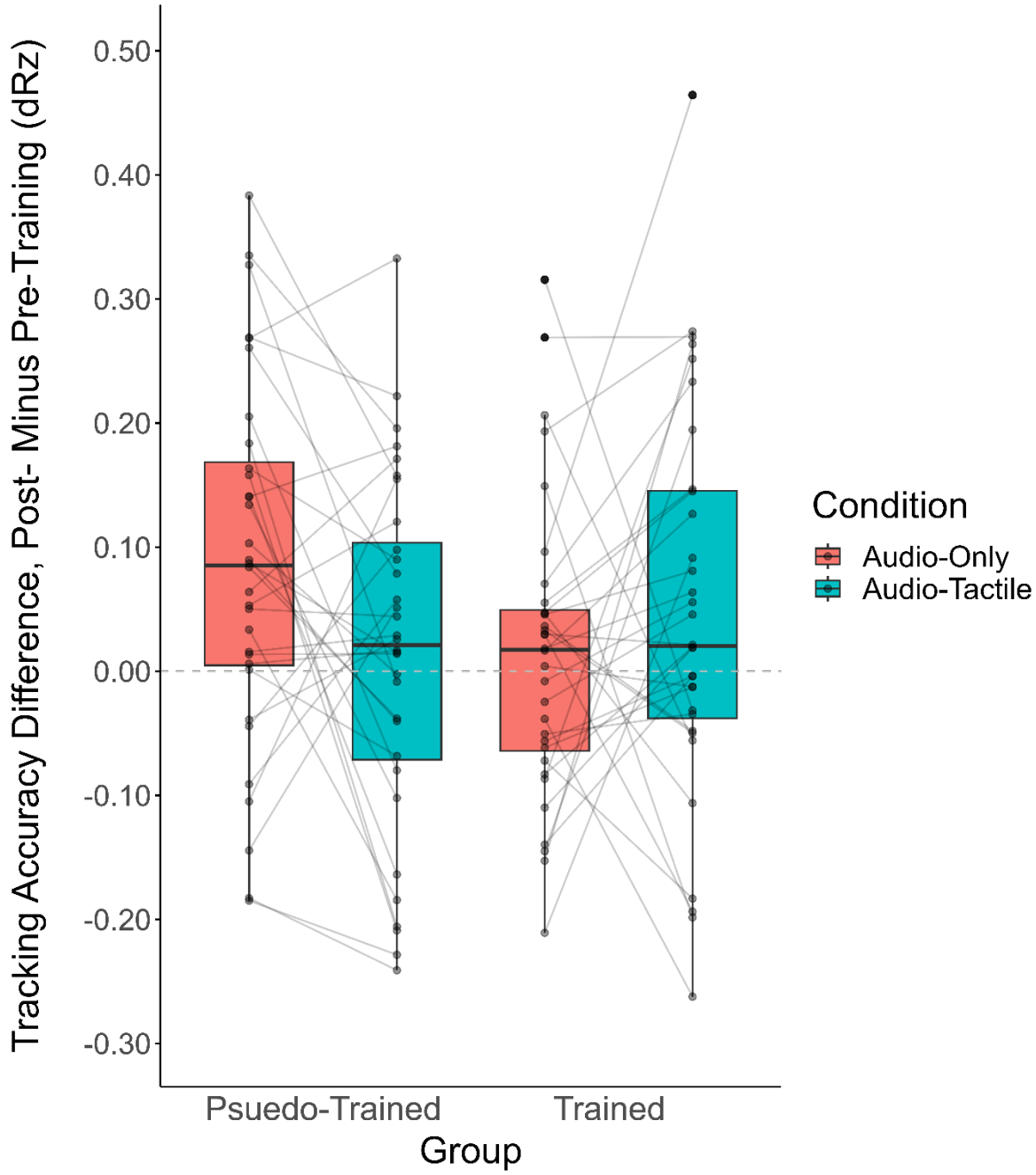


Figure 4.

Boxplots showing the median and interquartile ranges for the difference in speech intelligibility from post-training to baseline (post- minus pre-training, %) of participants in both the trained and the pseudo-trained groups, for both audio-only and audio-tactile speech.

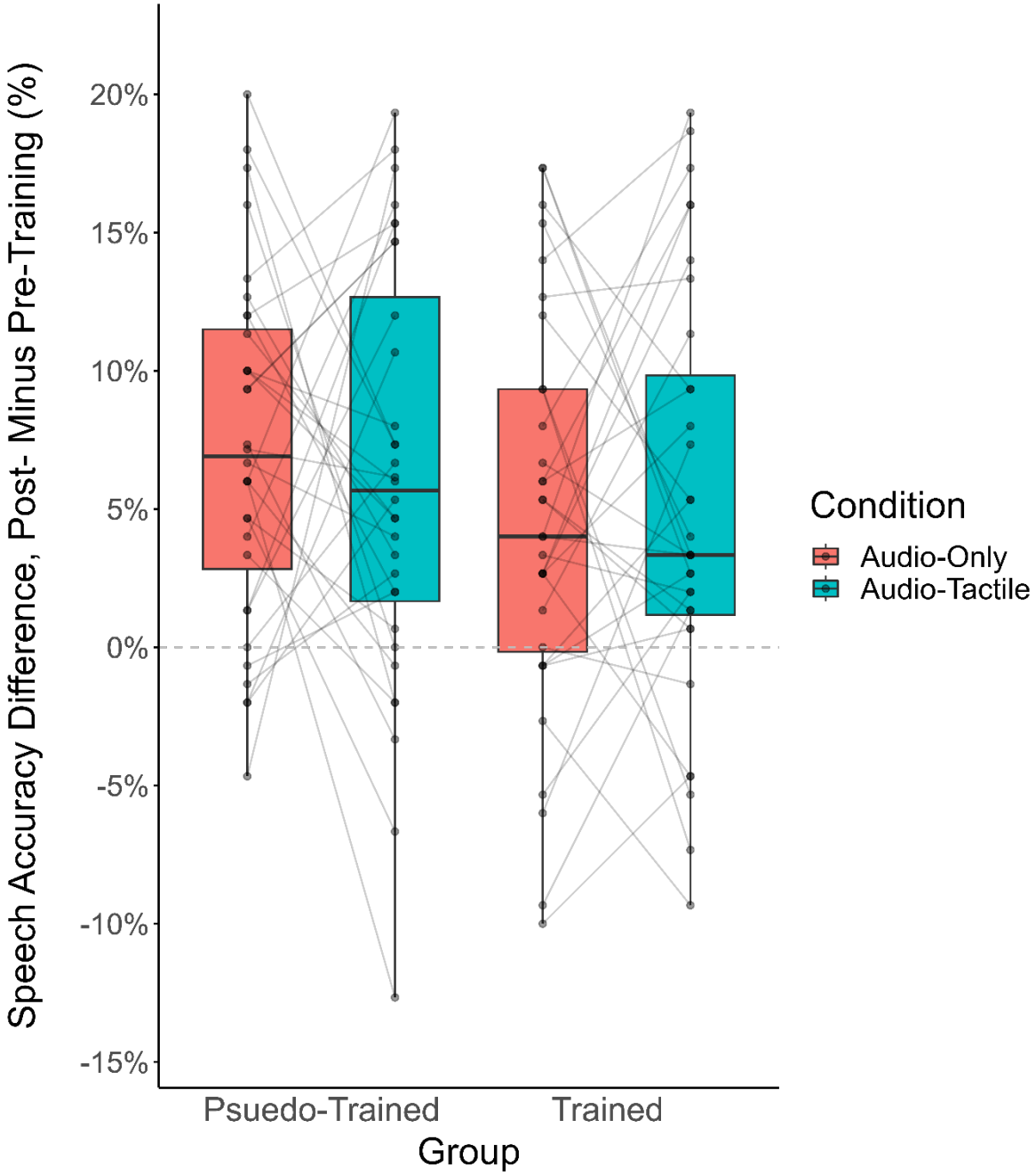


Figure 5.

Boxplots showing the median and interquartile ranges for the difference in tracking accuracy from the two-week follow-up task to post-training (follow-up minus post-training, dRz) of participants in both the trained and the pseudo-trained groups, for both audio-only and audio-tactile speech.

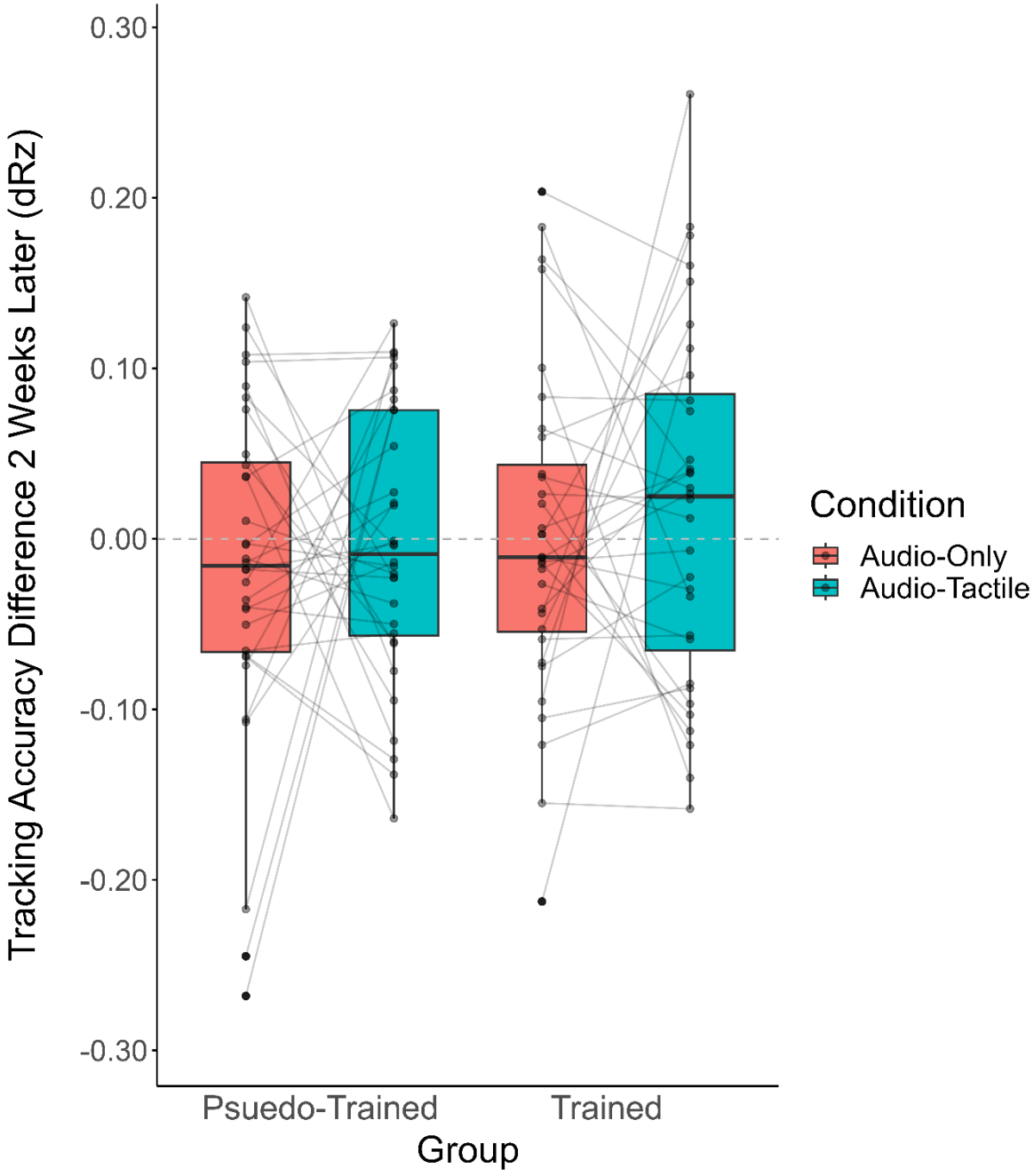
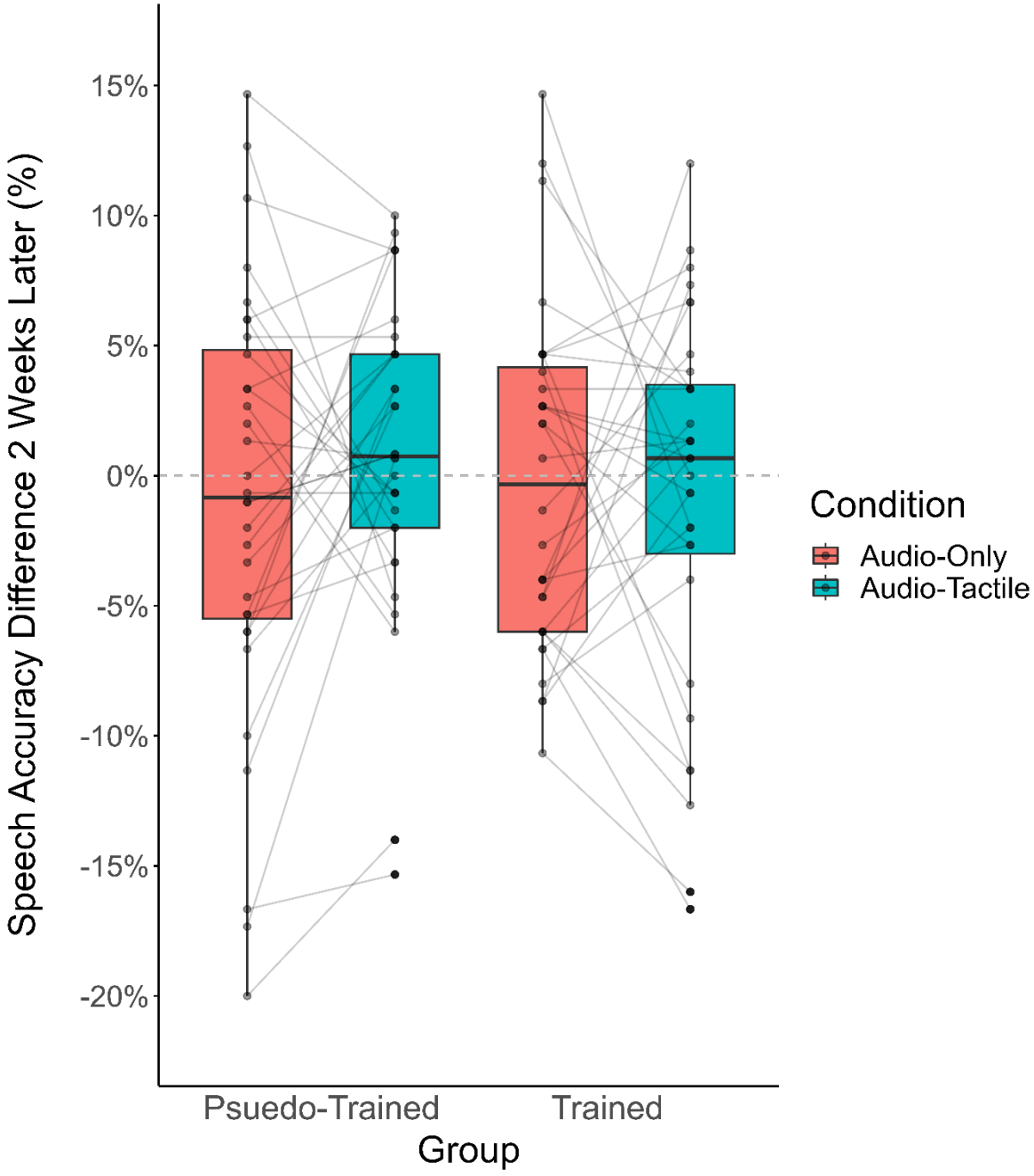


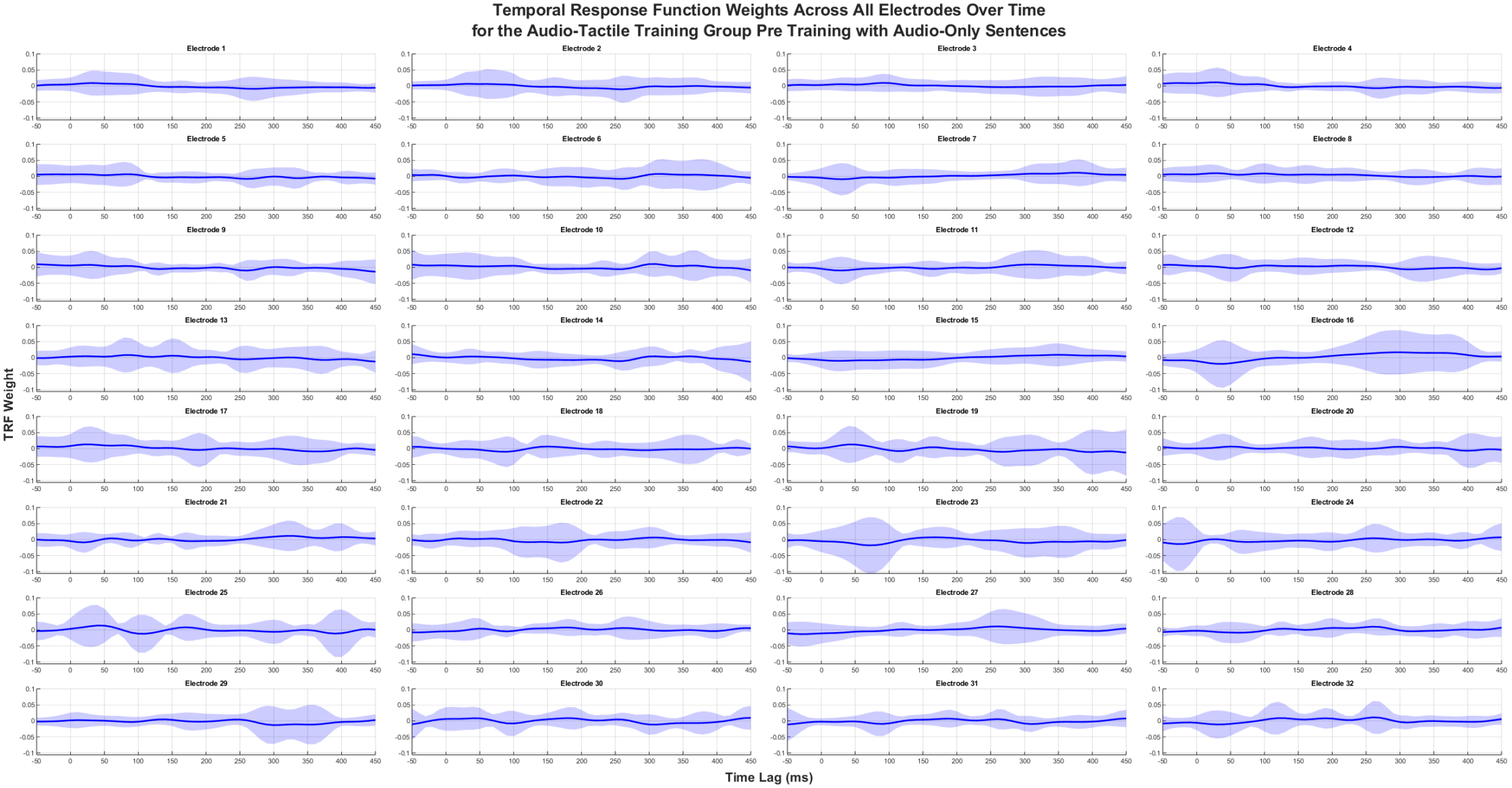
Figure 6.

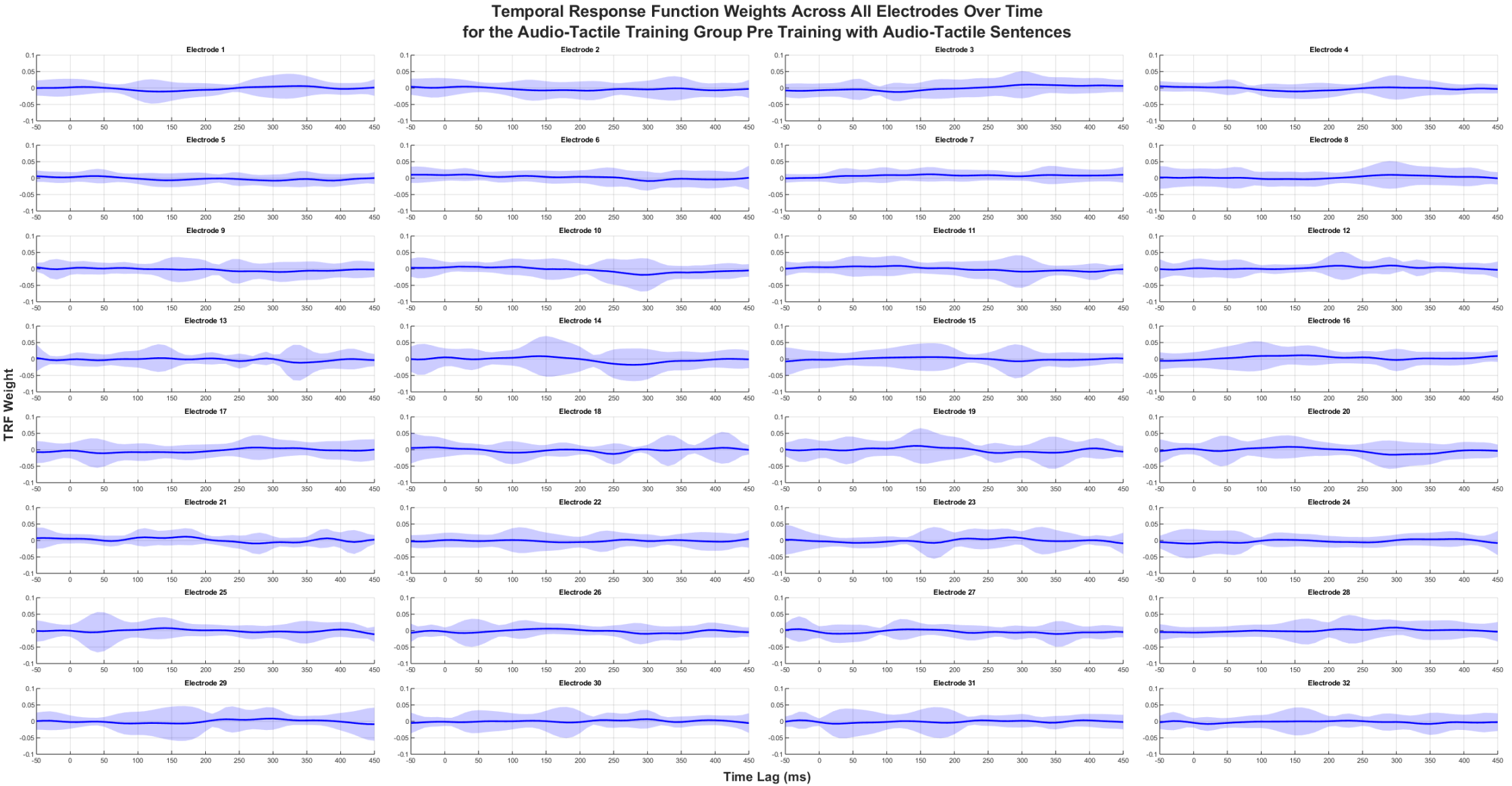
Boxplots showing the median and interquartile ranges for the difference in speech intelligibility from the two-week follow-up task to post-training (follow-up minus post-training, %) of participants in both the trained and the pseudo-trained groups, for both audio-only and audio-tactile speech.

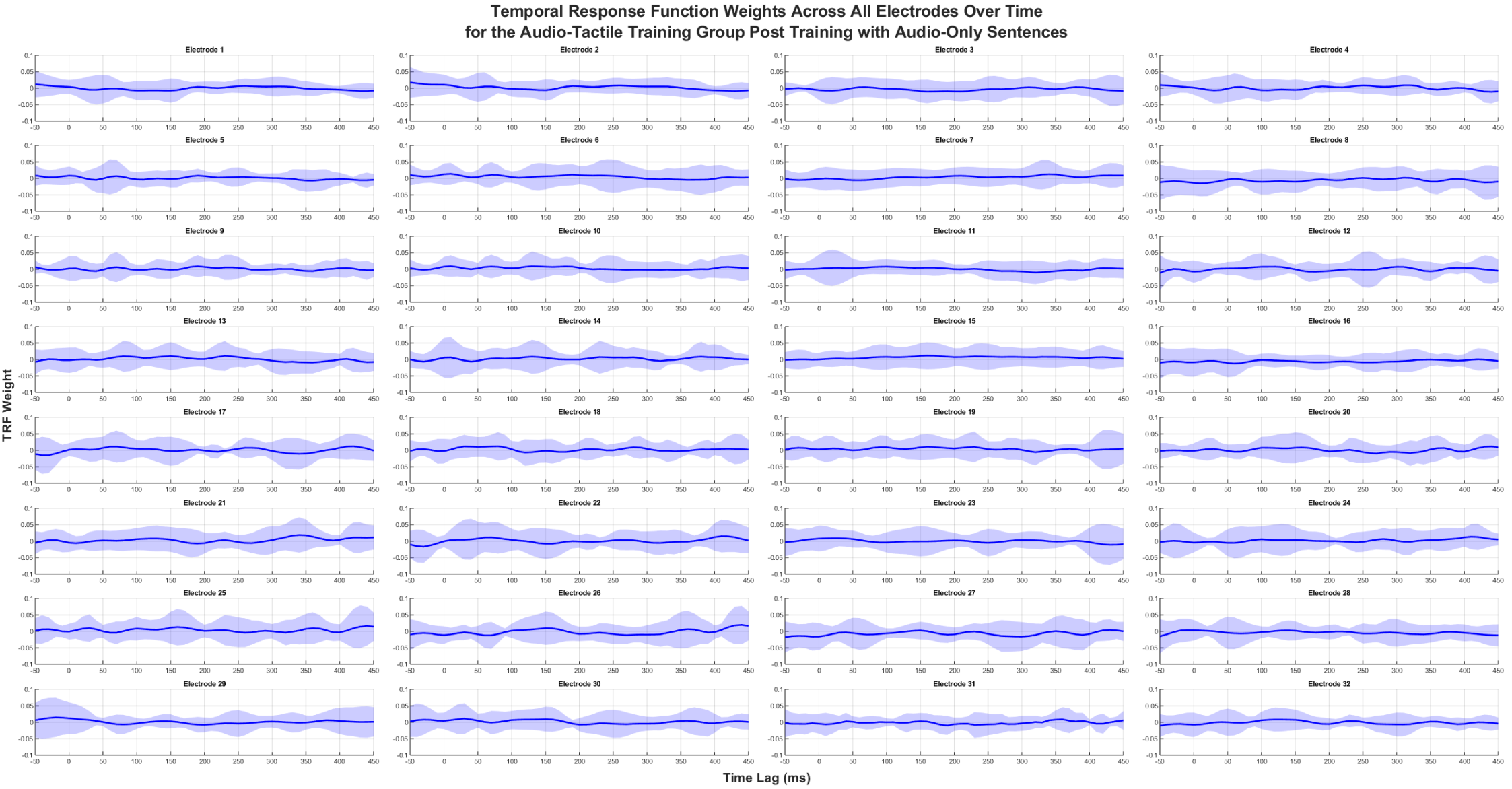


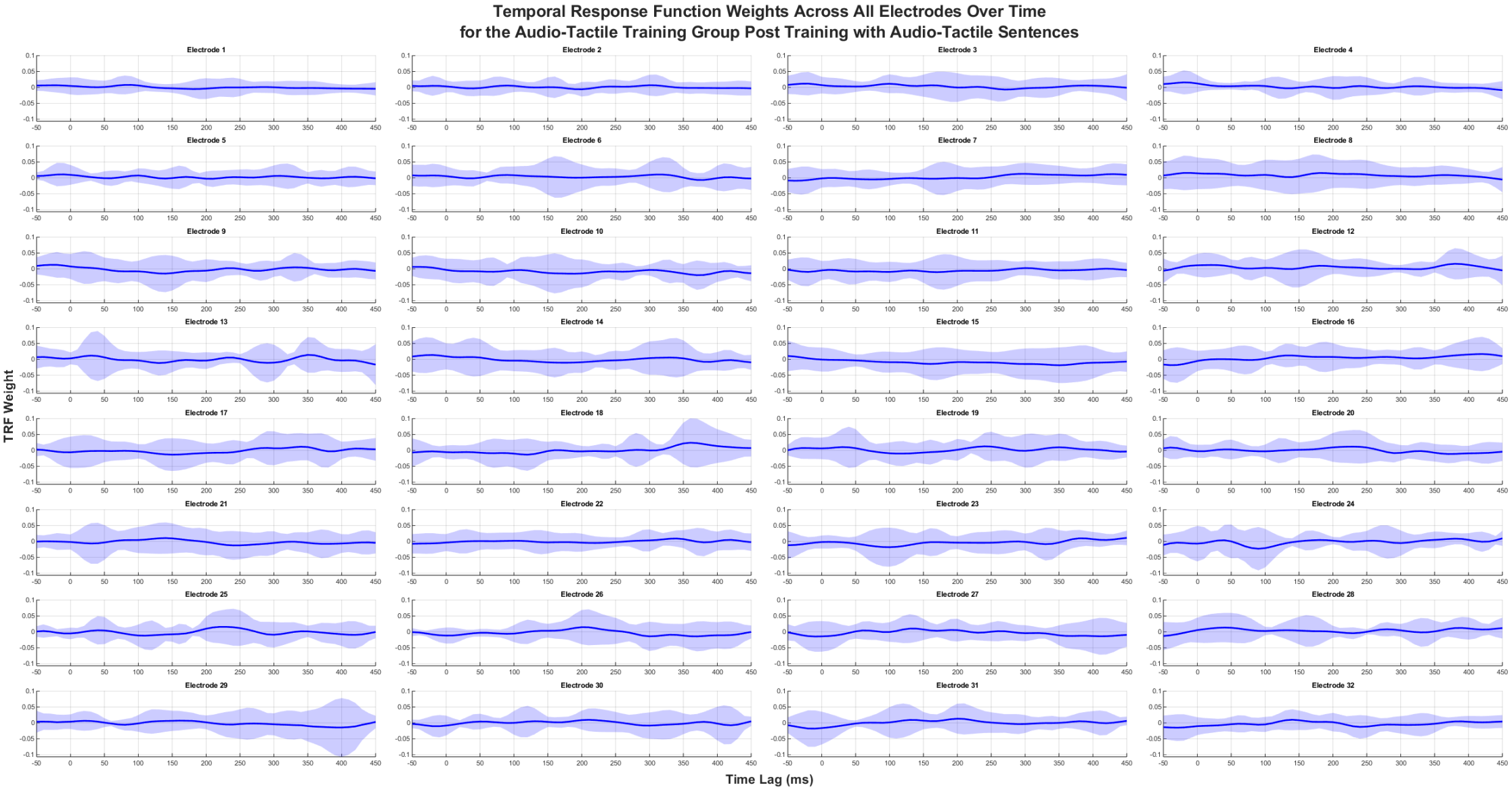
Supplementary Materials A

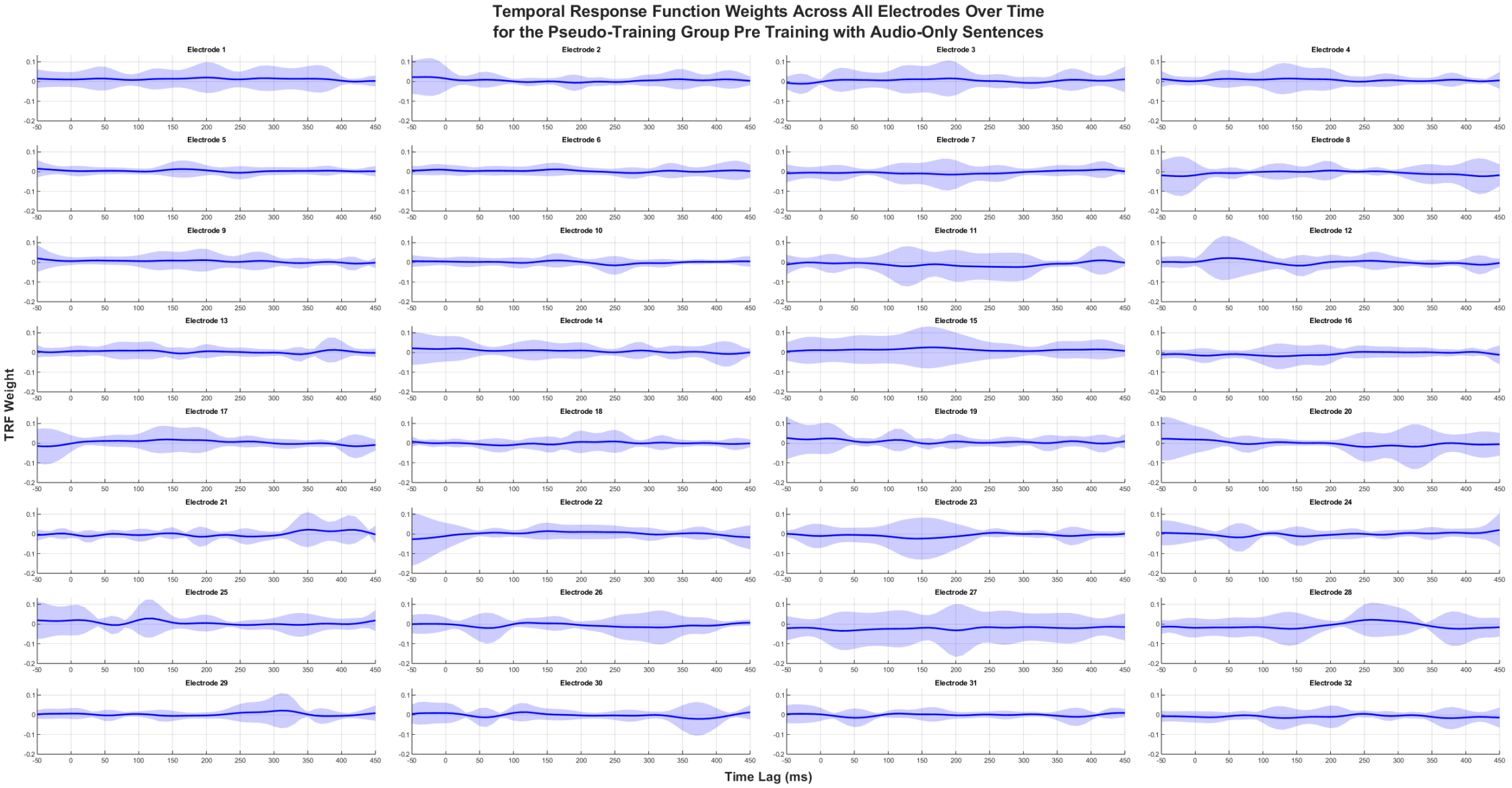
The following figures show the temporal response function (TRF) weights across all 32 electrodes for each training group (trained, and pseudo-trained), session number (pre-training, and post-training), and stimulation type (audio-only, and audio-tactile). In each figure, the solid blue line represents the grand average mean of TRF weightings across all participants, with variance displayed as one standard deviation away from the mean.

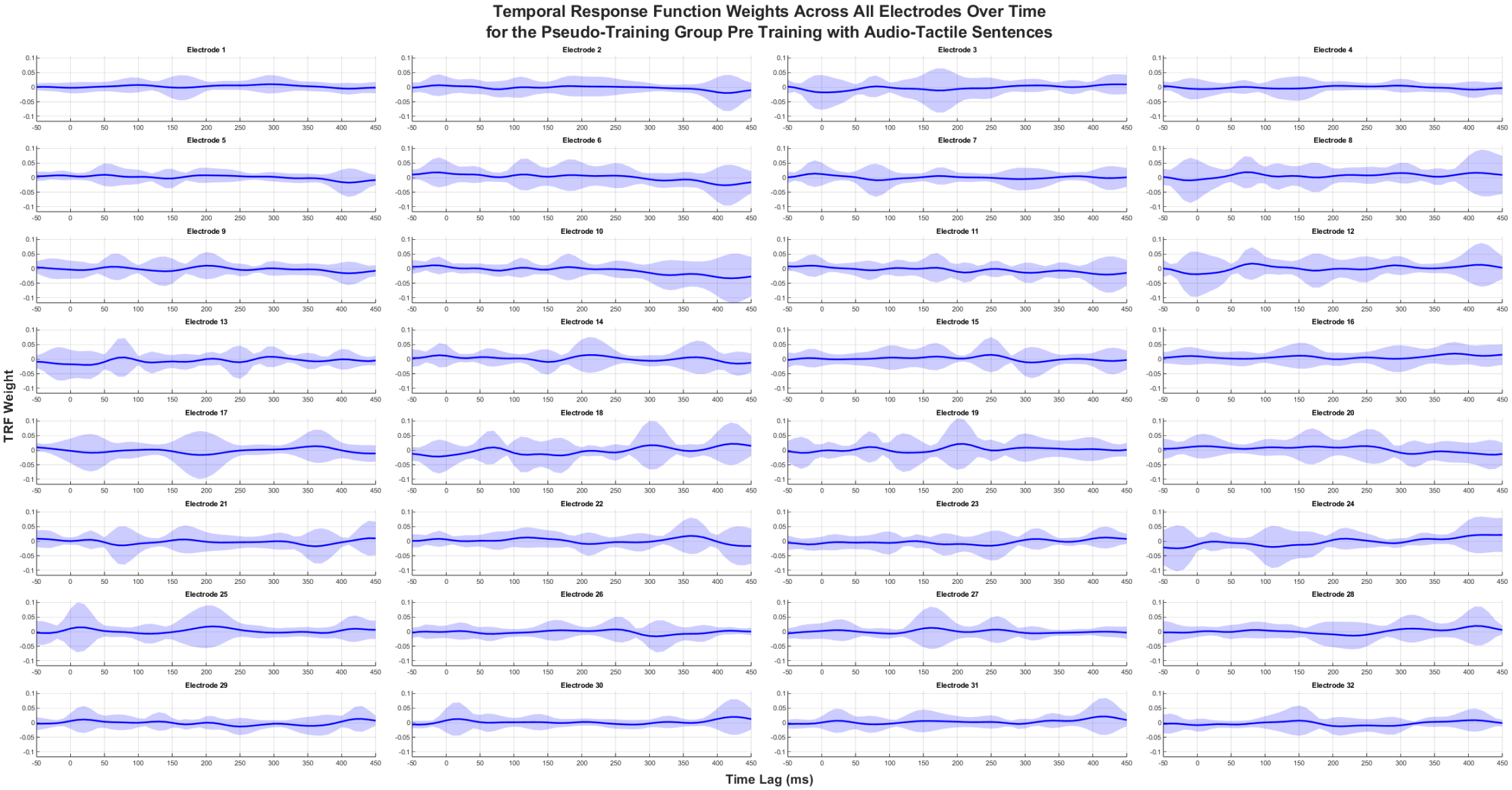

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Pre Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

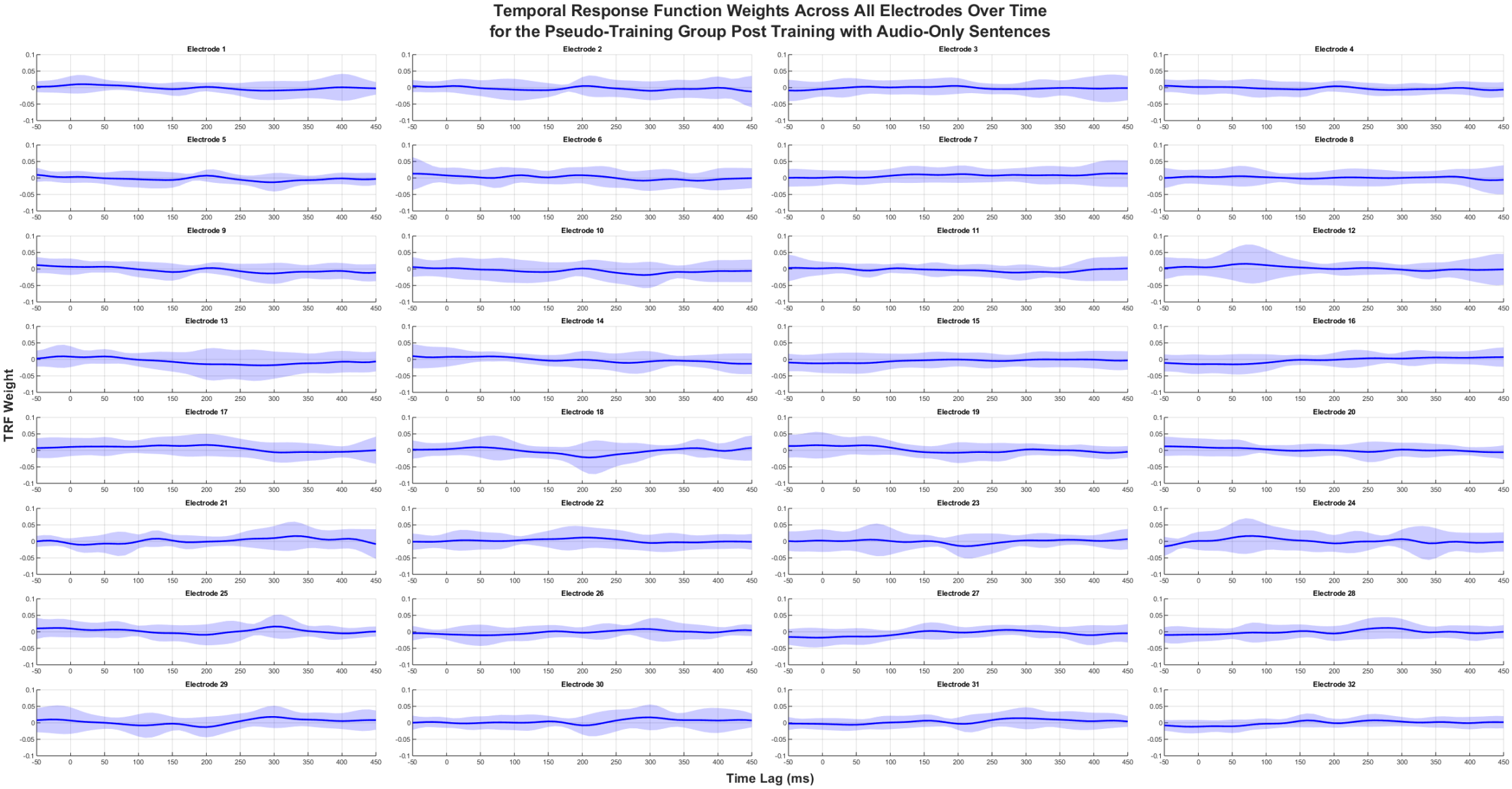

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Pre Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

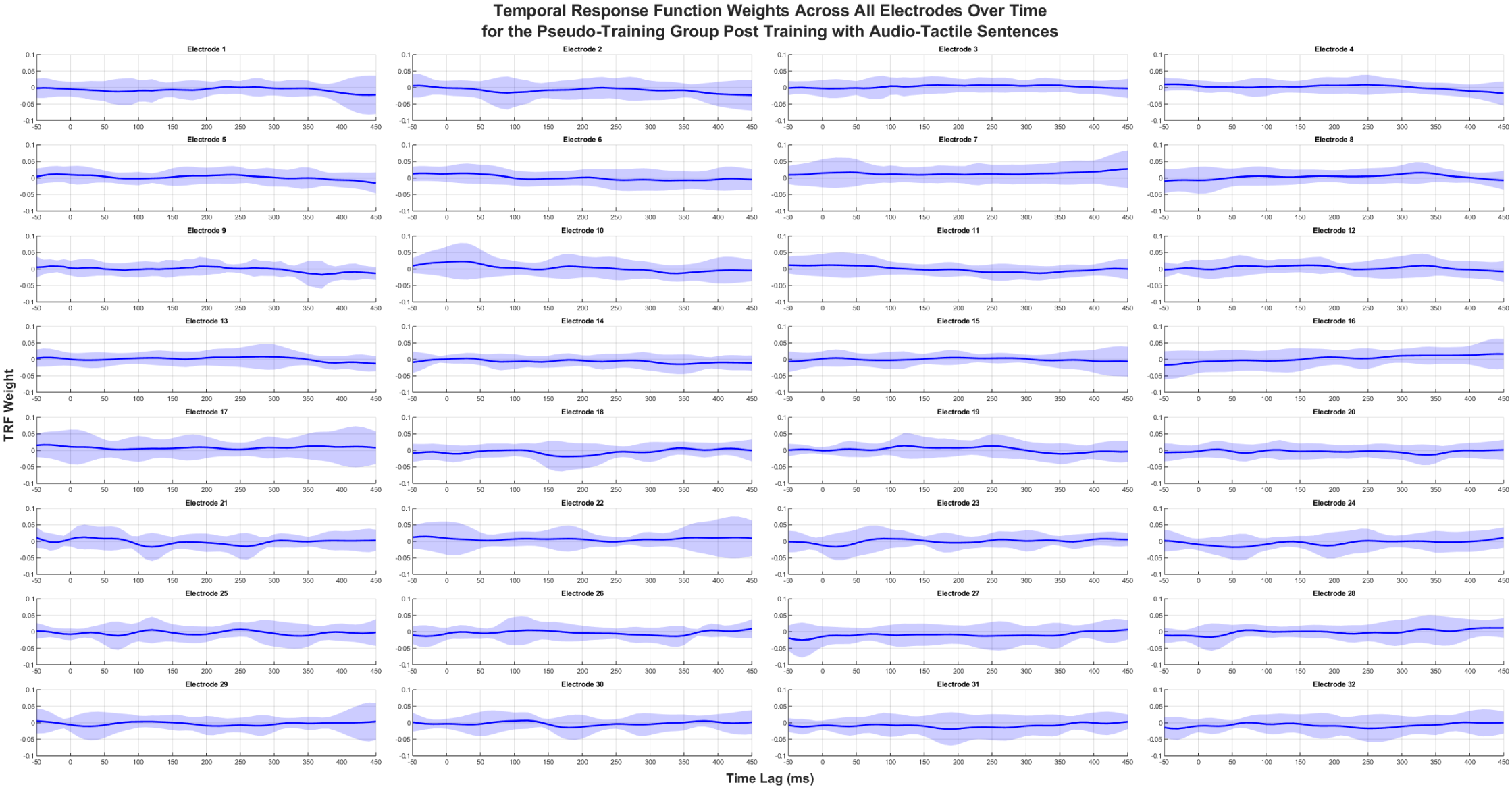

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Post Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Post Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Pre Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

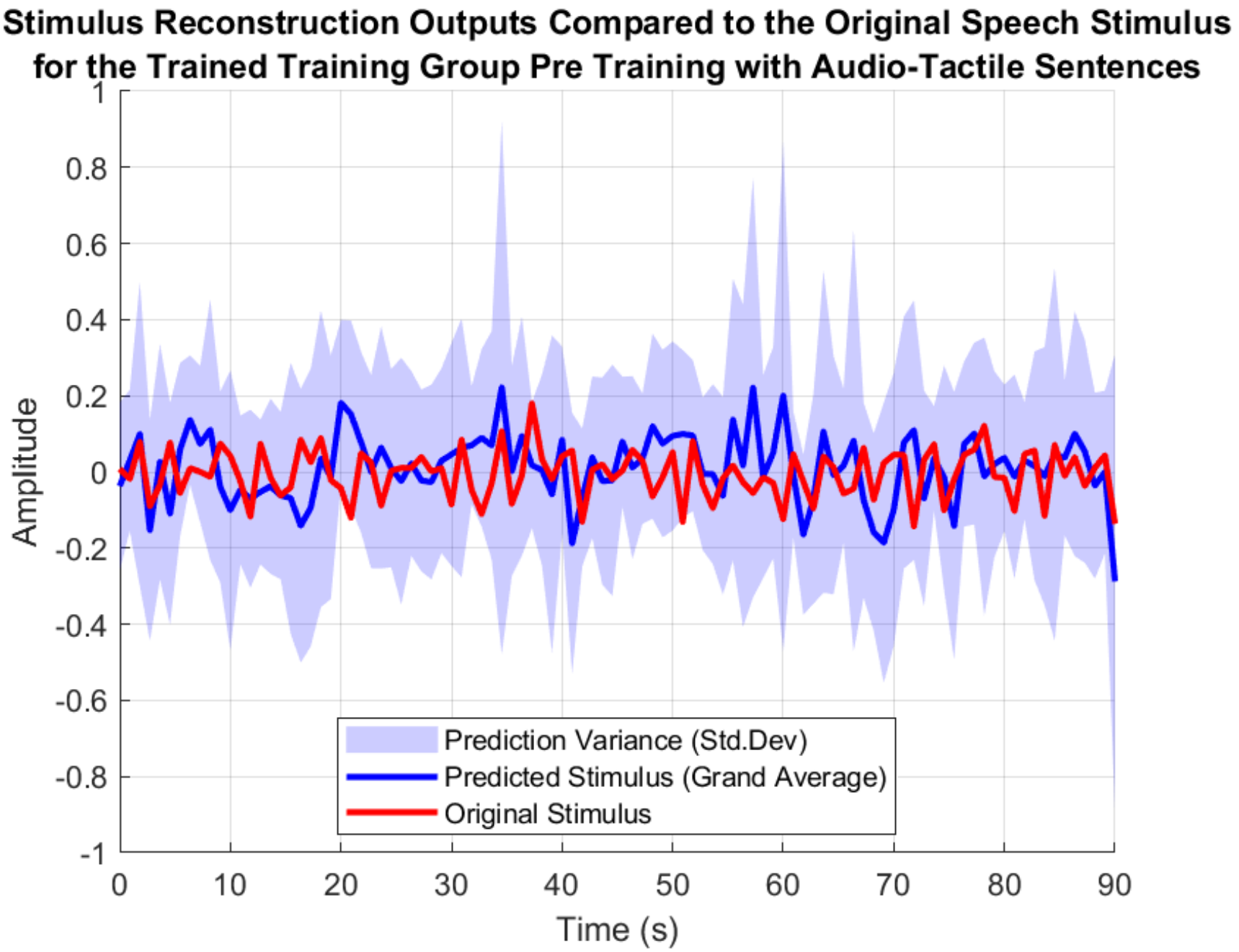

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Pre Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

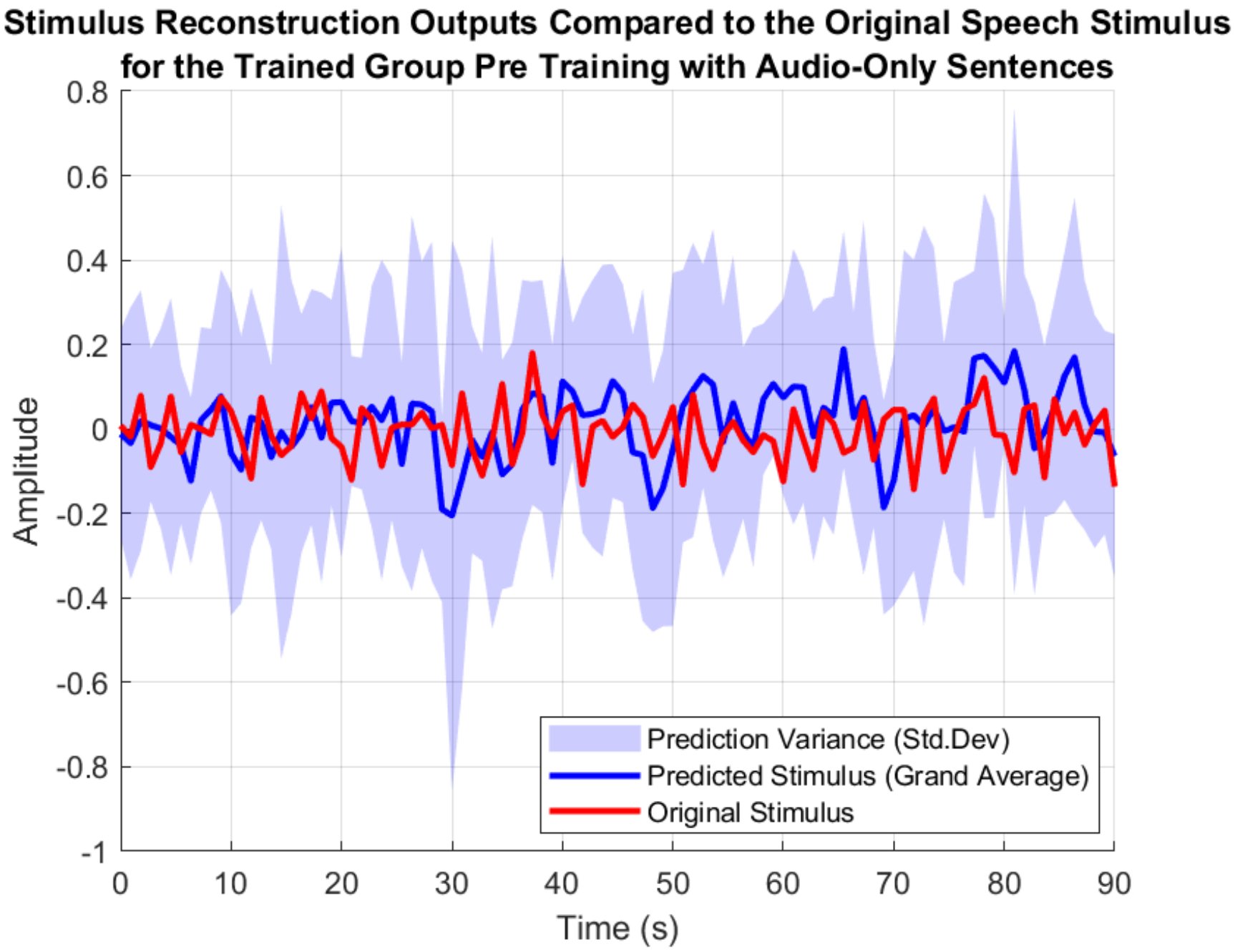

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Post Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

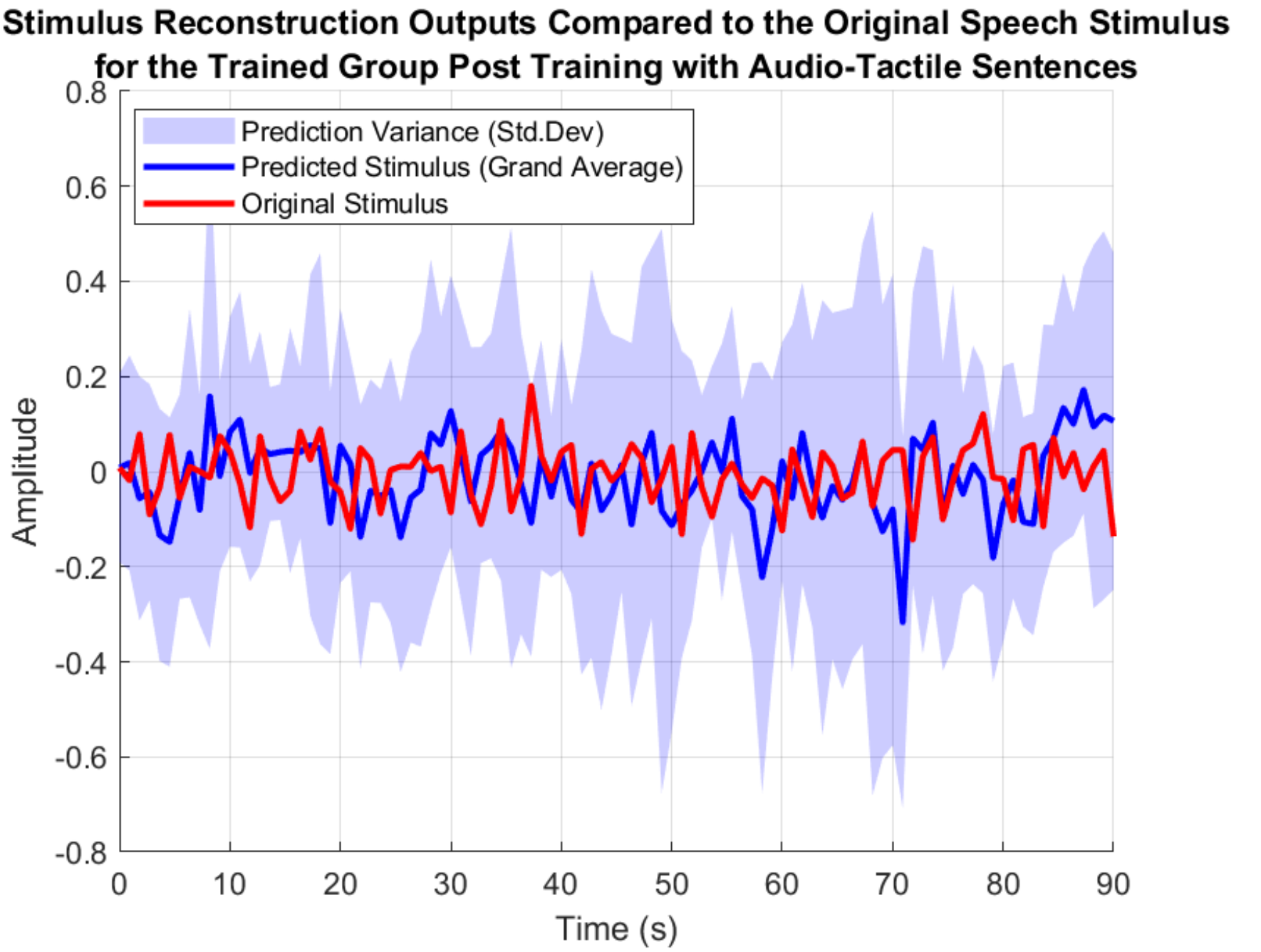

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Post Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

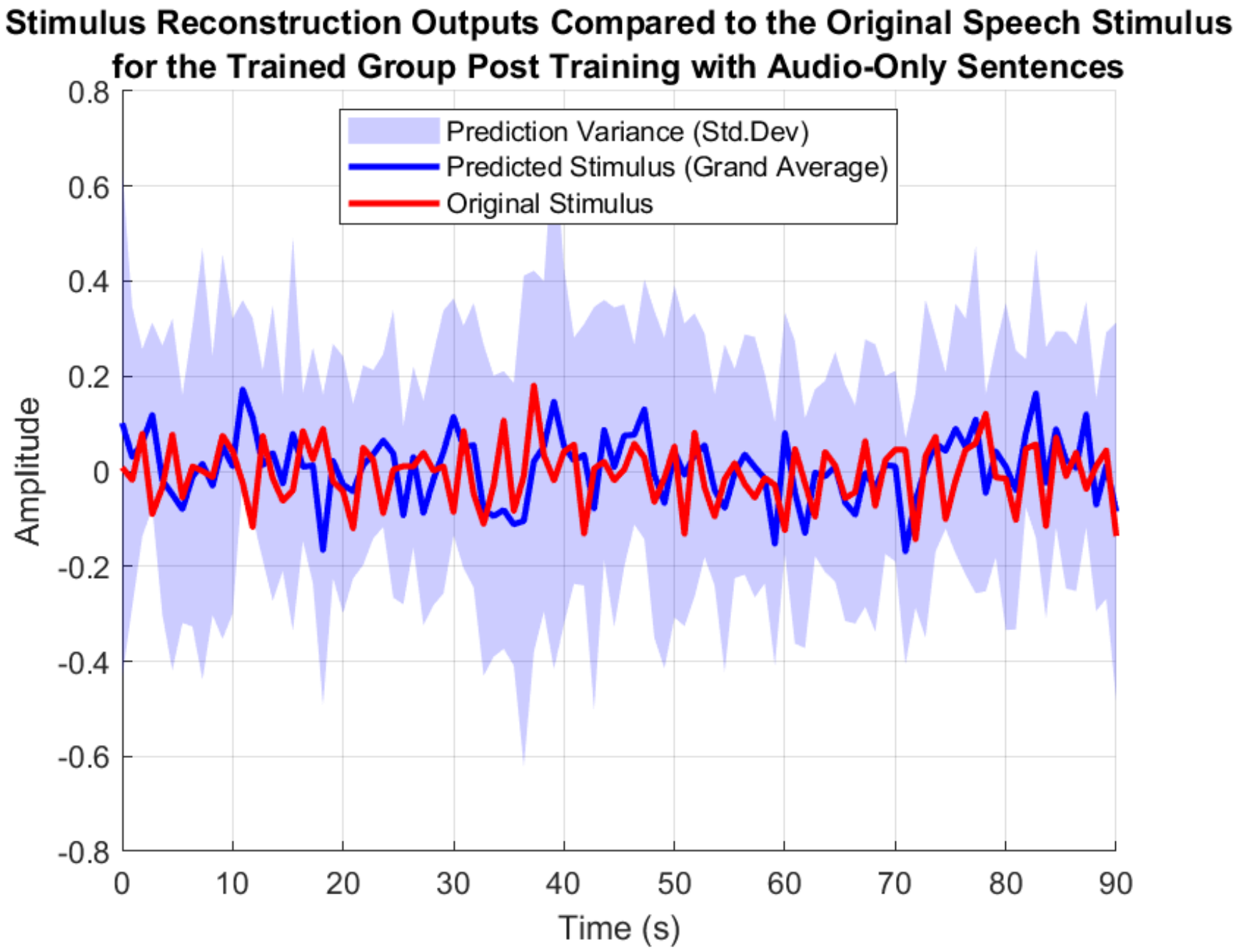
Supplementary Materials B

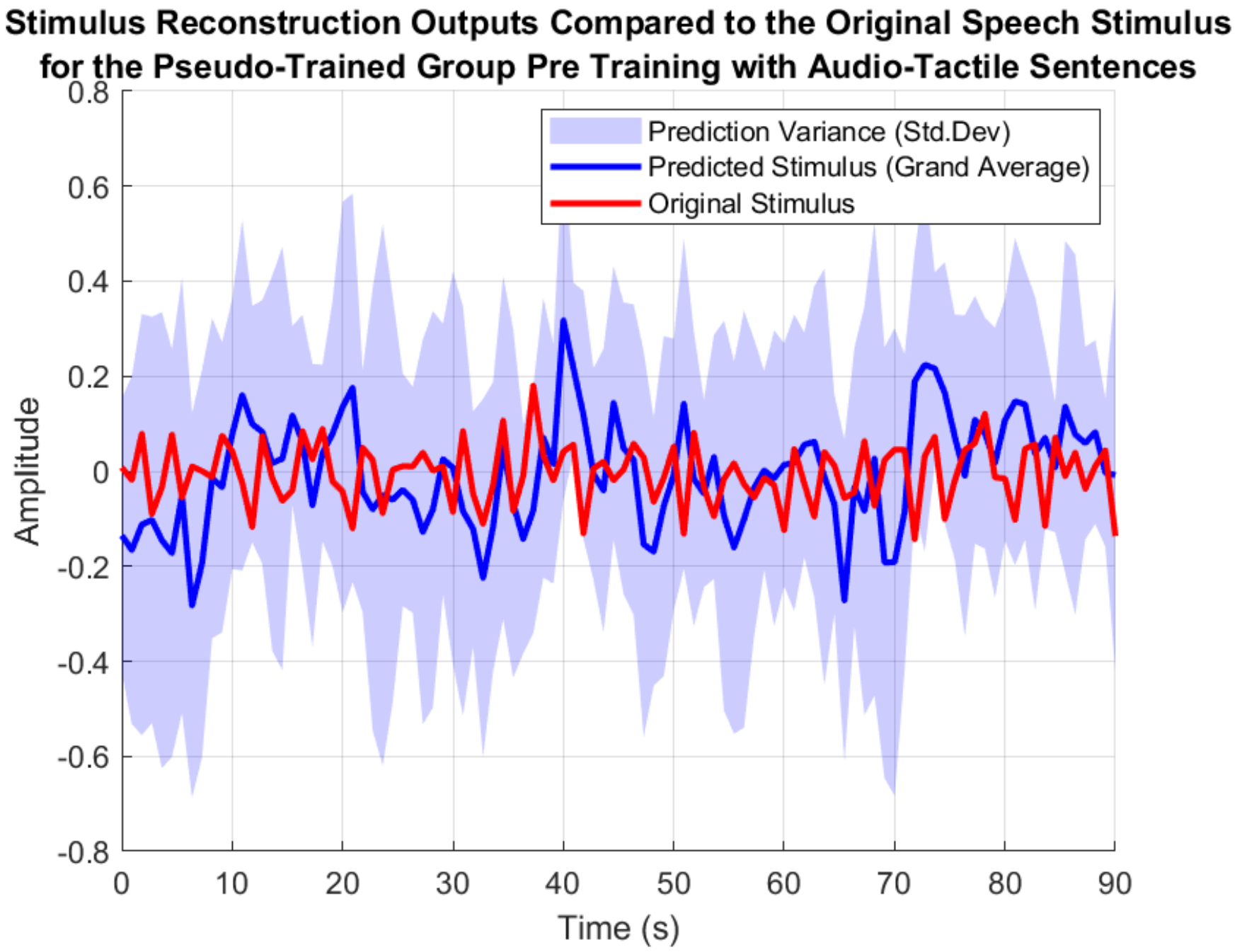
The following figures show the comparisons between the predicted stimuli from stimulus reconstructions and the original speech stimulus, for each training group (trained, and pseudo-trained), session number (pre-training, and post-training), and stimulation type (audio-only, and audio-tactile). In each figure, the solid blue line represents the grand average (mean) predicted stimulus across all participants, with variance displayed as one standard deviation away from the mean, whilst the solid red line represents the original stimulus (speech envelope).

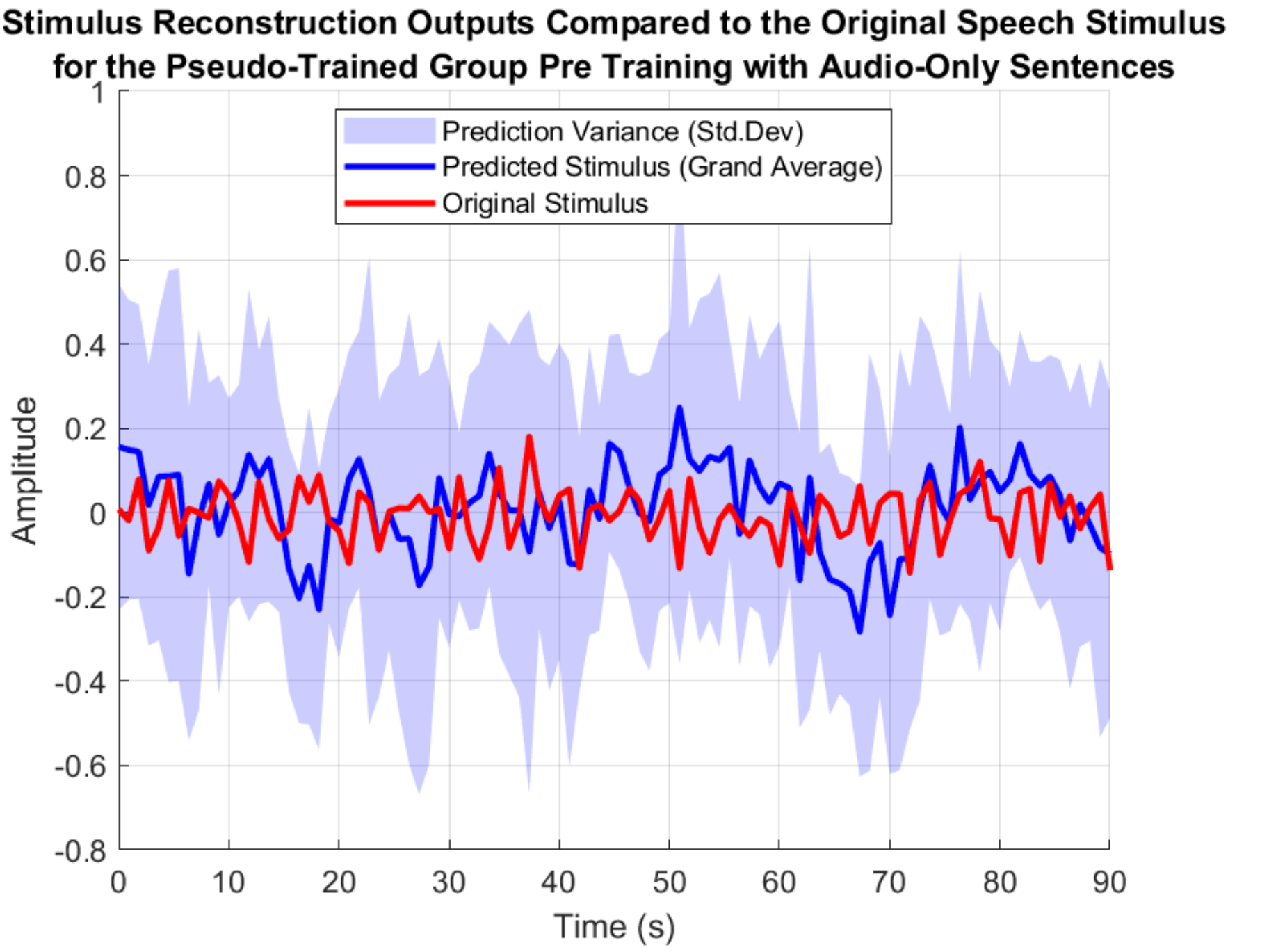

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Trained Training Group Pre Training with Audio-Tactile Sentences
Amplitude
Time (s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus

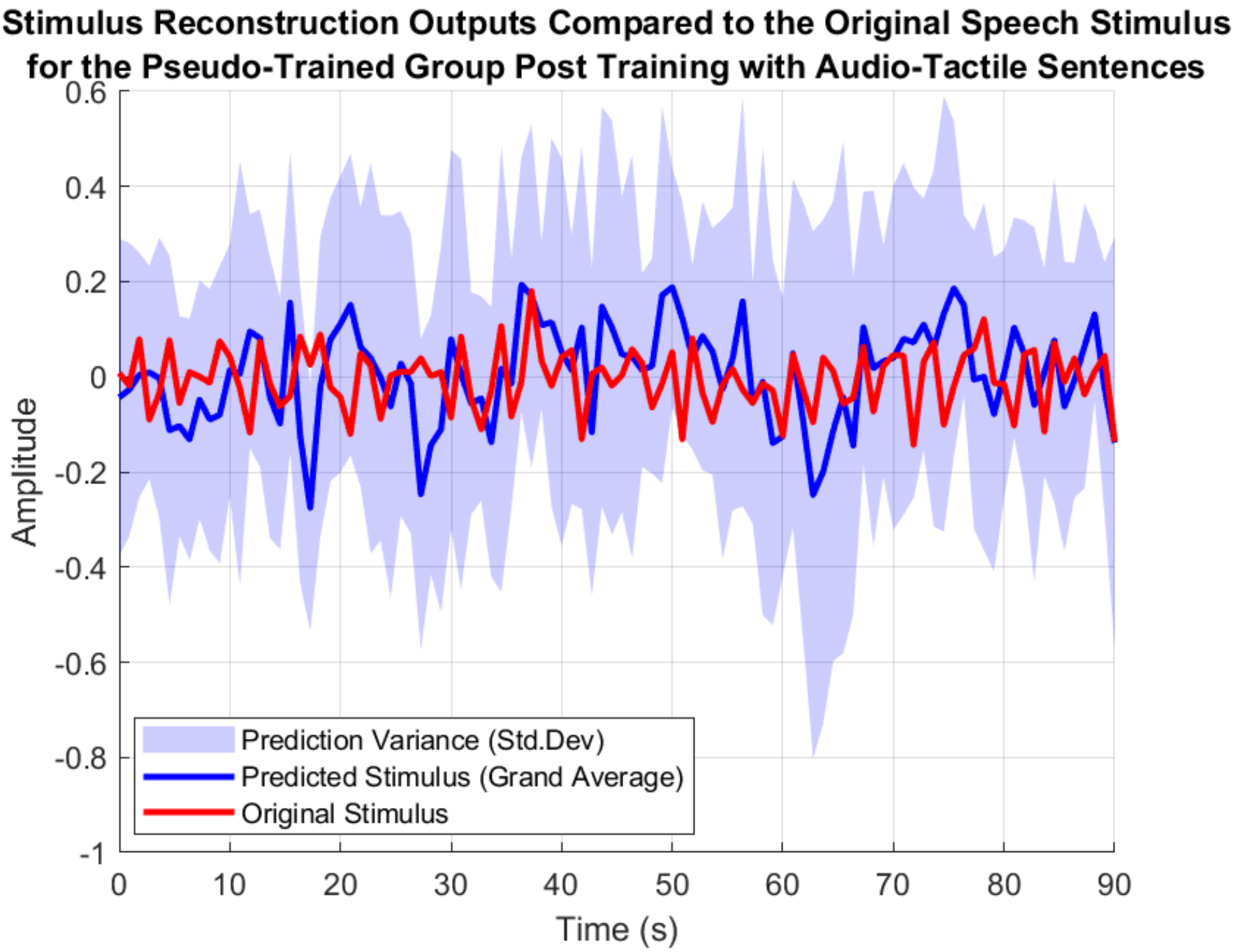

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Trained Group Pre Training with Audio-Only Sentences
Amplitude
Time (s)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus

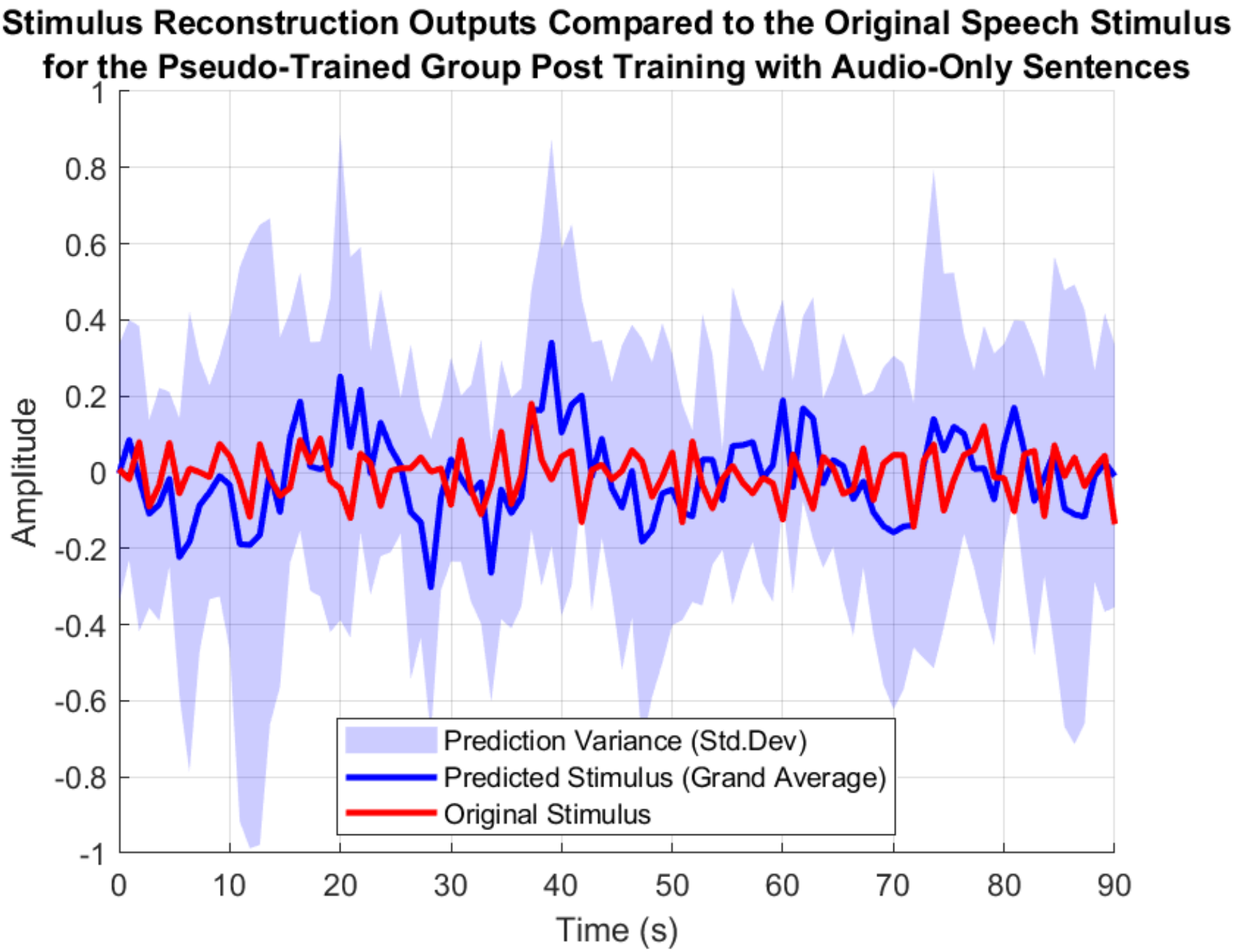

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Trained Group Post Training with Audio-Tactile Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
Time (s)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Trained Group Post Training with Audio-Only Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90
Time (s)


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Pseudo-Trained Group Pre Training with Audio-Tactile Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
Time (s)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Pseudo-Trained Group Pre Training with Audio-Only Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
Time (s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Pseudo-Trained Group Post Training with Audio-Tactile Sentences
Amplitude
Time (s)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Pseudo-Trained Group Post Training with Audio-Only Sentences
Amplitude
Time (s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus

References

An, H., Lee, J., Suh, M. W., & Lim, Y. (2023). Neural correlation of speech envelope tracking for background noise in normal hearing. *Frontiers in Neuroscience, 17*, 1268591.

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory, 9*(3), 138-150.

Bates, D., Maechler, M., Bolker, B., & Walker., S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw, 67*(1), 1-48.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25*(5), 402-412.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.53, retrieved August 2021 from http://www.praat.org/

BRITISH SOCIETY OF AUDIOLOGY (2018). Pure-tone air-conduction and bone conduction threshold audiometry with and without masking [Online]. Available at: https://www.thebsa.org.uk/resources/

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *NeuroImage, 233*, 117958.

Cieśla, K., Wolak, T., Lorens, A., Mentzel, M., Skarżyński, H., & Amedi, A. (2022). Effects of training and using an audio-tactile sensory substitution device on speech-in-noise understanding. *Scientific Reports, 12*(1), 3206.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Vincent, W., De Tiège, X., & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage, 184*, 201-213.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience, 39*(29), 5750-5759.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B. T., ... & Bertrand, A. (2021). EEG-based auditory attention decoding: Towards neuro-steered hearing devices. *arXiv preprint*. Accessed at: https://arxiv.org/abs/2008.04569

Geirnaert, S., Zink, R., Francart, T., & Bertrand, A. (2024). Fast, accurate, unsupervised, and time-adaptive EEG-based auditory attention decoding for neuro-steered hearing devices. *In Brain-Computer Interface Research: A State-of-the-Art Summary 11* (pp. 29-40). Cham: Springer Nature Switzerland.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. *Journal of Neuroscience*, *33*(4), 1417-1426.

Graetzer, S., Akeroyd, M. A., Barker, J., Cox, T. J., Culling, J. F., Naylor, G., ... & Viveros-Muñoz, R. (2022). Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus. *Data in Brief, 41*, 107951.

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of *Cognitive Neuroscience, 36*(1), 128-142.

Heil, P., & Peterson, A. J. (2015). Basic response properties of auditory nerve fibers: a review. *Cell and Tissue Research*, *361*(1), 129-158.

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., & Lizarazu, M. (2024). On the Speech Envelope in the Cortical Tracking of Speech. *NeuroImage*, 120675.

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *Elife, 8*, e48116.

Kong, Y. Y., Somarowthu, A., & Ding, N. (2015). Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. *Journal of the Association for Research in Otolaryngology, 16*, 783-796.

Kösem, A., Dai, B., McQueen, J. M., & Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage, 272*, 120040.

Larsby, B., Hällgren, M., Nilsson, L., & McAllister, A. (2015). The influence of female versus male speakers' voice on speech recognition thresholds in noise: Effects of low- and high-frequency hearing impairment. *Speech, Language and Hearing, 18*(2), 83-90.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 245.

Montoya-Martínez, J., Vanthornhout, J., Bertrand, A., & Francart, T. (2021). Effect of number and placement of EEG electrodes on measurement of neural tracking of speech. *PLoS One, 16*(2), e0246769. https://doi.org/10.1371/journal.pone.0246769

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Peter, G., Catriona, M., & Phillip, A. (2019). *Power analysis for generalised linear mixed models by simulation* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/simr/index.html

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage, 198*, 181-197.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences, 106*(26), 10598-10602.

Pontifex, M. B., Gwizdala, K. L., Parks, A. C., Billinger, M., & Brunner, C. (2017). Variability of ICA decomposition may impact EEG signals when used to remove eyeblink artifacts. *Psychophysiology, 54*(3), 386-398.

Qualtrics. (2005). *Qualtrics software*, Provo, Utah, USA. Copyright@2021, Current version: 09-21. Retrieved from: https://www.qualtrics.com

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*, 116134.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2024). *Analysis of factorial experiments* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/afex/index.html

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2022). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering, 18*(6), 066054.

Tawfik, S., Hassan, D. M., & Mesallamy, R. (2015). Evaluation of long term outcome of auditory training programs in children with auditory processing disorders. *International Journal of Pediatric Otorhinolaryngology, 79*(12), 2404-2410.

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in school-age children: the roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research, 57*(2), 556-565.

Vanthornhout, J., Decruy, L., & Francart, T. (2019). Effect of task and attention on neural tracking of speech. *Frontiers in Neuroscience, 13*, 977.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, *19*(2), 181-191.

Chapter 4

4 Does top-down audio-tactile speech-in-noise training affect speech-envelope tracking accuracy and intelligibility?

Linking Statement:

In the previous chapter, we had established further supporting evidence for audio-tactile stimulation providing clear initial enhancements to neural speech-envelope tracking. Unexpectedly, this enhancement was not apparent for the audio-tactile training group post-training, nor was a benefit to speech intelligibility seen. To further understand the potential contributions of audio-tactile integration to speech perception, this chapter will discuss a top-down variant on the training paradigm from Chapter 3. The implications of this top-down implementation on neural speech tracking accuracy and intelligibility will be discussed.

Author Note: *This work was produced in collaboration Dr. Helen Nuttall, Prof. Christopher Plack, Prof. Lars Hausfeld, Prof. Lars Riecke, and technician Barrie Usherwood as co-authors. This paper highlights preliminary analyses of thirty young adults as part of an extended pilot to investigate top-down audio-tactile training benefits to speech perception. Data collection of the remaining thirty participants calculated to achieve a powered sample will continue beyond submission of this thesis to support the submission of a publication in collaboration with the co-authors.*

Statement of Authorship

Chapter 4 (Paper Three):	Does top-down audio-tactile speech-in-noise training affect speech-envelope tracking accuracy and intelligibility?
Authors:	Brandon O'Hanlon, Lars Hausfeld, Lars Riecke, Barrie Usherwood, Christopher J Plack, and Helen E Nuttall
Publication status:	Published Publication has been accepted Publication has been submitted **Unpublished/Unsubmitted but in manuscript style**
Reference:	Not published.
Student/Principle Author:	Brandon Lee O'Hanlon
Contribution:	Theoretical conceptualization; study design; data collection; statistical data analysis; manuscript development; manuscript revisions based on supervisor feedback.

Principle Author Signature: Date: 14.04.2025

Through signing this statement, the Co-authors agree that:

(a) The student's contribution to the above papers is correct

(b) The student can incorporate this paper within the thesis

(c) The contribution of all co-authors for each paper equals 100% minus the contribution of the student.

Co-author Name:	Lars Hausfeld
Co-author Contributions:	Collaborator. Contributed theoretical knowledge and advice on cortical speech-envelope tracking methodology and audio-tactile training design.

Co-author Signature: Date: 16/04/2025

Co-author Name:	Lars Riecke
Co-author Contributions:	Collaborator. Contributed theoretical knowledge and advice on cortical speech-envelope tracking methodology and audio-tactile training design.

Co-author Signature: Date: 17/04/2025

Co-author Name:	Barrie Usherwood
Co-author Contributions:	Collaborator. Development and creation of audio-tactile stimulation device.

Co-author Signature: Date: 16/04/2025

Co-author Name:	Christopher J Plack
Co-author Contributions:	Co-supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 25/04/2025

Co-author Name:	Helen E Nuttall
Co-author Contributions:	Primary supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 28/04/2025

4.1 Abstract

Neural speech-envelope tracking in the auditory cortex – wherein neural activations synchronise with fluctuations in the speech envelope – is improved with audio-visual and audio-tactile speech compared to audio-only speech. Whilst visual cues can also improve speech intelligibility; evidence has shown no benefit behaviorally with tactile cues. Even with bottom-up short-term audio-tactile training, no further benefit to speech intelligibility was seen. Understanding the potential benefits of tactile stimulation is crucial for developing new methods of aiding speech perception in difficult listening conditions in the absence of relevant visual cues. We hypothesised that top-down audio-tactile training may provide benefit to speech intelligibility and further enhancements to tracking accuracy. Here, we present a preliminary pilot analysis of thirty young participants (ages: 19-30) who participated in a single-session electroencephalography experiment. Participants were given a sentence-in-noise recognition task with audio-tactile and audio-only stimuli. They then received either top-down audio-tactile training (discriminating between congruent and incongruent audio-tactile speech) or bottom-up audio-only training (passively listening to audio-only speech) in a single training session lasting 25 – 40 minutes, with feedback after trials. Finally, they completed the speech-in-noise recognition task again with audio-tactile and audio-only sentences. There was no significant interaction between timepoint (pre- and post-training) and group (audio-tactile training and audio-only training) for either tactile benefit to neural tracking accuracy (TbRz) or tactile benefit to speech intelligibility (TbSI). These preliminary findings suggest that a single-session top-down training paradigm is not sufficient for enhancing audio-tactile integration for either training group. However, data from a fully powered sample should be analysed before drawing final conclusions.

4.2 Does top-down audio-tactile speech-in-noise training affect speech-envelope tracking accuracy and intelligibility?

When listening to speech in difficult listening conditions, such as in background noise, visual information from lipreading can be integrated with auditory information from listening to improve speech intelligibility (Maier *et al.*, 2011). Visual cues also benefit neural tracking (Golumbic *et al.*, 2013), a process wherein neurons fire action potentials in synchrony with auditory features like the speech envelope (Heil & Peterson, 2015; Issa *et al.,* 2024). This is seen particularly in the delta and theta ranges (Bröhl & Kayser, 2020; Etard & Reichenbach, 2019) and is further highlighted with recent studies showing that facemask-wearing reduces tracking accuracies for audio-visual speech but not for audio-only speech (Haider *et al.*, 2024). Despite this, relevant visual information is not always present when listening. In these cases, the sense of touch may be useful for speech integration (Riecke *et al*., 2019; Fletcher *et al.*, 2020; Cieśla *et al*., 2022; O'Hanlon *et al.*, in prep., 2025). Cieśla *et al*. (2022) showed clear tactile benefits to speech intelligibility but found no significant difference between participants given audio-tactile training and participants given audio-only training. Whilst this was not investigated alongside neural tracking accuracy, this work does show tactile benefit to speech intelligibility both before and after training: there were no significant differences in this tactile benefit post-training between audio-tactile and audio-only training groups, however. On the other hand, using electroencephalography O'Hanlon *et al.* (in prep., 2025) found that, at baseline, audio-tactile speech produced an enhancement in neural tracking accuracy compared to audio-only speech. They found no tactile benefit at baseline to speech intelligibility, in line with similar audio-tactile research (Riecke *et al.*, 2019). Participants were then given three short training sessions in this experiment, each involving a speech discrimination task. During training, participants were presented with speech-in-noise and either relevant congruent tactile stimulation (trained group) or irrelevant

incongruent tactile stimulation (pseudo-trained group). No increase in tracking accuracy or speech intelligibility was seen in either group post-training.

It is known that speech perception involves a combination of dynamic bottom-up and top-down processing (Zekveld *et al.*, 2006; Diekhof *et al*., 2009). Indeed, both bottom-up and top-down focused training paradigms can successfully lead to improvements in speech intelligibility (Gohari *et al.*, 2023). As detailed by Gohari *et al.* (2023), a top-down training paradigm might include memory-based training (Ingvalson *et al.*, 2015; Schneiders *et al.*, 2012) or speech-in-noise training (Fletcher *et al.*, 2020; Ciesla *et al*., 2022; O'Hanlon *et al.*, in prep., 2025) to make use of top-down selective attention and contextual speech cues. Top-down modulation of bottom-up sensory decoding may occur through corticofugal projections (Asilador & Llano, 2021). Through these corticofugal projections, the auditory cortex utilises existing knowledge about our understanding of speech to influence bottom-up sensory processing, potentially through projections from the primary auditory cortex to subcortical generators like the inferior colliculus (Souffi *et al.*, 2021; Oberle *et al.*, 2022; Ford *et al.*, 2024). A bottom-up training paradigm on the other hand might include temporal integration (Zerr *et al.*, 2019) or phonemic training (Schumann *et al*., 2015) to make use of sensory speech processing.

Arguably, the training provided by O'Hanlon *et al.* (in prep., 2025) engaged both bottom-up and top-down processes, with auditory specific training through speech-in-noise discrimination utilising top-down processes and the audio-tactile integration through tactile stimulation utilising bottom-up processes. Namely, the speech-in-noise discrimination task used presented participants with target speech embedded within noise and asked them to identify keywords from said target speech in response. As described by Gohari *et al.* (2023), training with speech-in-noise engages top-down selective attention processes (see also Talsma *et al.*, 2010) to attend to the target speech whilst suppressing noise and can also draw

upon top-down contextual cues from working memory to aid in discrimination of keywords in a sentence (Lad *et al.*, 2020; Vermeire *et al.*, 2019). The tactile stimulation, however, was not crucial to identifying the correct keywords in the task, as the tactile element was assumed to benefit listening through congruent exposure alone and thus temporally integrated. Therefore, training to integrate the tactile stimulation with the target speech was likely driven by bottom-up sensory processing. Evidence suggests that there is a dynamic interaction between bottom-up and top-down processing in speech perception, with top-down processes modulating bottom-up in competition for limited resources (Amitay *et al.*, 2014; Huang & Elhilali, 2020). This may explain why audio-tactile training was ineffective, as the auditory-focused top-down processes during training may have modulated the tactile focused bottom-up processes. Further evidence shows that bottom-up training of audio-tactile integration has not always been successful; Rizza *et al.* (2018) trained participants to associate tactile stimulation with specific speech phonemes, finding no evidence of this form of bottom-up associative training leading to integration between the auditory and tactile sensory inputs.

Conversely, electro-haptic stimulation training in cochlear implant users led to improvements in speech intelligibility using a similar training paradigm to O'Hanlon *et al.* (Fletcher *et al.*, 2020). Electro-haptic stimulation is different to the tactile stimulation used in other audio-tactile based training studies (Rizza *et al.*, 2018; Riecke *et al.*, 2019; Cieśla *et al.,* 2022; O'Hanlon *et al.*, in prep., 2025), as it utilises stimulation to areas of the forearm rather than through envelope-shaped stimulation to the fingertips, but still encapsulates a form of audio-tactile integration. The effectiveness of similar training seen in Fletcher *et al.* may be due to the sample population of cochlear implant users tested, as evidence shows cochlear implant users to have enhanced multisensory integration versus normal-hearing listeners (Rouger *et al*., 2007). Thus, these cochlear implant users may have been more tuned to utilising haptic information to benefit speech perception than non-cochlear implant users.

Another point of consideration is the configuration of noise localisation within this study; electro-haptic benefits from training were only seen when the noise was located separately to the presented speech (ipsilateral or contralateral to the participant's implant) and not present when the noise accompanied the speech in a central location. This could further highlight a discrepancy between top-down and bottom-up training elements. For the central location, it would have been difficult to identify whether the haptic stimulation was congruent to the target speech or the noise during training, as both were present from the same source. In the case of the lateral conditions, where the target speech and noise are separated in source location instead, the haptic stimulation may have been more easily perceived as congruent to the target speech and incongruent to the noise, resulting in trained top-down selective attention of the haptic element in these conditions (Talsma *et al*., 2010). When looking at other top-down training paradigms, it was shown by Woodruff *et al*. (2024) that top-down training improved sensory processing speeds compared to bottom-up training, which did not show processing speed improvements at all, potentially mirroring electro-haptic benefit in lateral versus central localisation. With audio-tactile training, a top-down variant might include presenting both congruent and incongruent tactile stimulation during training and requiring participants to actively attend to both types of stimulation to determine which one was congruent. This would ensure that top-down attentional processes are utilised for audio-tactile integration during listening. Thus, a top-down audio-tactile training paradigm may provide benefits to neural tracking and speech intelligibility that were not seen in O'Hanlon *et al*. (in prep., 2025) for a normal-hearing population. We tested the following primary hypotheses:

(i): Benefit from speech-shaped tactile stimulation to cortical tracking of the speech envelope in-noise will be enhanced post-training versus pre-training for those who have

undergone top-down training with audio-tactile stimulation, compared to those given audio-only training.

(ii): Tactile benefit to speech intelligibility of speech-in-noise will be enhanced post-training versus pre-training for those who have undergone top-down training with audio-tactile stimulation, compared to those given audio-only training.

4.3 Materials and Methods

4.3.1 Participants

Participants were recruited from the Lancaster University campus and the surrounding Lancaster area. 30 participants passed the eligibility criteria and took part in the study (ages: 19-30 years; $M_{age} = 21$; $SD_{age} = 2.32$; 12 males, 18 females). Whilst 60 were determined in sample size calculations (see below) for sufficient power, here a preliminary pilot of 30 participants is presented, with 15 in the audio-tactile training group and 15 in the audio-only group. As such, the results of this experiment are preliminary, and discussion is limited to findings with insufficient power until a full dataset is analysed. Inclusion criteria were that all participants: were between the ages of 18 and 35, were right-handed only, were monolingual native speakers of British English, and had normal hearing. Hearing was measured using pure tone audiometric air conduction testing, bilaterally across 250 to 8000 Hz frequencies and following the British Society of Audiology guidelines (BSA, 2018). All participants had calculated hearing thresholds below 20 dB HL. Exclusion criteria were that no participants: had nerve damage to the fingertips, motor problems in the hands, missing fingers on their hands, or peripheral neuropathy in the hands or fingertips. Participants were screened for eligibility using Qualtrics before the experiment began. Ethical approval was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (approval reference: FST-2024-0766-SA-1, project ID: 0766).

4.3.2 Sample Size Calculations

Sample size was calculated using 'Power R Sample Size (*pwrss*)', an R based shinyapp for calculating power and statistical sample sizes (Bulus, 2023: https://pwrss.shinyapps.io/index/). A small to medium effect size of R squared = .17 was assumed for this study and a Bonferroni-corrected significance criterion of $p < .025$ was used

as we were testing two main hypotheses. A logistic regression sample size calculation was used on *pwrss* as this was the closest matching option to our experimental design. 60 participants were deemed sufficient for a power of .80 at an alpha level of .025, with 30 in each group.

4.3.3 Experimental Design

Participants took part in a single electroencephalography (EEG) session. Participants were randomised into either the 'Audio-tactile Training' group or the 'Audio-only Training' group before testing began. Throughout the pre-training task, training task, and post-training task, participants were presented with a variety of sentences in noise, both audio-only and audio-tactile. This led to a 2 (Training Group; audio-tactile training, or audio-only training) x 2 (Timepoint; pre-training, or post-training) mixed-factors design to evaluate these data and test two primary hypotheses.

4.3.4 Materials

All target sentence and noise audio were wideband, with sampling rates of 44.1 kHz and bandwidths of up to 22.05 kHz. Four-talker babble noise was created first using sentences from the Clarity Speech Corpus (Graetzer *et al.*, 2022), containing a mixture of male and female voices at random. This was done using a MatLab script that randomly selected four sentences from the corpus of various durations, took a random 4000 ms segment from the selected sentences, and then placed all four segments onto the same waveform track. Ten 4000 ms audio files were created using this script, which were used in experimental trials. The target sentence stimuli used sentences also from the Clarity Speech Corpus, which in turn were selected from the British National Corpus, with 210 selected in total. However, as these needed to be all from the same speaker and of a similar duration, it was decided that the chosen sentences would be re-recorded using the primary researcher as the target speaker.

The sentences were recorded along with a 10-minute segment of Alice in Wonderland that was also spoken by the primary researcher. Recordings were made using a HyperX Quadcast external USB microphone, set to a cardioid polar direction. Target sentences were between 3400 and 3650 ms long and always played 250 ms after the noise began. All sentences and noise files were loaded into Audacity together and normalised to 60 dB SPL. The noise files were then presented in the experiment between -10 and 10 dB SNR at 0.5 dB increments, creating 41 possible difficulty levels relative to the 60 dB SPL target sentences.

The SNR selected for the participant was calculated using a speech-in-noise test. This was a custom adaptation of the QuickSIN test (Etymonic) used to personalise the difficulty of the experiment so that participants could correctly recognise approximately 50% of keywords from target sentences, referred to as the Speech Recognition Threshold (SRT). Typically, during QuickSIN, the participant would be required to verbally respond to full sentences and correctly responded keywords would be recorded and used in SNR calculations. However, in this study, participants discriminated between keywords via a button press (see Procedure). This meant that the participant could select correct keywords by chance, resulting in the experiment being set at a lower SNR and potentially lowering response accuracy down towards chance level. Furthermore, the original QuickSIN sentences were spoken by a female speaker, which can be harder to discern in mixed-gender multi-talker noise than a male speaker such as was used for the experiment's target sentences (Larsby *et al.*, 2015). Therefore, the QuickSIN test was adapted so that the method of response and the target speaker of the test matched the target speaker of the experiment. A modified calculation method was used which added an additional +2 dB to each participant's final SRT scores, due to previous experiments with these stimuli showing a lower speech intelligibility baseline of approximately 40% (O'Hanlon *et al.*, in prep., 2025). Therefore, to increase this baseline closer to expected SRT accuracy, the following formula was used:

- SRT = Starting dB level + 2 + (dB step value/2) - Total Keywords Correct
- Therefore, SRT = 5.75 - Total Keywords Correct

Tactile stimulation was provided to the right index finger of participants in audio-tactile and training conditions using a lab-built tactile device. This device used a hard drive accentuator to generate small horizontal movements to the finger based on slow fluctuations in electric potential, which were set to match slow fluctuations in the speech envelope. The device used a soft ring around the finger to keep it in place during stimulation and was insulated around the area of contact with participants to reduce potential electrical interference with the EEG. Furthermore, it was attached to a handrest to make stimulation more comfortable throughout the experiment. For the creation of audio-tactile stimuli, speech envelopes were extracted from all 210 sentences using Praat (Boersma & Weenink, 2021) via Hilbert transformation. These extracted envelopes were then loaded back into Audacity to be normalised. For audio-tactile sentences, the sentence and its respective envelope were combined into a single waveform file. This was done by having the left channel of a stereo track be the sentence audio, and the right channel just the sentence envelope. For audio-only sentences, the sentence was placed in the left channel and the right channel was left silent. When played, the left channel was split to a pair of EEG-compatible insert earphones (Etymonic ER3-14A Ear Tips) which played the audio in mono format to both ears, whilst the right channel was split to the tactile stimulation device. These earphones had transducers that were housed in electromagnetic shielding to prevent stimulus artefacts in the data.

An eligibility screening and initial consent form were created and hosted using Qualtrics (Qualtrics, 2005). The experiment was designed using PsychoPy Builder (v2022.1.4; Peirce *et al.*, 2019) with custom coding elements. Conditions on the pre- and post-training tasks were counterbalanced, with some participants experiencing the audio-tactile condition first and others the audio-only condition. No sentences were repeated across

any task or trial. Sentences order lists were randomised into 10 possible sets which were counterbalanced between training groups. To ensure that the same noise file played alongside the same sentences to avoid potential biases with some sentences being harder to understand in some noise files, all sentences were paired with one of the possible 10 noise files. Therefore, when sentences were randomised into sentence sets, they came attached with the same noise files across those sets for consistency. Sentence and noise files were played to the participant with the same onset during the same PsychoPy routine. A 32-channel BioSemi EEG kit (BioSemi, NL) was used to record neural activity whilst the participant listened to stimuli. More channels were not needed for reconstruction validity, as shown by Montoya-Martínez *et al.* (2021). The kit utilised an ActiveTwo AD-box amplifier and recorded brain activity at a 2048 Hz sampling rate. ActiveTwo 32 channel standard caps were used with standard 32-channel 10%-system electrode locations. Cap sizes used ranged from 52 to 60 cm head circumference. Signa Gel electrolyte gel was used to improve electrode impedance, with acceptable impedance values targeting 30 kΩ or less during setup which was monitored using ActiView software (BioSemi, NL).

4.3.5 Procedure

Participants were blinded regarding their group assignment. Participants completed pre-screening via Qualtrics (Qualtrics, 2005). Once participant eligibility was confirmed, participants were invited to the lab. A pure tone audiometry assessment was conducted using a calibrated audiometer and following BSA guidelines (British Audiology Society, 2018). Participants were then introduced to EEG and the tactile device used for the experiment. They were presented with two randomly selected sentences with relevant tactile stimulation. Participants' individual tactile perception threshold was measured by adjusting the force of the tactile device and using a subjective assessment of participants' ability to distinguish the movements of the device relevant to the speech they were listening to and if the device was

comfortable. If the intensity needed adjusting, it was done by the researcher. Once a threshold was found, the device was set above threshold by one incremental level. This tactile intensity level was recorded and was kept consistent for each participant throughout all their testing sessions.

4.3.6 Speech-in-Noise Test

Participants then took part in the customised speech-in-noise test. To ensure that any potential distracting effects of the EEG cap were considered during the calculation of the participant's individualised SNR, the EEG cap was set up for the speech-in-noise test despite not recording data. Participants were also asked to place their right index finger into the tactile stimulation device. Again, this device was not switched on during this speech-in-noise test. However, having the index finger of their dominant hand placed on an unusual device may affect speech discrimination performance in later tasks, and so was considered when calculating the participant's SRT. The researcher provided the verbal instructions for the test alongside written instructions that the participant had to read and click through to proceed to the task. Participants were told that they were to listen for the researcher's voice and to ignore the background noise of other people speaking at the same time. The researcher was present for every session conducted in the study so that their voice as the target sentence was recognisable. The same example sentence was given to all participants: 'The dog ran down the long road'. They were then told that after hearing the sentence in noise, four keywords would pop up on the screen, one in each corner. One of these keywords was the first keyword heard in the target sentence, whilst the other three were semantically or phonetically similar. In the context of the example target sentences, the correct choice was clarified to participants as 'dog', and other options may be 'god' or 'cat'. They were told to press the corners of the number pad to select the corners of the screen. As an example, if 'dog' was in the top right, they would press the top right of the lab keyboard's number pad, which was always '9'. If it

was in the bottom left, they were told to press ‘1’. As explained to the participant, once the selection was made, the four keywords would disappear and four more would show up. One of these would be the second keyword, ‘ran’. Participants would keep making selections until all five sets of keywords were presented. After this, the next target sentence would play.

For the speech-in-noise test, there were six trials in each block: the SNRs were always: 10, 5, 0, -5, -10, and -15 dB. Participants were told to guess if they were unsure. There were five blocks in total. The first two blocks acted as practice and familiarisation for the participant. The last three blocks were used to calculate the participant’s SRT for the experiment. This was done by summing the number of correctly discriminated keywords from each sentence per block and subtracting that value from 14.5. This was based on the formula from the QuickSIN method (see Materials). As this was done for each of the three test blocks, the average SRT was calculated between them to provide the final SNR value that was then used for every session with the participant moving forward.

4.3.7 Tactile Familiarisation

Next, participants were given brief familiarisation with the tactile device. This was to provide a period of exposure with the tactile stimulation before the main task began, allowing an extended opportunity for participants in both training groups to attempt to understand how the tactile stimulation related to target speech. Familiarisation was done at the phoneme, word, and sentence level to provide participants with as much opportunity to relate the envelope-shaped tactile stimulation with the listened-to speech as much as possible. Starting at the phoneme level, participants were given audio-tactile stimulation of a phoneme, such as ‘ba’ or ‘ka’. The phoneme was visually displayed on the screen during listening. Then, a full word related to that phoneme played, such as ‘baker’, repeating three times. And finally, a full short sentences akin to those in the task, such as ‘The baker was baking loaves of bread’

played, repeating twice. Each participant completed four sets of familiarisation trials, with each set including phonemes, words, and sentences.

4.3.8 Pre-Training Task

The pre-training task acted as the baseline and consisted of three elements: a short passive listening task using a story excerpt from 'Alice in Wonderland' (see 'Materials'), a block of 30 speech sentence discrimination trials in noise that were audio-only, and a block of 30 sentences in noise that were audio-tactile. The order of the audio-only and audio-tactile blocks was counterbalanced across participants, whilst the story was always presented as the first part of the task. The story segment was presented without noise and was split into two five-minute parts, with a break and a content question after each to ensure participants were paying attention. This was a multiple-choice answer question, such as 'What did the rabbit pull out of its waistcoat?', with four answers to select from using a mouse click. Participants could not proceed without selecting the correct answer. In a case where participants failed to select the correct answer, they listened to the segment again with the same content question. EEG data were recorded throughout this task. Once the story task finished, the researcher would let the participants know whether they would be next listening to audio-tactile or audio-only sentences. This was so that participants were not surprised if the device was active during sentence listening. Participants were reminded at this stage to try not to blink during speech listening and to remain as comfortable as possible. The audio-only and audio-tactile sentence tasks worked the same as the speech-in-noise test, except that the SNR for all sentences was set based on the participant's SRT level. Furthermore, after each sentence trial, the participant was given a voluntary break screen and a progress bar indicating how many sentences were left for the block. These were implemented to ensure that participants had ample opportunity to get comfortable and take a break as they needed, as well as an opportunity for the researcher to adjust any electrodes that may have gone noisy mid-testing.

4.3.9 Training Task

The training task consisted of a single session that was approximately 25 – 40 minutes in duration. Here, both training groups received a different training task. In both groups, two sets of 30 sentences were presented, with the first set being speech played without any noise and the second set being speech played with the same four-talker babble as the previous tasks at the same SNR. For the audio-tactile training group, each sentence was played twice. On either the first or second playing of the sentence, congruent tactile stimulation was provided that matched the sentence heard. The other sentence was accompanied by incongruent tactile stimulation that matched an irrelevant sentence to the one heard. The participants were then asked to select the sentence for which the audio and tactile stimulation correctly matched by pressing either 'a' or 'b' on a keyboard (two-alternative forced choice), with 'a' referring to the first sentence that played and 'b' the second. After making their selection, they were presented with a feedback screen which let them know if their response was correct, were played the sentence again with the congruent stimulation, and had the sentence visually written in full on the screen for them to review. This third playing of each sentence in the feedback stage was always played without noise, even in the speech-in-noise condition. Participants in this group were required to correctly identify all congruent sentences in both the set with and without noise. If a sentence was incorrectly identified, it was added back into the set to play again until the response to the sentence was correct. Once all 30 sentences without noise and 30 sentences with noise were correctly identified, the training ended. This typically occurred within 25 – 40 minutes. If 45 minutes had passed and not all sentences had been correctly identified, the training ended. This was due to limited resources and to prevent exhaustion. Two participants in this group ended their training without identifying all sentences in noise, with one failing to identify four sentences in noise and another five.

For the audio-only training group, participants also listened to 30 sentences without noise and 30 sentences with noise. Each sentence was played twice. This time, however, no tactile stimulation was provided and there were no differences between either playing of each sentence. Instead, participants were advised to passively observe each sentence and listen as closely as possible. There was no active task to complete, although a 'feedback' screen was still shown to participants after each trial, which played the sentence again without noise and presented the sentence in full visually to the participant. To ensure that participants were paying attention and reviewed every sentence in the 'feedback' stage, they would occasionally be presented with a follow-up question after a trial. This question would present a single keyword, such as 'dog', and ask participants to select either 'y' for 'yes' or 'n' for 'no' as to whether this keyword was spoken in the trial they last listened to.

4.3.10 Post-Training Task

The post-training task began immediately after training and consisted of the same speech discrimination task from the pre-training task, with audio-tactile and audio-only sentence blocks again in a counterbalanced order All sentences used across the experiment were unique. At the end of this task, participants were debriefed on the true aims of the study and paid to compensate them for their time.

4.4 Statistical Analyses

4.4.1 Variables

There were two independent variables. The training group variable refers to which training group the participant was placed in, either the audio-tactile training group or the audio-only training group. The timepoint variable refers to when the data was collected in the experiment: either pre-training or post-training. The training group was a between-subjects factor, with 15 participants placed in the audio-tactile training group and 15 in the audio-only training group. For dependent variables, both tactile benefit to speech intelligibility (TbSI) and tactile benefit cortical speech-envelope tracking accuracy (TbRz) were measured. These were derived from speech intelligibility (SI) and tracking accuracy (Rz) respectively. SI was defined as the percentage of correctly discriminated keywords in a sentence trial, which was averaged over all 30 sentence trials in a condition. TbSI was then calculated by taking the SI of the audio-tactile and audio-only sentence conditions in both the pre- and post-training tasks and using the following formula:

(Audio-tactile SI – Audio-only SI) / (1- Audio-tactile SI)

To calculate Rz, the multivariate temporal response function toolbox (mTRF, Crosse *et al.,* 2016) was used. This process involved utilising a decoder function in the mTRF toolbox to reconstruct an estimation of the target sentence speech envelope based on the inputs of collected neural data and then correlate this estimated envelope with the original stimulus envelope. This correlation was used as the measure of Rz. Similarly to TbSI, TbRz was calculated using audio-tactile and audio-only conditions by the following:

(Audio-tactile Rz – Audio-only Rz)/(1- Rz)

4.4.2 Pre-processing and Decoding

EEG data was pre-processed using EEGLab (Delorme, & Makeig, 2004) in MatLab. Initially, the data were recorded at a sampling rate of 2048 Hz, with no online filtering. Data were recorded throughout each condition, with a new recording file being made per condition. Using EEGLab, the data were first resampled to 100 Hz and filtered using a Finite Impulse Response (FIR) filter with a low pass at 1 Hz, before independent components analysis (ICA) was run. The spheres and weight matrices outputted by the ICA were saved to be used for a future decomposition. This method of early ICA was selected as our target frequency range of 0.5 – 15 Hz included delta below 1 Hz, which is susceptible to slow-drift distortion with extended infomax ICA (Pontifex *et al.*, 2017). The raw data were then reloaded back into EEGLab and resampled to 100 Hz again. The data were filtered to our target range next using a FIR filter, with a low pass at 15 Hz and a high pass at 0.5 Hz. Next, the data were re-referenced using the average. The previously decomposed ICA weights and spheres that were determined from the first loading of the data were then placed on this second iteration. ICLabel (Pion-Tonachini *et al.,* 2019) was used to automatically flag components for muscle, eye, heart, line-based, and channel-based noise, with boundaries for all set at 85%. These flagged components were then removed before finally the stimulus presentation periods were extracted using the onset and offset of each sentence file played. To remove the onset of event-related potentials, the first second of each sentence trial was removed. The result was a three-second epoch per trial. For the 'Alice in Wonderland' story segments, the same pre-processing steps were used. However, the first two seconds of each five-minute segment were removed instead, resulting in two 298-second epochs.

4.4.3 Speech-envelope Tracking Accuracy (Rz)

Rz was obtained using the stimulus reconstruction method via the multivariate Temporal Response Function toolbox in MatLab (Crosse *et al.*, 2016). This method of

reconstruction uses a backwards approach with a decoder for the neural data. For cross-validation, the method of 'leave-one-trial-out' was chosen (see Riecke *et al.*, 2019). As we were using a low SNR that matched participants' individual SRTs, outputs of the reconstruction method were expected to be lower for sentence trials than in previous literature. Furthermore, due to the quicker sentence duration, each sentence trial could not provide enough EEG data alone for valid reconstruction. The required amount of EEG data for valid envelope reconstruction is not entirely clear in the literature, with some referencing 60 seconds as sufficient for 87.5% accuracy (Biesmans, *et al.*, 2016). A comparative look between EEG and Magnetoencephalography (MEG) suggests that EEG requires as much as three times the duration of MEG for valid reconstruction, coming to approximately 120 seconds (Destoky, *et al.*, 2019). It is essential to provide as much EEG data to the decoder as possible, with a minimum duration of somewhere between 60 and 120 seconds in mind. This meant that for all 30 sentences in a condition, we would need every epoch available in that condition to be combined for a more reliable reconstruction. By stitching together epochs, however, we run the risk of training the decoder on the 'seams' of the individual epochs, which may provide inefficient decoder parameters when it comes to calculating the final speech-envelope tracking accuracy value. To alleviate this issue and the issue of low SNR during sentence listening, the two five-minute story segments in clear speech were used to train the decoder first and output optimal parameters for the reconstruction of the 30 stitched-together sentences. This provided an optimal regularisation parameter (lambda) and number of 'folds' or 'segments' (nf), which were then applied as the parameters for sentence reconstruction. Reconstruction outputs were averaged across all leave-one-trial-out validations to provide a final Rz value for each session's conditions.

4.4.4 Model for Analysing Neural Data

For testing hypothesis (i), linear mixed-effects models (LMERs) were used taking TbRz as the dependent variable. For our first hypothesis, we expected that RbRz would see a greater increase post-training for the audio-tactile training group, compared to the audio-only training group. The ID of the participants and the sentence list assigned to them were loaded as random factors. The LMER model was as follows:

TbRz ~ Group + Timepoint + Group*Timepoint + (1|ID) + (1|sentence))

To accept this hypothesis, we would expect to see a significant interaction between group and timepoint, with this interaction showing significant increases in TbRz in the audio-tactile training group post-training compared to the audio-only training group.

4.4.5 Model for Analysing Behavioural Data

For hypothesis (ii), a generalised LMER model (GLMER) was used as our accuracy scores were bound based on choice-selection in the speech discrimination task. We expected that RbSI would see a greater increase post-training for the audio-tactile training group, compared to the audio-only training group. The ID of the participants and the sentence list assigned to them were loaded as random factors. The GLMER model was as follows:

TbSI ~ Group + Timepoint + Group*Timepoint + (1|ID) + (1|sentence))

To accept this hypothesis, we would expect to see a significant interaction effect between group and timepoint, with this interaction showing significant increases in TbSI in the audio-tactile training group post-training compared to the audio-only training group.

4.4.6 Pre-registration

The study was pre-registered on the Open Science Framework (OSF) before data collection began. Further details on the data simulation methodology and the preregistration

itself can be found at: https://osf.io/38kqt. There was a deviation from this pre-registration involving the wording of the two hypotheses (i, and ii), which originally read as there being an enhancement of TbRz (i) and TbSI (ii) post-training versus pre-training for those given audio-tactile training, “but not for those given audio-only training”. Upon reflection, these could be interpretated as two hypotheses each, as they would require two tests to reject the null: one to see if audio-tactile training increased TbRz and TbSI respectively, and another to show that audio-only training had no impact on TbRz and TbSI respectively. These hypotheses were modified to state “compared to those given audio-only training” instead, to better convey the single test per hypothesis chosen above that relies on a significant interaction effect to reject the null.

4.5 Results

4.5.1 Effect of Top-Down Training with Tactile Stimulation on Speech-Envelope Tracking Accuracy

Table 1 shows the mean Rz pre- and post-training of audio-tactile and audio-only sentences for both the audio-tactile training group and the audio-only training group. Figure 1 shows the TbRz differences from pre- to post-training for both the audio-only (pre-training mean = -.008, post-training mean = -.01) and audio-tactile (pre-training mean = -.008, post-training mean = +.01) training groups. The preliminary LMER analysis for testing hypothesis (i) indicated no significant main effect of group ($\beta = .00$, $t = -.001$, 95% $CI = [-.02, .08]$, $p > .05$), timepoint ($\beta = -.01$, $t = -.28$, 95% $CI = [-.04, .05]$, $p > .05$), or the interaction between group and timepoint ($\beta = .03$, $t = .81$, 95% $CI = [-.09, .04]$, $p > .05$). Hence, these data do not provide support for hypothesis (i).

4.5.2 Effect of Top-Down Training with Tactile Stimulation on Speech Intelligibility

Table 2 shows the mean SI pre- and post-training of audio-tactile and audio-only sentences for both the audio-tactile training group and the audio-only training group. Figure 2 shows the TbSI differences from pre- to post-training for both the audio-only (pre-training mean = +.7%, post-training mean = -18%) and audio-tactile (pre-training mean = -2%, post-training mean = -18%) training groups. The preliminary GLMER analysis for testing hypothesis (ii) indicated a significant main effect of timepoint (pre-training or post-training; $\beta = -.20$, $t = -3.52$, 95% $CI = [-.09, -.31]$, $p < .025$). There was no significant main effect of group (trained or pseudo-trained; $\beta = -.05$, $t = -.80$, 95% $CI = [-.16, .07]$, $p > .05$), or the interaction between group and timepoint ($\beta = .05$, $t = .57$, 95% $CI = [-.11, .20]$, $p > .05$). Hence, these data do not provide support for hypothesis (ii).

4.6 Discussion

Top-down audio-tactile training was used to investigate potential tactile benefit to speech intelligibility and neural tracking accuracy in comparison to audio-only training. This differed from previous work by O'Hanlon *et al.* (in prep., 2025), which investigated tactile training for audio-tactile speech with a paradigm that did not involve selective attentional processes for distinguishing tactile stimulation in the training task, as only congruent tactile stimulation (trained group) or incongruent tactile stimulation (pseudo-trained group) was present during a speech-in-noise discrimination task. In this top-down audio-tactile training study, participants were assigned to either an audio-tactile training or audio-only training group and were presented with audio-tactile and audio-only sentences in noise during a speech discrimination task both pre- and post-training. It was hypothesised that the audio-tactile training group, when presented with audio-tactile sentences post-training, would see an increase compared to the audio-only training group in tactile benefit to neural tracking accuracy (TbRz) for hypothesis (i) and tactile benefit to speech intelligibility (TbSI) for hypothesis (ii). Results uncovered no significant benefit to TbRz or TbSI post-training for either group. Whilst means and standard deviations of each condition do show a general increase in speech intelligibility post-training for audio-only sentences in both groups (see Table 2), this was not the case for audio-tactile sentences. These results however are preliminary, as only 30 out of the target sample of 60 participants were tested at the time of analysis.

As such, until a suitable sample size for power is obtained, no conclusion can be made to reject hypothesis (i) and (ii). These preliminary results do offer early insight into the lack of effectiveness of the new top-down audio-tactile training paradigm. Currently, with no significant change present in TbRz or TbSI post-training in either group, it can be speculated that speech-relevant tactile stimulation was not successful in benefitting neural

representations or our understanding of speech in difficult listening conditions for this limited sample. Training benefits were seen on average for audio-only sentences alone (see Table 2), however. The lack of tactile benefit may be due to the study taking place over only a single session, as opposed to a multisession training study with multiple sessions over multiple days, as with O'Hanlon *et al*. (in prep., 2025). Allowing time for participants to consolidate memory after training may have provided further tactile benefits (Atienza *et al.*, 2002; Molloy *et al*., 2012). Whilst top-down training has even been beneficial to speech perception immediately after training ended and was retained for at least a week after (Drouin & Theodore, 2022), consolidation after sleep has been evidenced to stabilise training benefits faster than without in the context of decoding noise-vocoded speech (Drouin *et al*., 2023). Drouin *et al.* (2023) presented two groups with the same training paradigm for understanding noise-vocoded speech, with one group starting training in the morning and another in the evening. Participants took part in a pre-training, training, and immediate post-training task akin to this paper's present procedure. Participants then had a second post-training assessment 12 hours later, which for the morning group was the same day and the evening group the following morning. Here, Drouin *et al*. found that there was immediate training benefit in both groups, but only the evening group maintained these benefits in the 12-hour post-assessment. In a 1-week follow-up, both groups saw training benefits stabilise. Likely, the morning group experienced fatigue due to a period of wakefulness following training, resulting in a loss of training benefit that was restored following sleep consolidation in the 1-week follow-up. Given the long duration of the present top-down audio-tactile experiment, taking approximately 2 hours and 30 minutes to complete, it may be that participants in the audio-tactile training group did gain tactile training benefit, but the length of the study session stretched into a similar period of consequential wakefulness to that of Drouin *et al*.'s

morning group, leading to a decrease in audio-tactile performance immediately post-training. It may be that tactile benefit would have been seen in this group after sleep consolidation.

On the other hand, the lack of sleep consolidation post-training does not adequately explain why performance with audio-only sentences seemed to improve immediately post-training for both groups regardless of task length. One alternative explanation for the lack of tactile benefit could be that the audio-tactile training group may not have been fully able to discern differences between the congruent and incongruent stimulation during said task, as not all participants were able to successfully identify the congruent stimulation for all 60 sentences within 45 minutes. Although, if this was the case, it would be expected that more participants than only two would have failed to complete the training in time. Furthermore, there is evidence to suggest that auditory learning is strengthened with more difficult, demanding training tasks (Ahissar & Hochstein, 1997; Moore & Amitay, 2007). More recent evidence shows that training is even further strengthened with easy-to-hard presentation, wherein easier trials within training are presented first to participants before more difficult trials (Wisniewski *et al.*, 2019; Wisniewski *et al.*, 2024). The top-down training task presented to the audio-tactile training group in this experiment reflected an easy-to-hard presentation by presenting clear speech trials before speech-in-noise trials. As this is a preliminary analysis, however, further sample testing would be required to speculate on training difficulty based on the proportion of fully completed training tasks in this training group. A more plausible explanation for the results is that the top-down audio-tactile training was beneficial to the pool of sentences used within the training session only. These sentences would have repeated upon an incorrect response, leading to multiple repeats in the pool that a participant found difficult to initially discern tactile congruency with. It is entirely possible as a result that tactile benefit from this training may only be present for this same or similar set of sentences, with no transfer of learning seen to untrained sentences in noise (see also:

Bieber & Gordon-Salant, 2021; Bugannim *et al.*, 2019; Banai & Lavner, 2019). As the post-training task always used a pool of different sentences to the pre-training and training tasks, this could explain why tactile benefit did not improve post-training.

Currently, the preliminary implications of this experiment highlight a lack of tactile benefit from top-down audio-tactile training. These findings represent underpowered analyses and cannot be taken as a conclusion until a full sample size is collected. Looking at effect sizes, however, the corrected marginal (variance explained by the fixed effects only) r-squared values (see Nagakawa *et al.*, 2017) for the LMER for TbRz was very low (r^2 = .03) in comparison to priori sample size calculations (r^2 = .17). The GLMER analysis for TbSI (r^2 = .17) was comparable to priori sample size calculations (r^2 = .17). This may indicate that further analysis on a fully powered sample could see significant changes in the analysis for TbRz, whereas the results of TbSI analyses are less likely to change with an increased sample. If these findings remain consistent with a powered sample, future research should focus on potential differences in tactile stimulation delivery to understand why benefit was not seen here with speech-envelope shaped tactile stimulation to the fingertips versus other works using electro-haptic (Fletcher *et al*., 2020) and vibro-tactile (Ciesla *et al.*, 2022) devices. Furthermore, a multisession training approach should be considered to ensure that tactile benefits are not masked by a lack of sleep consolidation post-training. It would also be beneficial to test the impact of using the same stimulus set post-training versus a stimulus set different to the training pool, as this would provide crucial information on the generalisability of tactile training benefits to wider speech environments.

4.6.1 Conclusion

In conclusion, the preliminary analysis presented for this study fails to identify notable tactile benefit to either neural tracking accuracy or speech intelligibility of speech-in-noise sentences post-training. This is apparent for both the audio-tactile and audio-only

training groups. This work provides further conflicting evidence of the benefit of audio-tactile speech integration using speech-envelope shaped tactile stimulation to the fingertips. This may highlight this method's insufficient ability – even with top-down training - to provide tangible tactile benefit to audio-tactile speech compared to other forms of stimulation such as electro-haptic and vibro-tactile. Future research should investigate other possible tactile devices and training paradigms to improve neural representations and speech intelligibility in difficult listening environments. However, these conclusions should be considered with care until a fully powered sample is obtained and these preliminary findings are re-analysed.

Tables and Figures

Table 1.

Means and Standard Deviations (SD) of neural tracking accuracies pre- and post-training in audio-tactile and audio-only conditions, for both the audio-tactile training group and the audio-only training group.

		Pre-training		Post-training	
Group	Sentence Type	Mean (Rz)	SD (Rz)	Mean (Rz)	SD (Rz)
Audio-tactile Training	Audio-tactile	.06	.04	.07	.07
	Audio-only	.07	.04	.06	.05
Audio-Only Training	Audio-tactile	.09	.06	.07	.03
	Audio-only	.10	.05	.09	.04

Table 2.

Means and Standard Deviations (SD) of speech intelligibility accuracies pre- and post-training in audio-tactile and audio-only conditions, for both the audio-tactile training group and the audio-only training group.

		Pre-training		Post-training	
Group	Sentence Type	Mean (%)	SD (%)	Mean (%)	SD (%)
Audio-tactile Training	Audio-tactile	54.89	27.38	55.38	29.12
	Audio-only	56.58	27.19	63.47	26.97
Audio-Only Training	Audio-tactile	55.11	27.50	51.42	27.78
	Audio-only	55.20	29.37	59.73	28.28

Figure 1.

Boxplots showing the median and interquartile ranges for the tactile benefit to tracking accuracy (TbRz) pre- and post-training in both the audio-tactile and audio-only training groups.

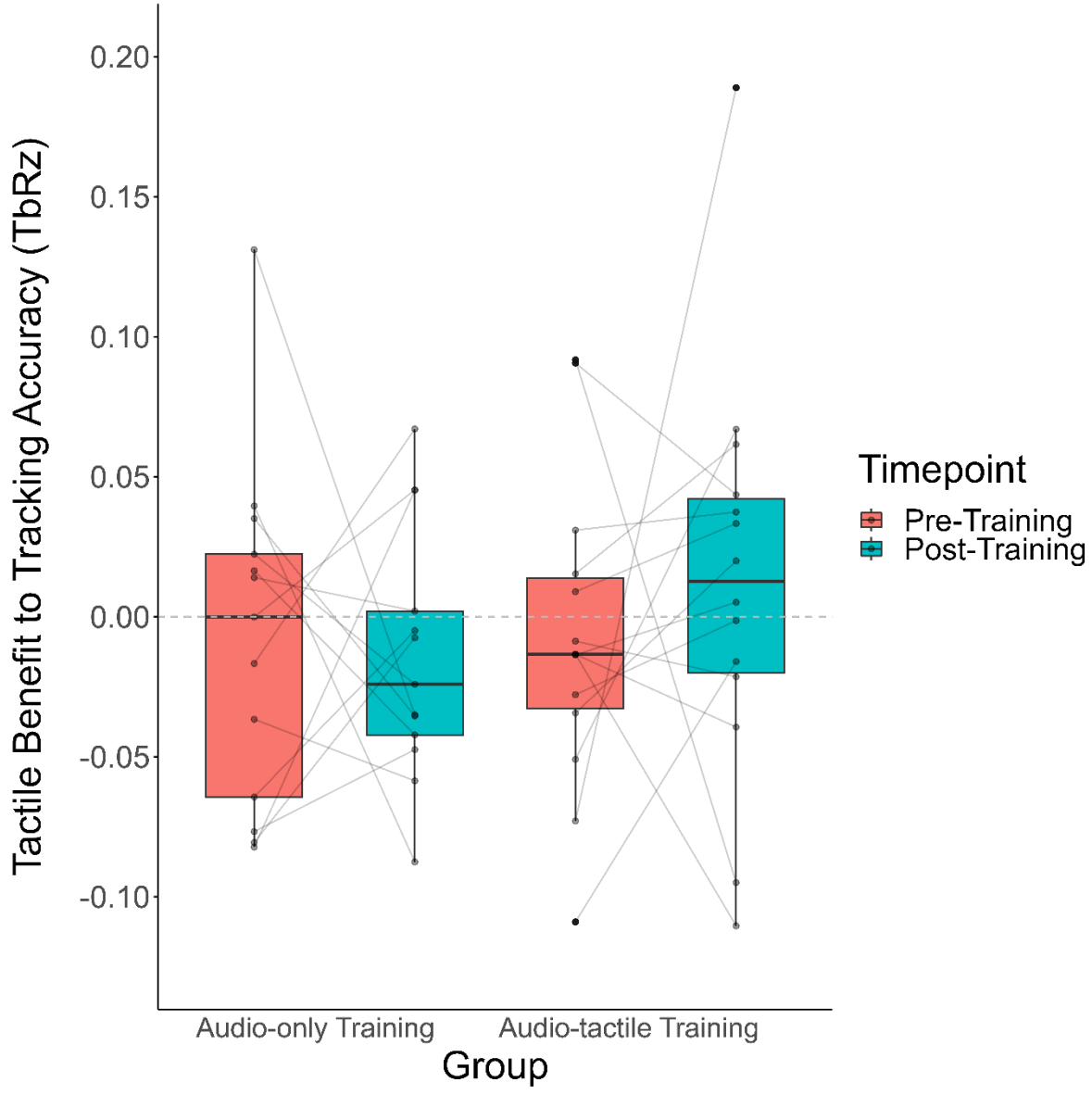
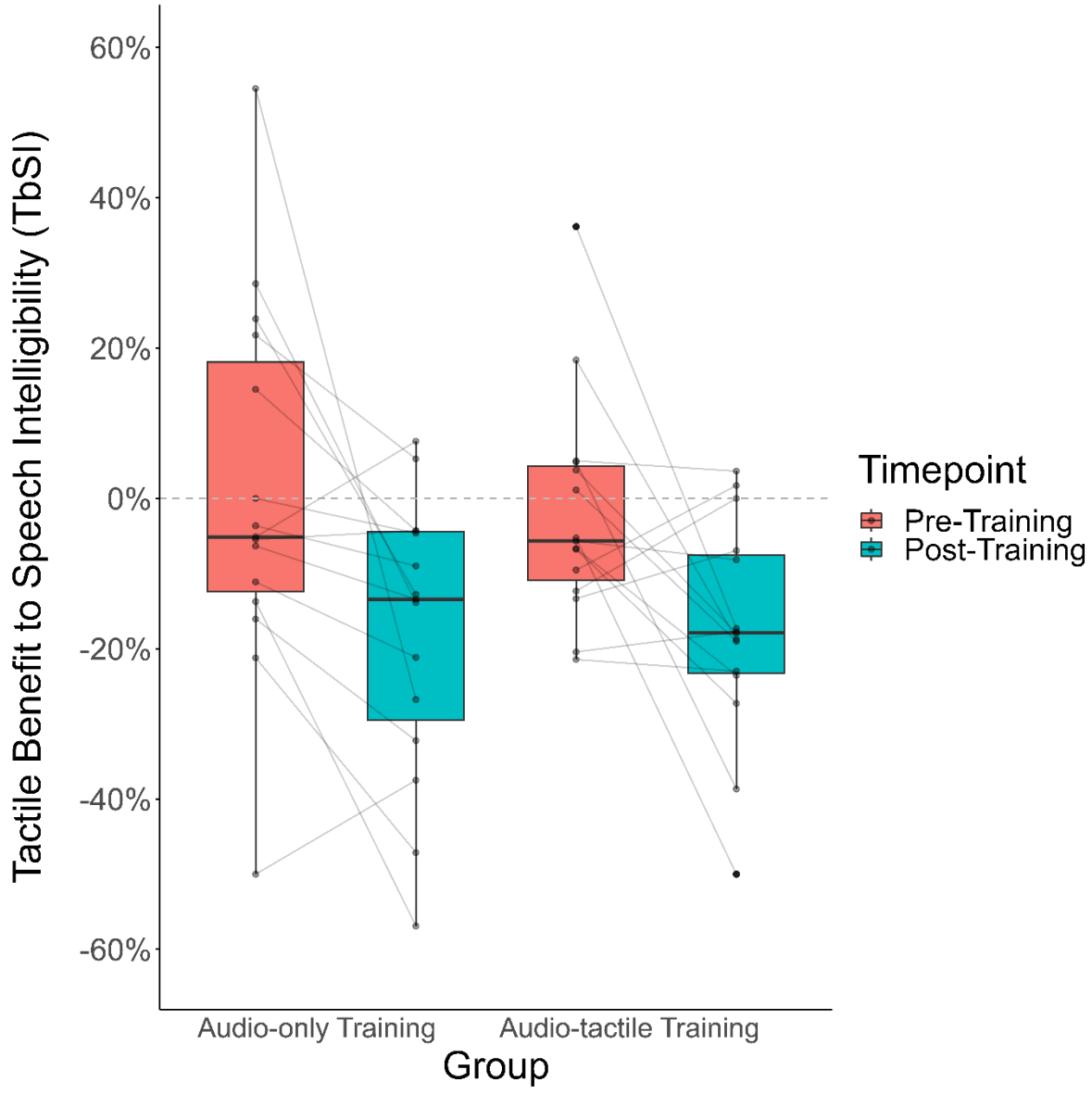


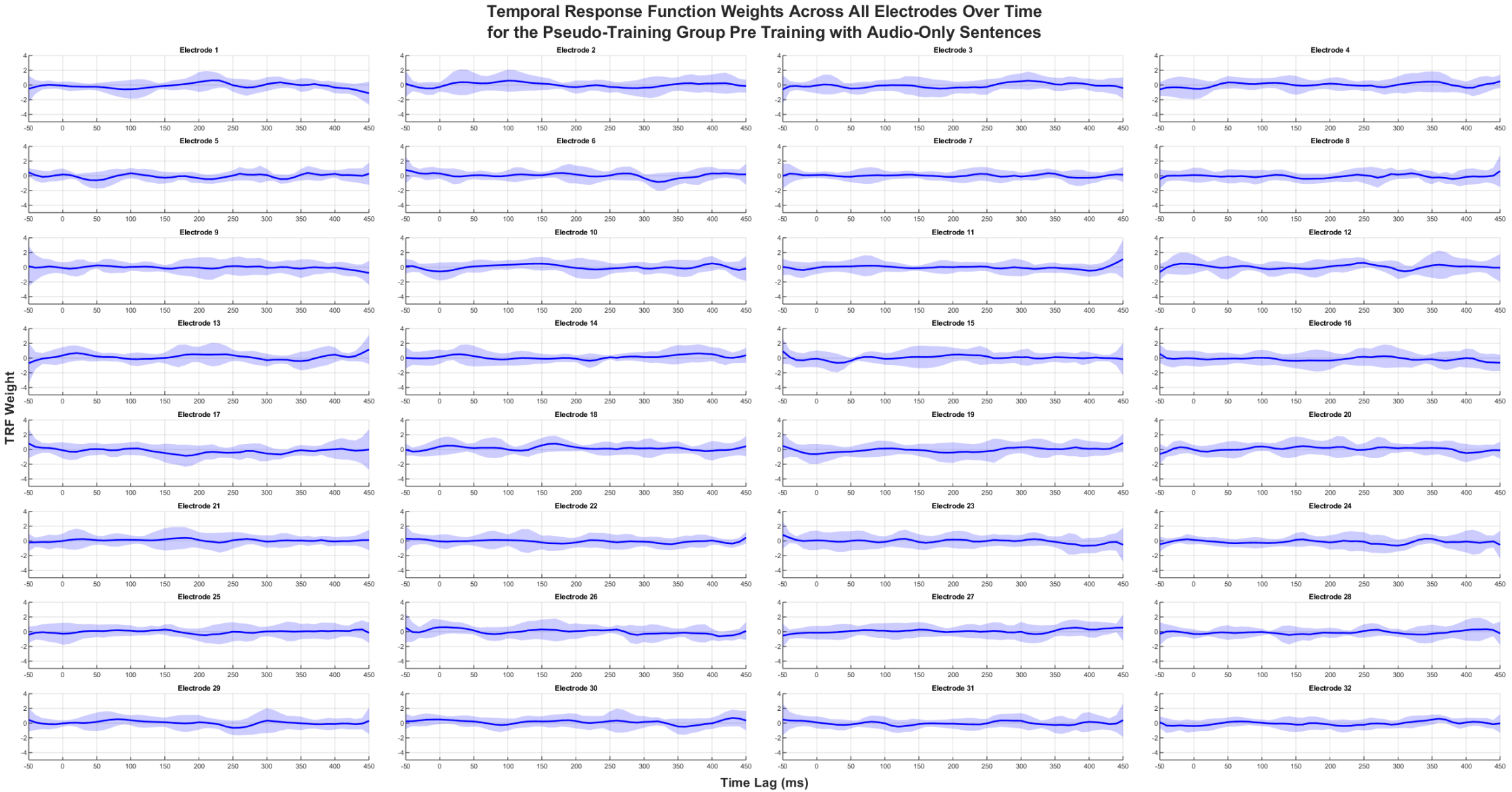
Figure 2.

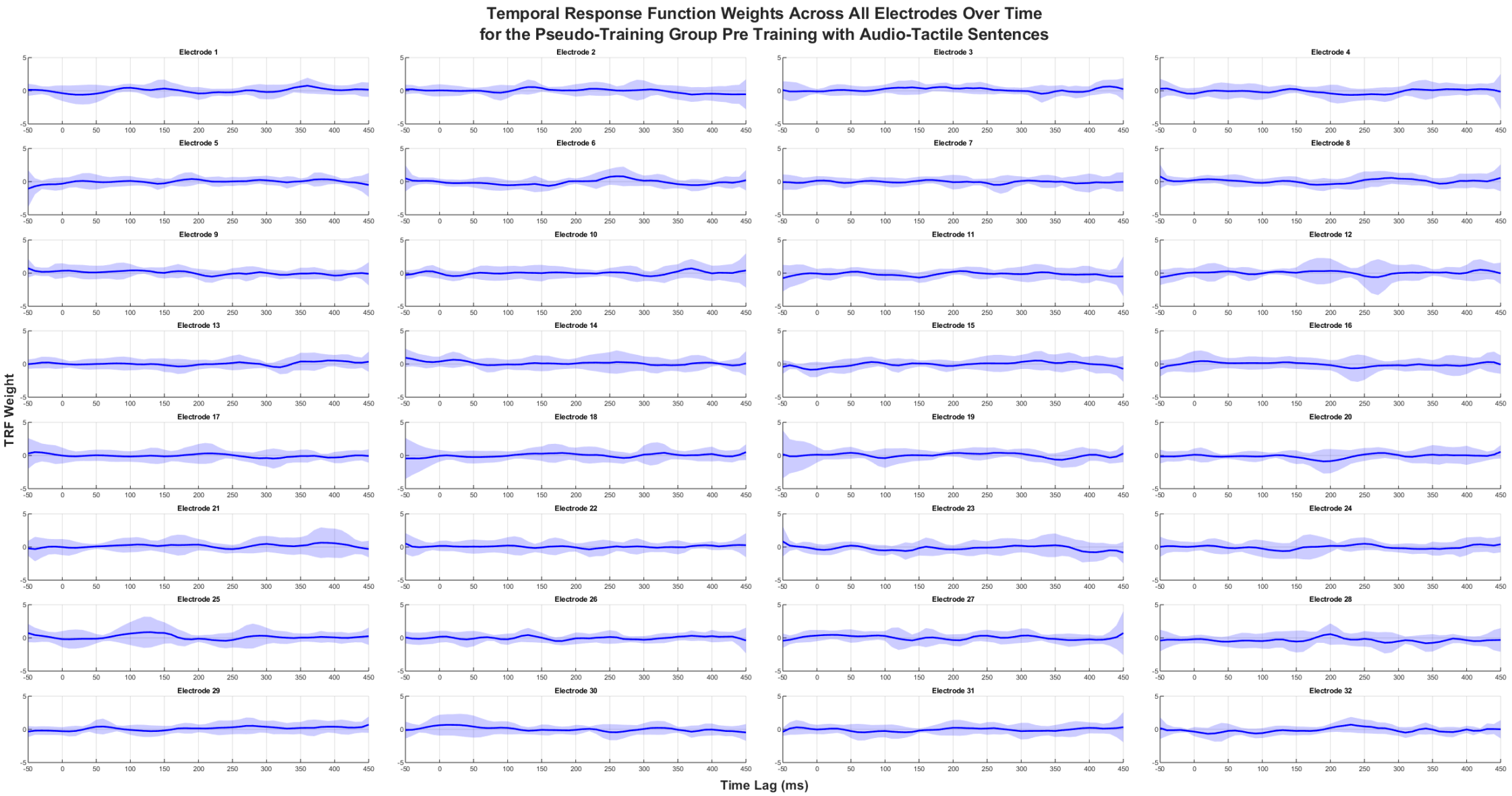
Boxplots showing the median and interquartile ranges for the tactile benefit to speech intelligibility (TbSI) pre- and post-training in both the audio-tactile and audio-only training groups.

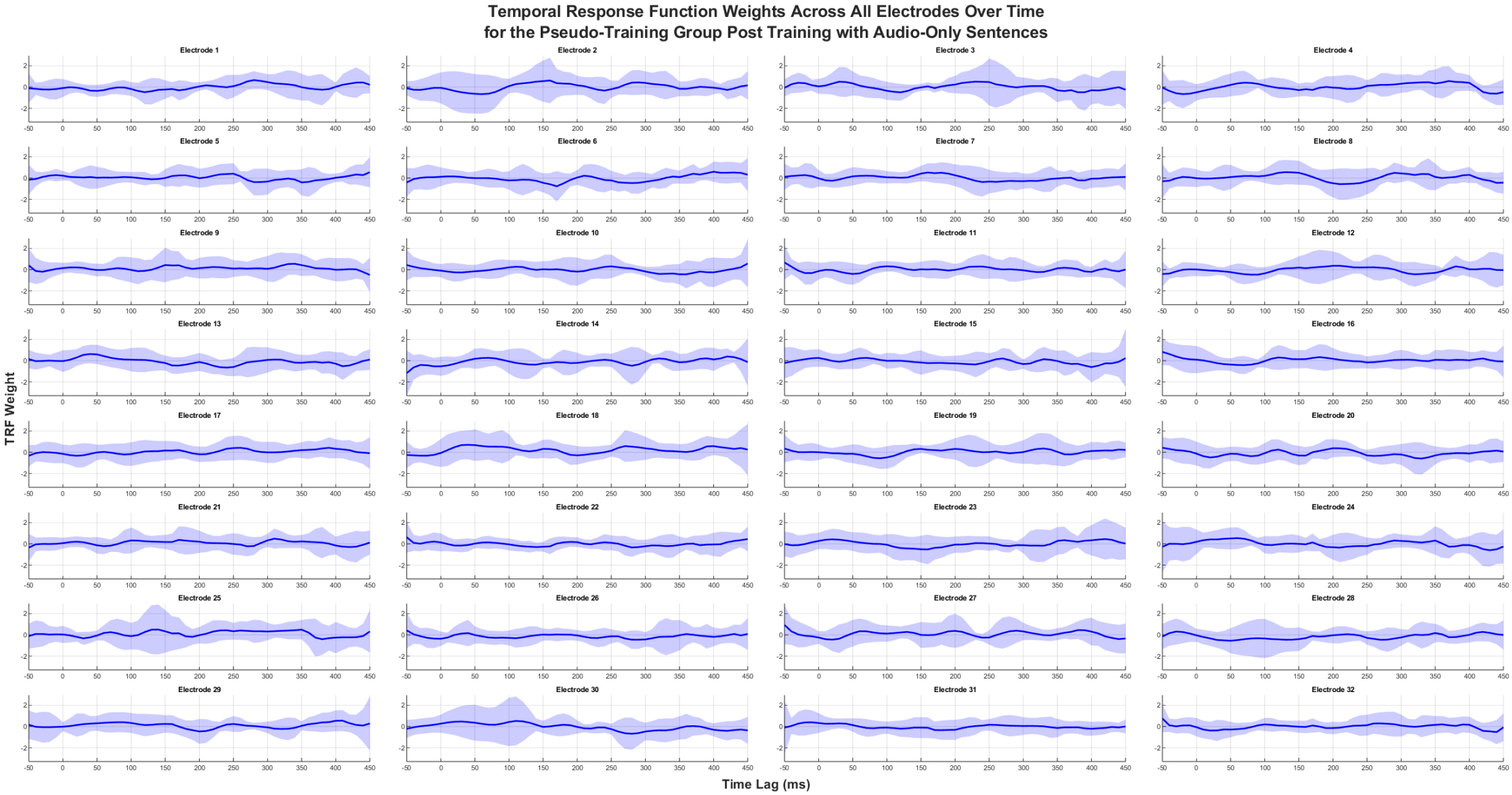


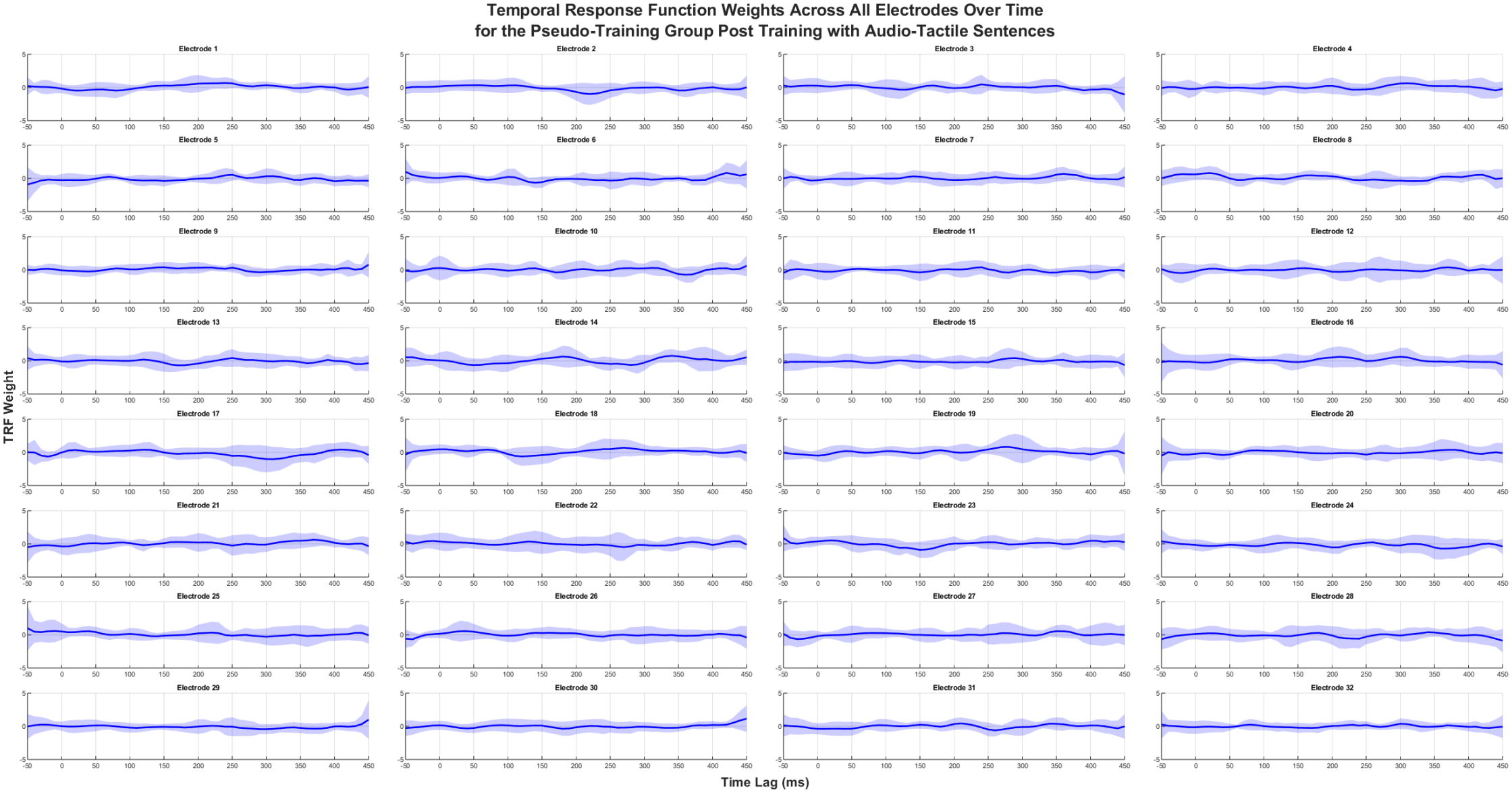
Supplementary Materials A

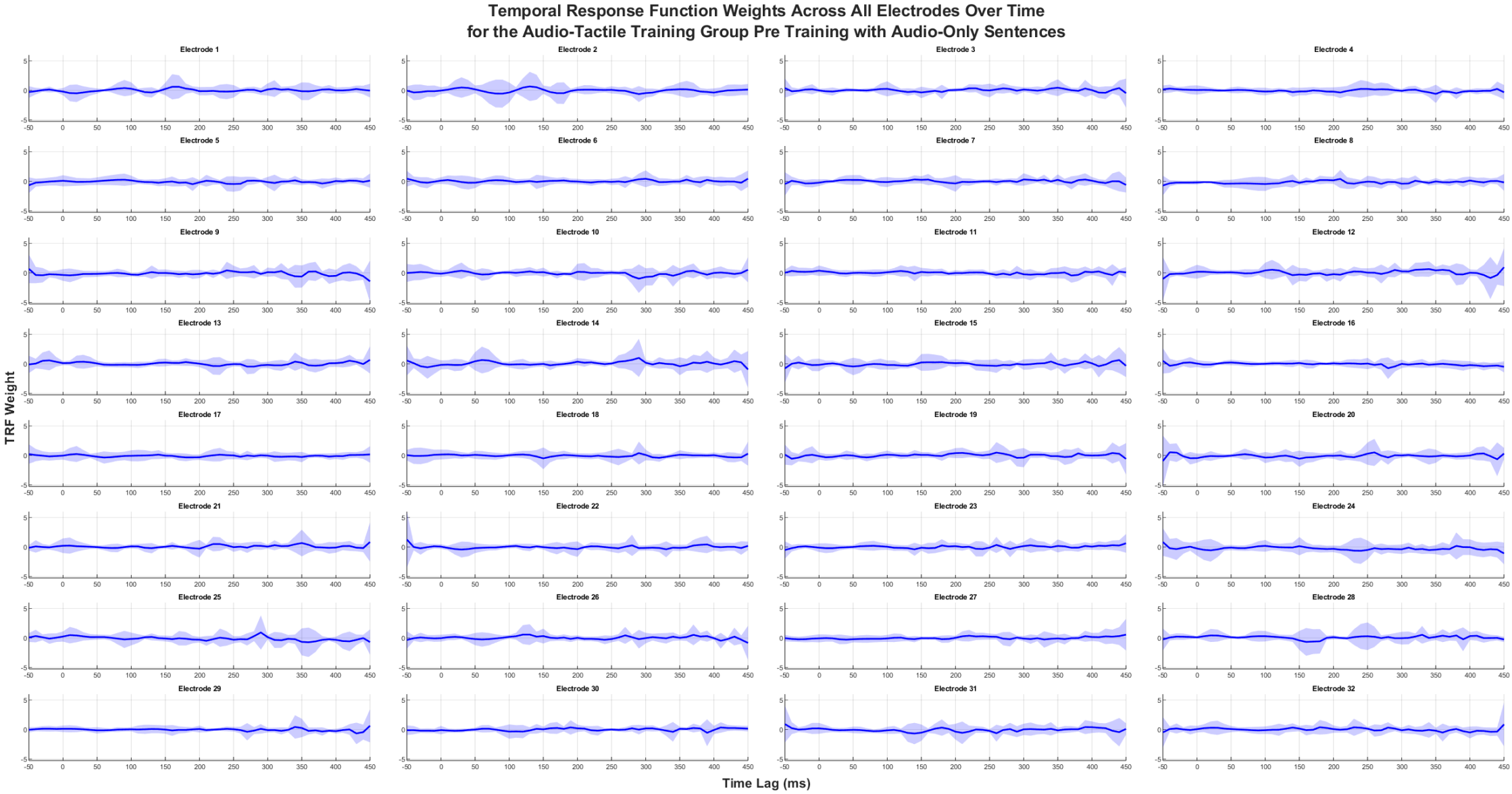
The following figures show the temporal response function (TRF) weights across all 32 electrodes for each training group (audio-only training, and audio-tactile training), session number (pre-training, and post-training), and stimulation type (audio-only, and audio-tactile). In each figure, the solid blue line represents the grand average mean of TRF weightings across all participants, with variance displayed as one standard deviation away from the mean.

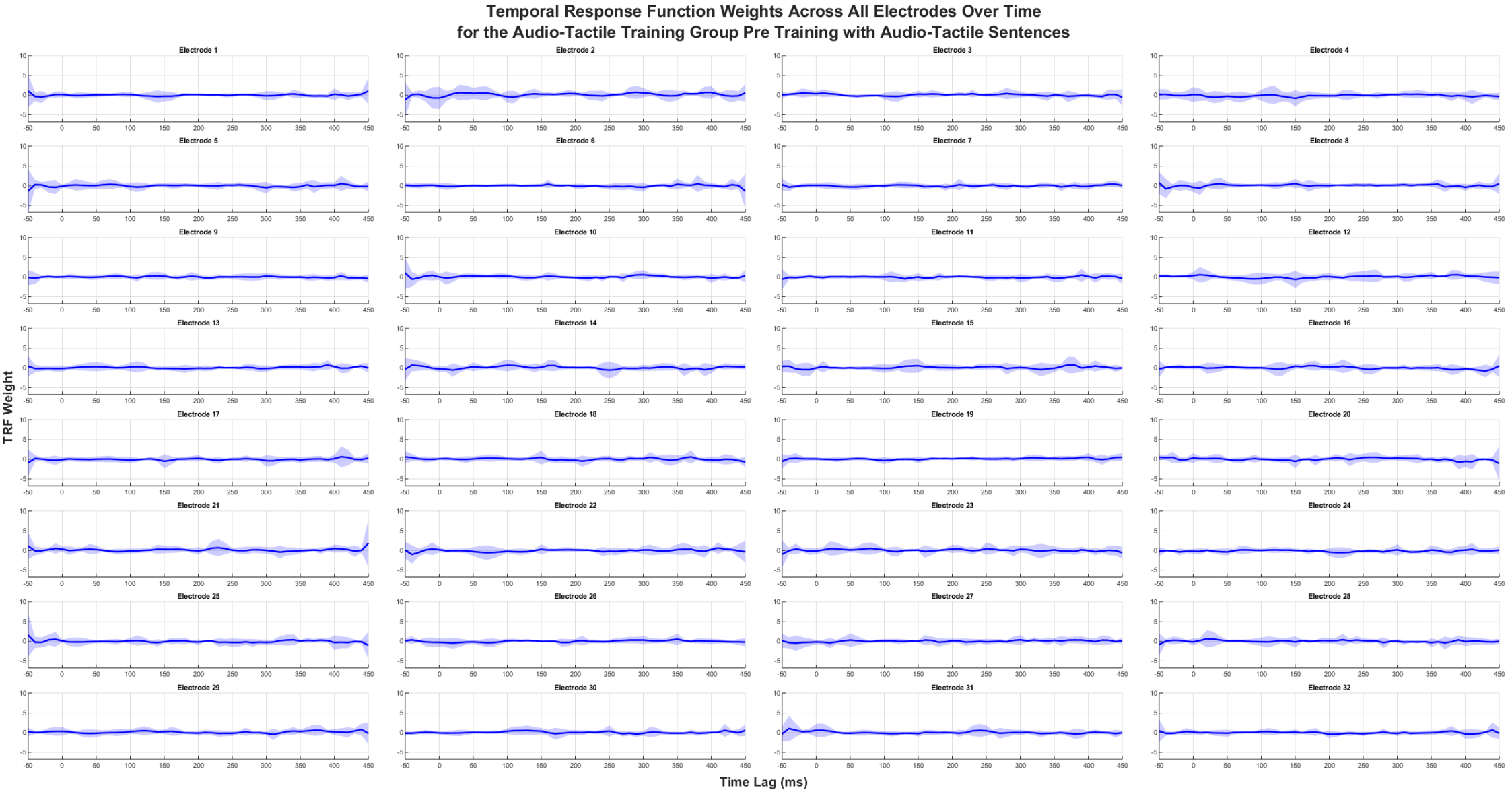

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Pre Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

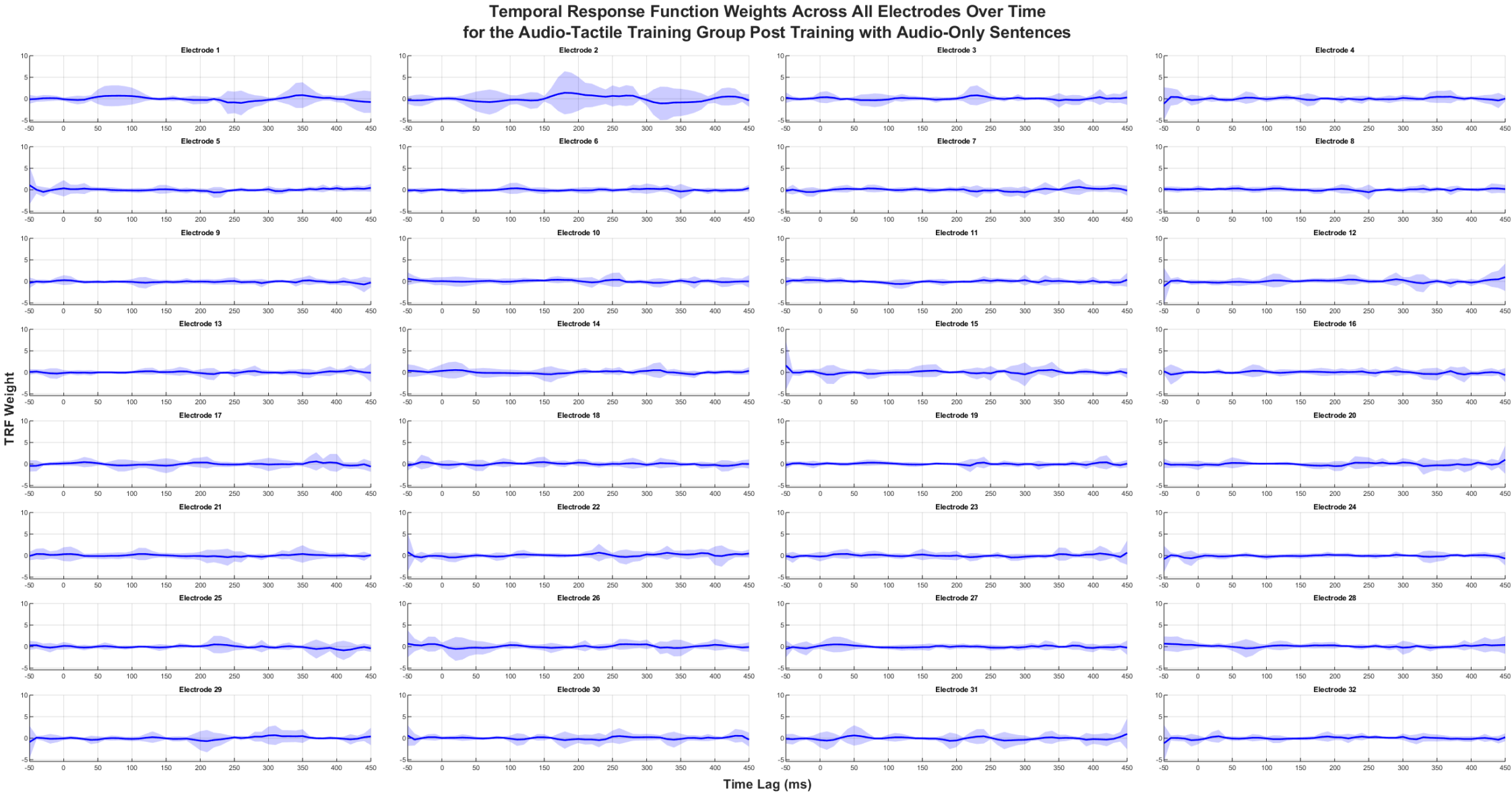

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Pre Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

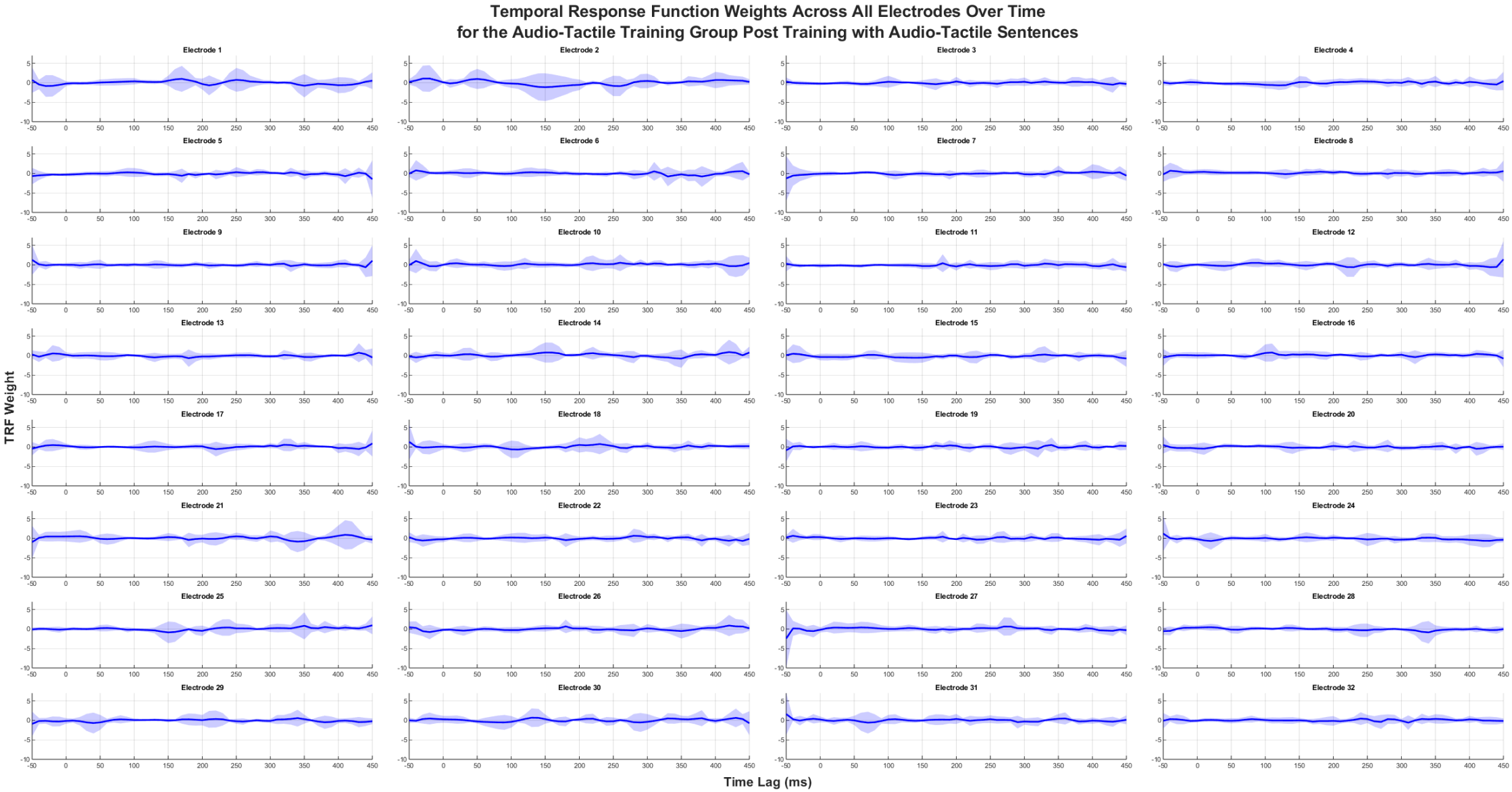

Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Post Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for the Pseudo-Training Group Post Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Pre Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

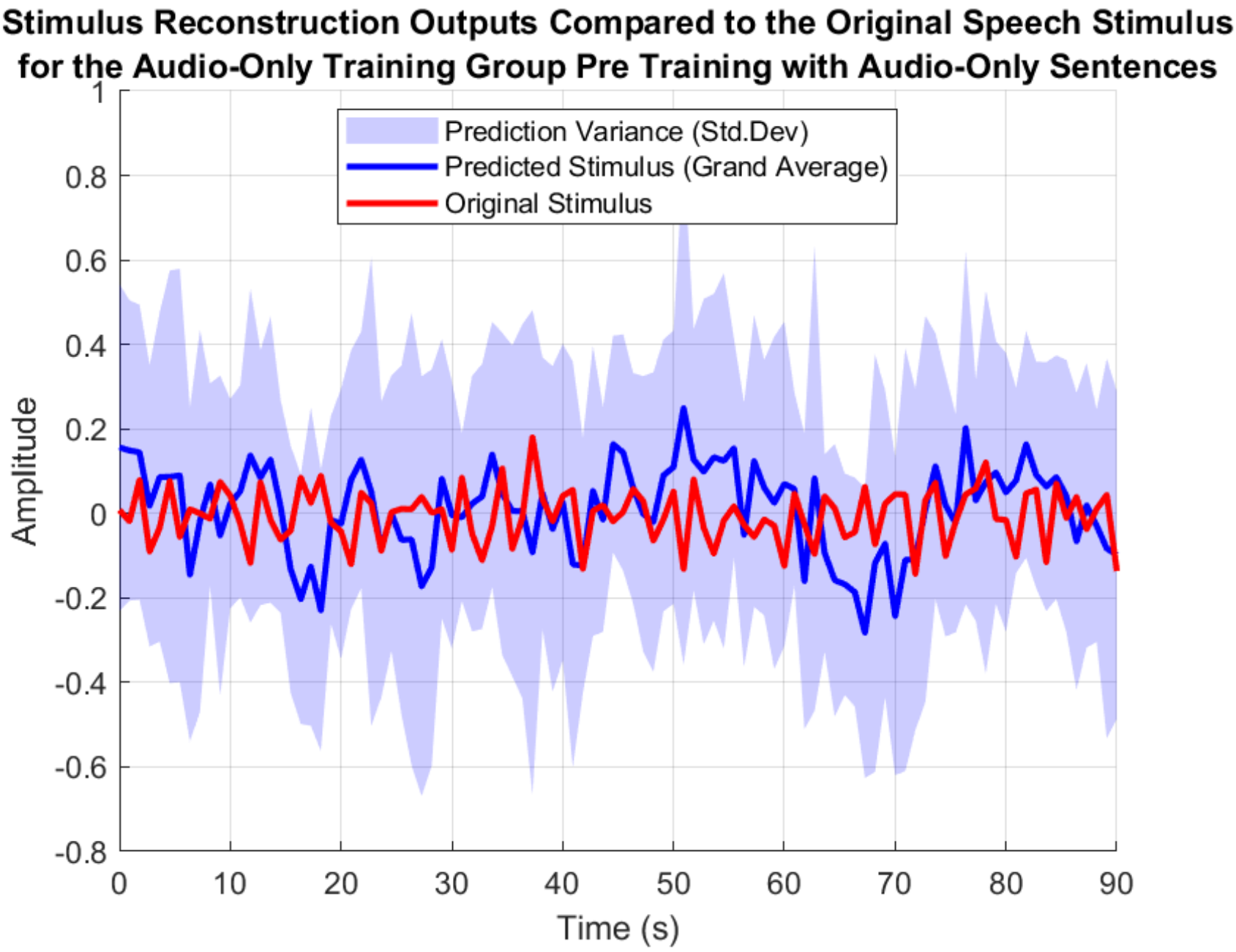

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Pre Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

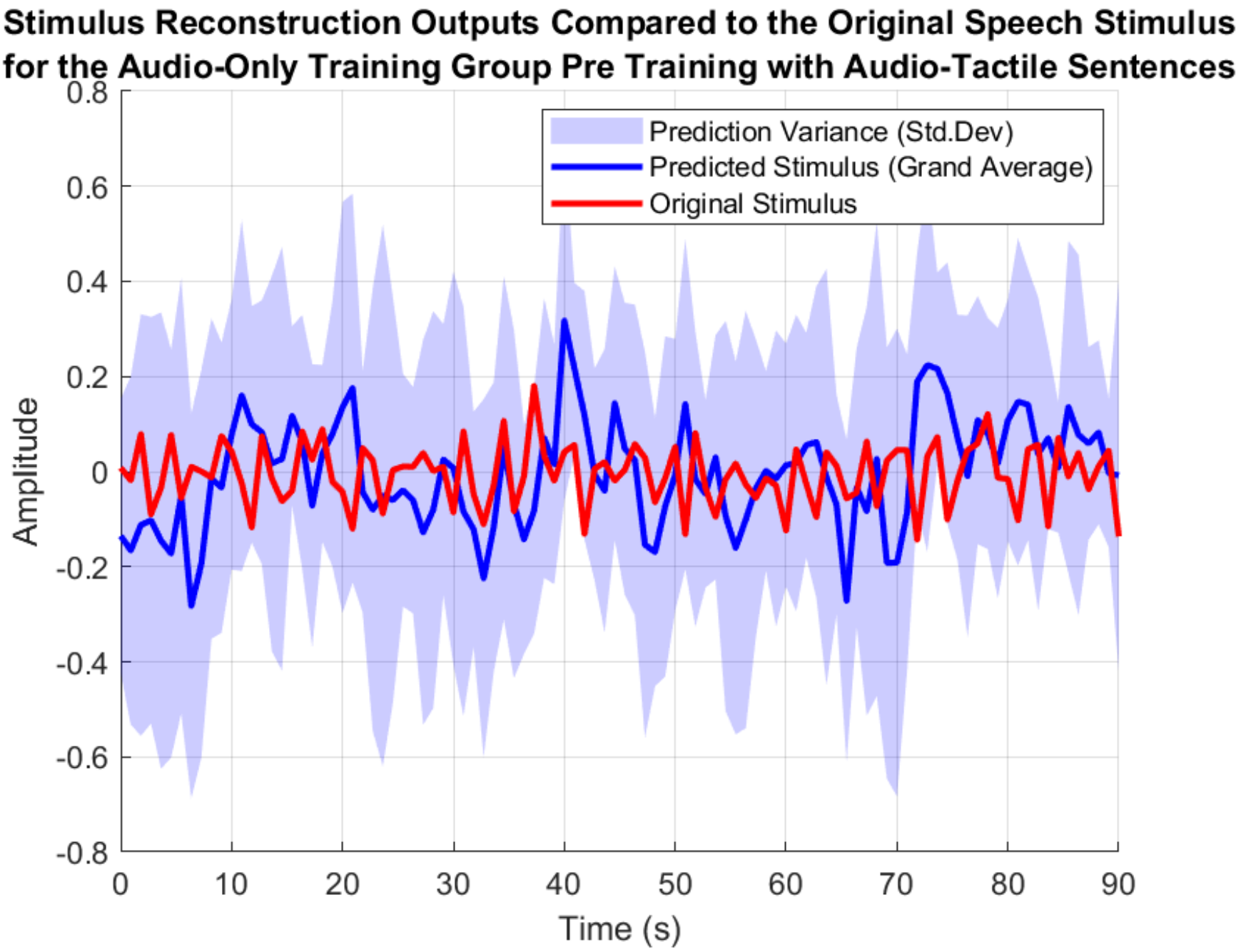

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Post Training with Audio-Only Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

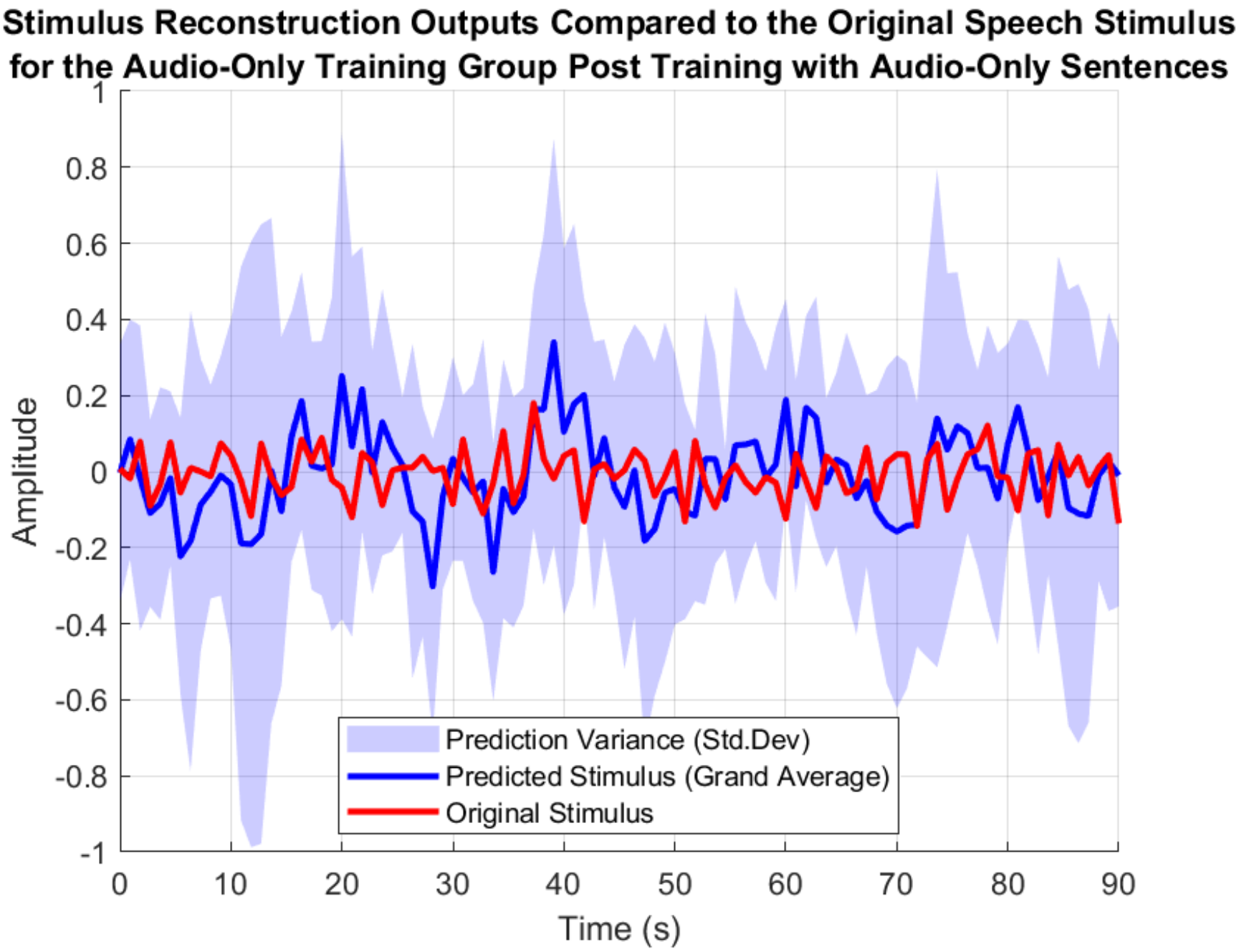

Temporal Response Function Weights Across All Electrodes Over Time
for the Audio-Tactile Training Group Post Training with Audio-Tactile Sentences
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
Electrode 7
Electrode 8
Electrode 9
Electrode 10
Electrode 11
Electrode 12
Electrode 13
Electrode 14
Electrode 15
Electrode 16
Electrode 17
Electrode 18
Electrode 19
Electrode 20
Electrode 21
Electrode 22
Electrode 23
Electrode 24
Electrode 25
Electrode 26
Electrode 27
Electrode 28
Electrode 29
Electrode 30
Electrode 31
Electrode 32
TRF Weight
Time Lag (ms)

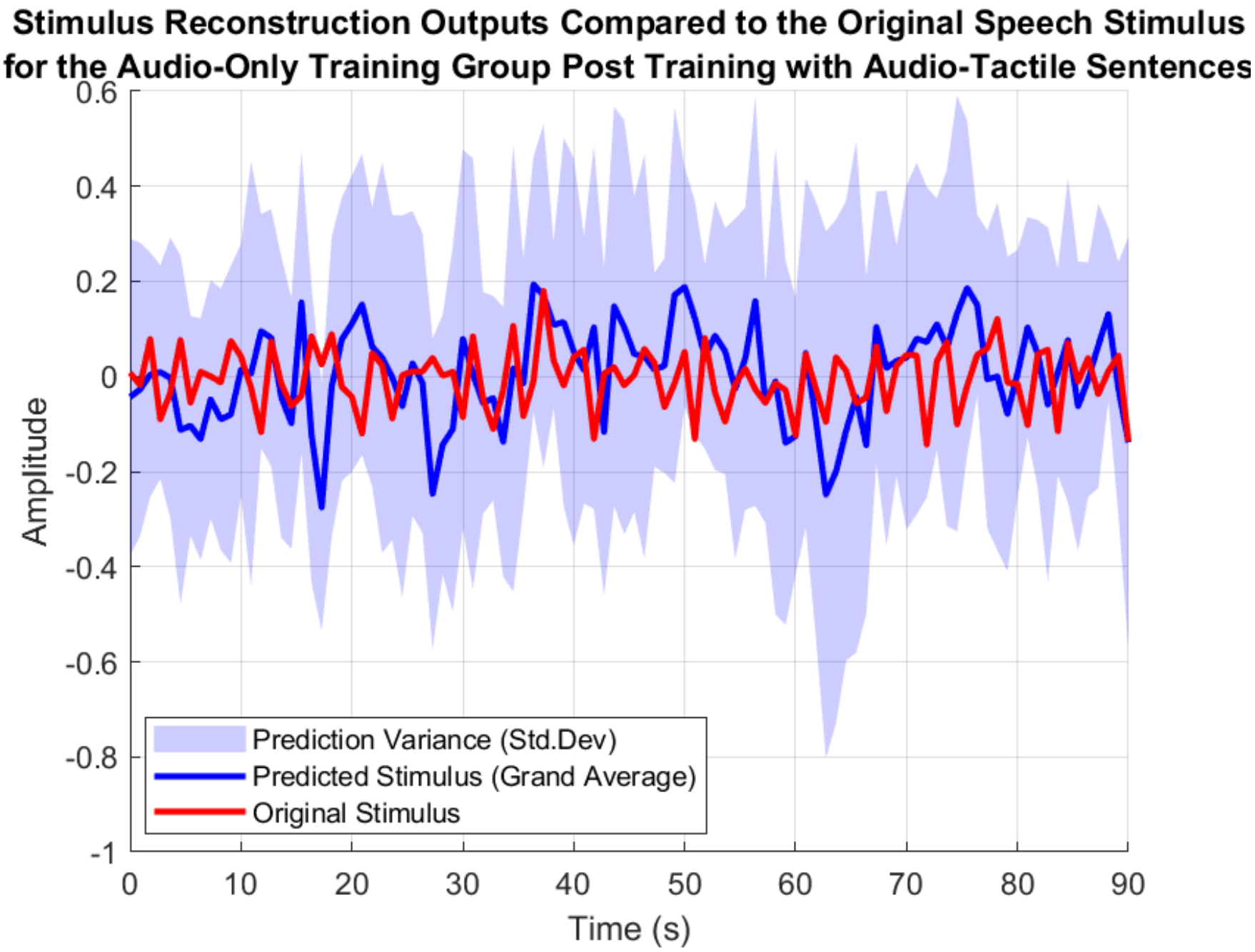
Supplementary Materials B

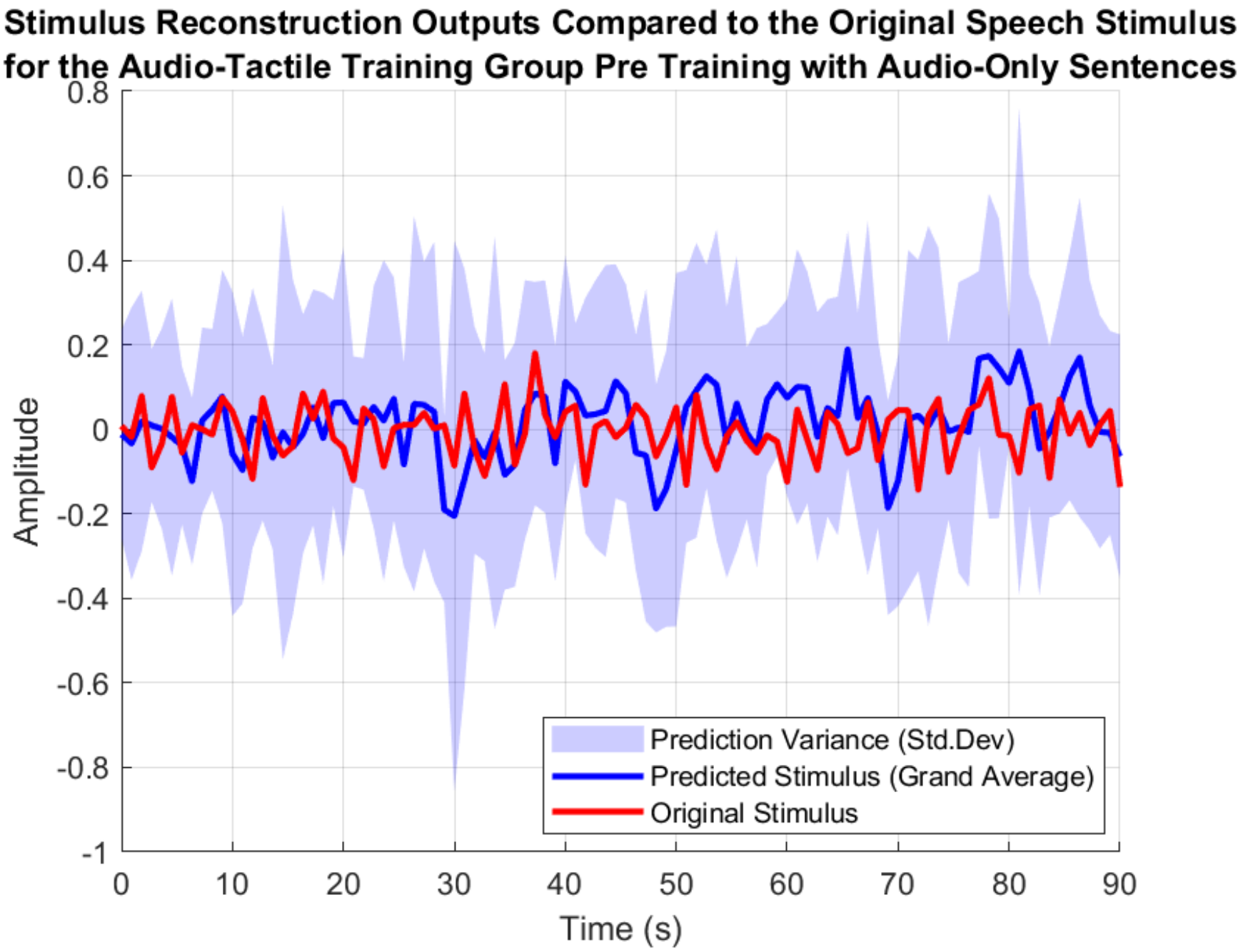
The following figures show the comparisons between the predicted stimuli from stimulus reconstructions and the original speech stimulus, for each training group (audio-only training, and audio-tactile training), session number (pre-training, and post-training), and stimulation type (audio-only, and audio-tactile). In each figure, the solid blue line represents the grand average (mean) predicted stimulus across all participants, with variance displayed as one standard deviation away from the mean, whilst the solid red line represents the original stimulus (speech envelope).

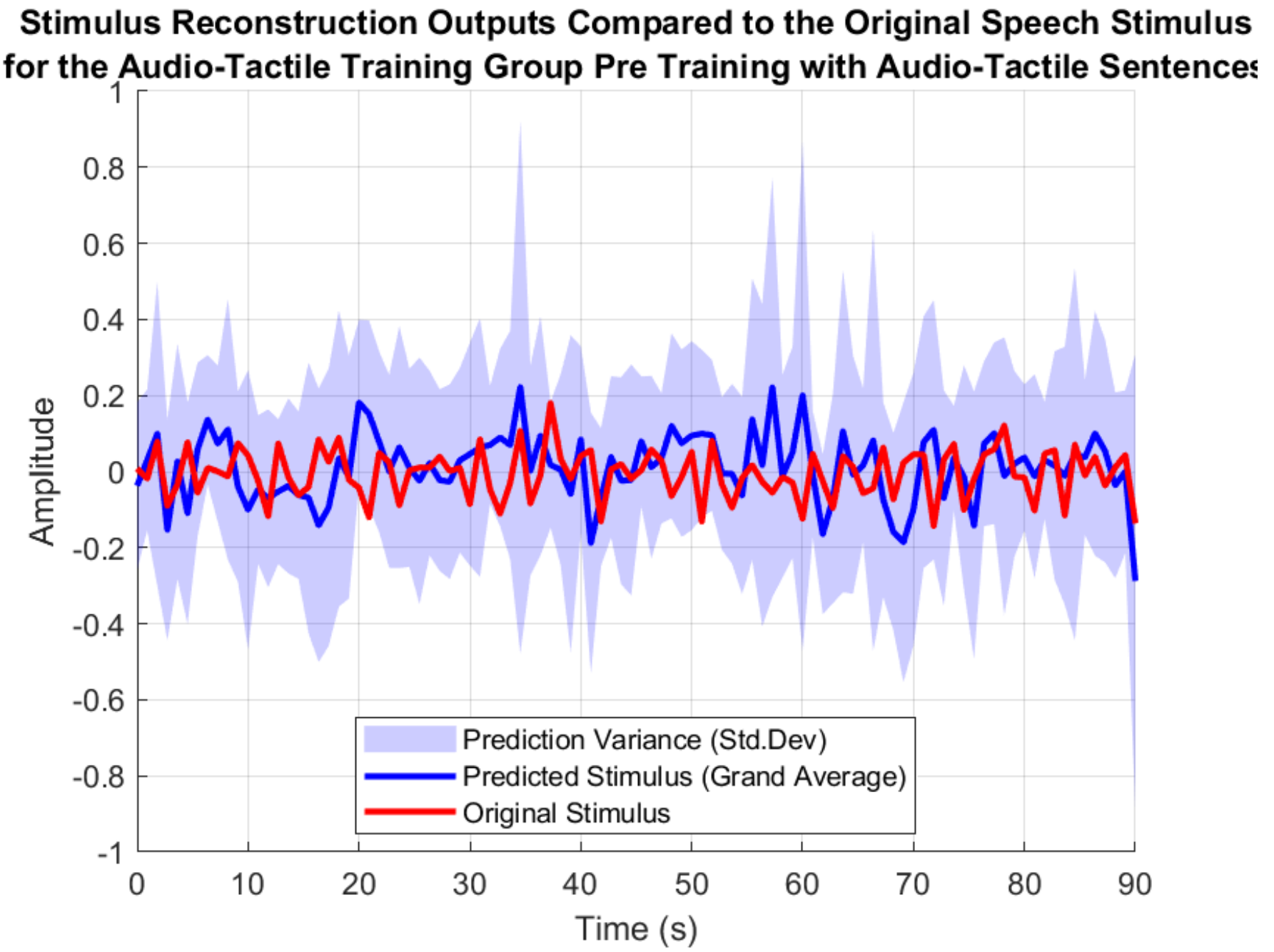

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Audio-Only Training Group Pre Training with Audio-Only Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90
Time (s)

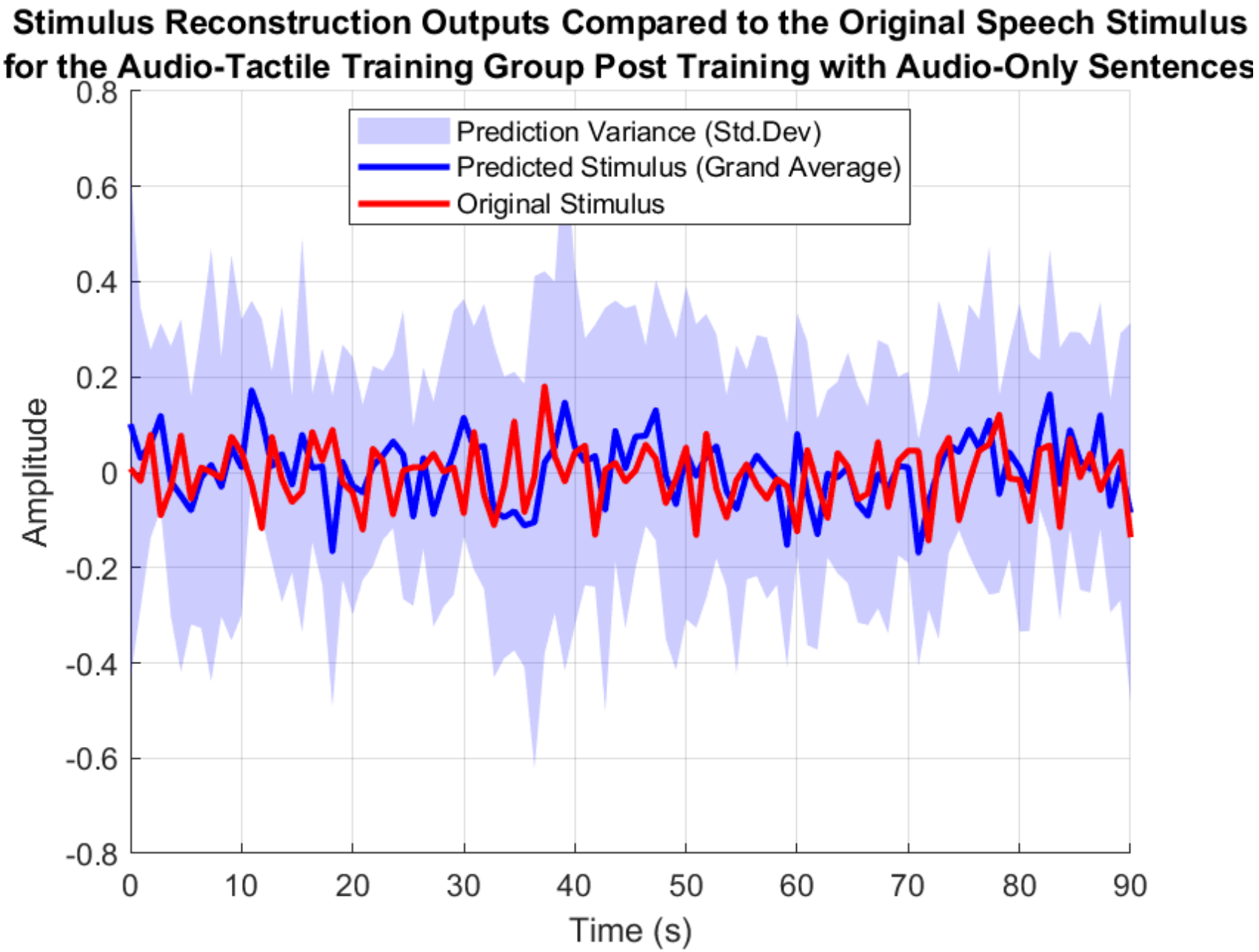

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus for the Audio-Only Training Group Pre Training with Audio-Tactile Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
Time (s)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Audio-Only Training Group Post Training with Audio-Only Sentences
Amplitude
Time (s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus

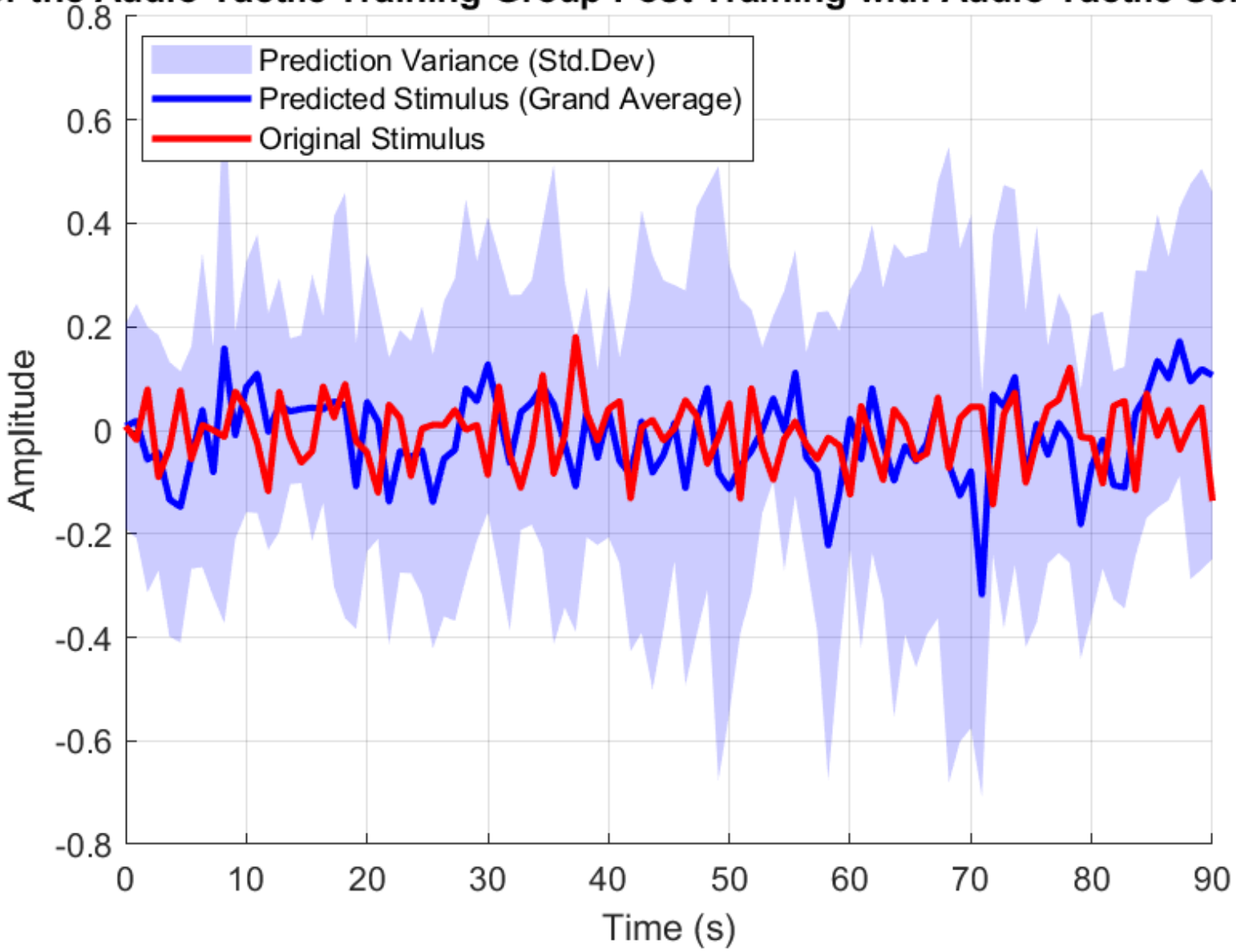

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Audio-Only Training Group Post Training with Audio-Tactile Sentences
Amplitude
Time (s)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus for the Audio-Tactile Training Group Pre Training with Audio-Only Sentences
Amplitude
Time (s)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Audio-Tactile Training Group Pre Training with Audio-Tactile Sentences
Amplitude
Time (s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus


Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus
for the Audio-Tactile Training Group Post Training with Audio-Only Sentences
Prediction Variance (Std.Dev)
Predicted Stimulus (Grand Average)
Original Stimulus
Amplitude
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60
70
80
90
Time (s)

Stimulus Reconstruction Outputs Compared to the Original Speech Stimulus for the Audio-Tactile Training Group Post Training with Audio-Tactile Sentence



References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387*(6631), 401-406.

Amitay, S., Zhang, Y. X., Jones, P. R., & Moore, D. R. (2014). Perceptual learning: Top to bottom. *Vision Research, 99*, 69-77.

Asilador, A., & Llano, D. A. (2021). Top-down inference in the auditory system: potential roles for corticofugal projections. *Frontiers in Neural Circuits, 14*, 615259.

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory, 9*(3), 138-150.

Banai, K., & Lavner, Y. (2019). Effects of stimulus repetition and training schedule on the perceptual learning of time-compressed speech and its transfer. *Attention, Perception, & Psychophysics, 81*(8), 2944-2955.

Bieber, R. E., & Gordon-Salant, S. (2021). Improving older adults' understanding of challenging speech: Auditory training, rapid adaptation and perceptual learning. *Hearing Research, 402*, 108054.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25*(5), 402-412.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.53, retrieved August 2021 from http://www.praat.org/

BRITISH SOCIETY OF AUDIOLOGY (2018). Pure-tone air-conduction and bone conduction threshold audiometry with and without masking [Online]. Available at: https://www.thebsa.org.uk/resources/

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *NeuroImage, 233*, 117958.

Bugannim, Y., Roth, D. A. E., Zechoval, D., & Kishon-Rabin, L. (2019). Training of speech perception in noise in pre-lingual hearing impaired adults with cochlear implants compared with normal hearing adults. *Otology & Neurotology, 40*(3), e316-e325.

Bulus, M. (2023). *Pwrss: Statistical power and sample size calculation tools*. R Package Version 0.3, 1.

Cieśla, K., Wolak, T., Lorens, A., Mentzel, M., Skarżyński, H., & Amedi, A. (2022). Effects of training and using an audio-tactile sensory substitution device on speech-in-noise understanding. *Scientific Reports, 12*(1), 3206.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Vincent, W., De Tiège, X., & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage, 184*, 201-213.

Diekhof, E. K., Biedermann, F., Ruebsamen, R., & Gruber, O. (2009). Top-down and bottom-up modulation of brain structures involved in auditory discrimination. *Brain Research, 1297*, 118-123.

Drouin, J. R., & Theodore, R. M. (2022). Many tasks, same outcome: Role of training task on learning and maintenance of noise-vocoded speech. *The Journal of the Acoustical Society of America, 152*(2), 981-993.

Drouin, J. R., Zysk, V. A., Myers, E. B., & Theodore, R. M. (2023). Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech. *Journal of Speech, Language, and Hearing Research, 66*(2), 720-734.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience, 39*(29), 5750-5759.

Fletcher, M. D., Song, H., & Perry, S. W. (2020). Electro-haptic stimulation enhances speech recognition in spatially separated noise for cochlear implant users. *Scientific Reports, 10*(1), 12723.

Ford, A. N., Czarny, J. E., Rogalla, M. M., Quass, G. L., & Apostolides, P. F. (2024). Auditory corticofugal neurons transmit auditory and non-auditory information during behavior. *Journal of Neuroscience, 44*(7).

Gohari, N., Dastgerdi, Z. H., Rouhbakhsh, N., Afshar, S., & Mobini, R. (2023). Training programs for improving speech perception in noise: A review. *Journal of Audiology & Otology, 27*(1), 1.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. *Journal of Neuroscience*, *33*(4), 1417-1426.

Graetzer, S., Akeroyd, M. A., Barker, J., Cox, T. J., Culling, J. F., Naylor, G., ... & Viveros-Muñoz, R. (2022). Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus. *Data in Brief, 41*, 107951.

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of *Cognitive Neuroscience, 36*(1), 128-142.

Heil, P., & Peterson, A. J. (2015). Basic response properties of auditory nerve fibers: a review. *Cell and Tissue Research*, *361*(1), 129-158.

Huang, N., & Elhilali, M. (2020). Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes. *Elife, 9*, e52984.

Ingvalson, E. M., Dhar, S., Wong, P., & Liu, H. (2015). Working memory training to improve speech perception in noise across languages. *The Journal of the Acoustical Society of America, 137*(6), 3477-3486.

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., & Lizarazu, M. (2024). On the Speech Envelope in the Cortical Tracking of Speech. *NeuroImage*, 120675.

Lad, M., Holmes, E., Chu, A., & Griffiths, T. D. (2020). Speech-in-noise detection is related to auditory working memory precision for frequency. *Scientific Reports, 10*(1), 13997.

Larsby, B., Hällgren, M., Nilsson, L., & McAllister, A. (2015). The influence of female versus male speakers' voice on speech recognition thresholds in noise: Effects of low- and high-frequency hearing impairment. *Speech, Language and Hearing, 18*(2), 83-90.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 245.

Molloy, K., Moore, D. R., Sohoglu, E., & Amitay, S. (2012). Less is more: latent learning is maximized by shorter training sessions in auditory perceptual learning. *PLoS One, 7*(5), e36929.

Montoya-Martínez, J., Vanthornhout, J., Bertrand, A., & Francart, T. (2021). Effect of number and placement of EEG electrodes on measurement of neural tracking of speech. *PLoS One, 16*(2), e0246769. https://doi.org/10.1371/journal.pone.0246769

Moore, D. R., & Amitay, S. (2007, May). Auditory training: rules and applications. In *Seminars in Hearing* (Vol. 28, No. 02, pp. 099-109). Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213.

O'Hanlon, B. L., Riecke, L., Hausfeld, L., Usherwood, B., Plack, C. J., & Nuttall, H. E. (in prep., 2025). *Effects of Short-Term Audio-Tactile Training on Cortical Speech-Envelope Tracking and Speech Intelligibility*. [Unpublished Manuscript]. Department of Psychology, Lancaster University.

Oberle, H. M., Ford, A. N., Dileepkumar, D., Czarny, J., & Apostolides, P. F. (2022). Synaptic mechanisms of top-down control in the non-lemniscal inferior colliculus. *Elife, 10*, e72730.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage, 198*, 181-197.

Pontifex, M. B., Gwizdala, K. L., Parks, A. C., Billinger, M., & Brunner, C. (2017). Variability of ICA decomposition may impact EEG signals when used to remove eyeblink artifacts. *Psychophysiology, 54*(3), 386-398.

Qualtrics. (2005). *Qualtrics software*, Provo, Utah, USA. Copyright@2021, Current version: 09-21. Retrieved from: https://www.qualtrics.com

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*, 116134.

Rizza, A., Terekhov, A. V., Montone, G., Olivetti-Belardinelli, M., & O'Regan, J. K. (2018). Why early tactile speech aids may have failed: No perceptual integration of tactile and auditory signals. *Frontiers in Psychology, 9*, 767.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences, 104(*17), 7295-7300.

Schneiders, J. A., Opitz, B., Tang, H., Deng, Y., Xie, C., Li, H., & Mecklinger, A. (2012). The impact of auditory working memory training on the fronto-parietal working memory network. *Frontiers in Human Neuroscience, 6*, 173.

Schumann, A., Serman, M., Gefeller, O., & Hoppe, U. (2015). Computer-based auditory phoneme discrimination training improves speech recognition in noise in experienced adult cochlear implant listeners. *International Journal of Audiology, 54*(3), 190-198.

Souffi, S., Nodal, F. R., Bajo, V. M., & Edeline, J. M. (2021). When and how does the auditory cortex influence subcortical auditory structures? New insights about the roles of descending cortical projections. *Frontiers in Neuroscience, 15*, 690223.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400-410.

Vermeire, K., Knoop, A., De Sloovere, M., Bosch, P., & van den Noort, M. (2019). Relationship between working memory and speech-in-noise recognition in young and older adult listeners with age-appropriate hearing. *Journal of Speech, Language, and Hearing Research, 62*(9), 3545-3553.

Wisniewski, M. G., Church, B. A., Mercado, E., Radell, M. L., & Zakrzewski, A. C. (2019). Easy-to-hard effects in perceptual learning depend upon the degree to which initial trials are “easy”. *Psychonomic Bulletin & Review, 26*, 1889-1895.

Wisniewski, M. G., Mercado III, E., & Church, B. A. (2024). You Can Make it Too Easy: Simulating Easy-To-Hard Effects in Auditory Perceptual Learning with an Unsupervised Neural Network Model of Plasticity. *Auditory Perception & Cognition, 7*(3), 219-233.

Woodruff, E., Poltronieri, B. C., de Albuquerque Sousa, L. P., de Oliveira, Y. G., Reis, M. A., Scoriels, L., & Panizzutti, R. (2024). Effects of bottom-up versus top-down digital cognitive training in older adults: A randomized controlled trial. *Archives of Gerontology and Geriatrics, 127*, 105552.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage, 32*(4), 1826-1836.

Zerr, M., Freihorst, C., Schütz, H., Sinke, C., Müller, A., Bleich, S., ... & Szycik, G. R. (2019). Brief sensory training narrows the temporal binding window and enhances long-term multimodal speech perception. *Frontiers in Psychology, 10*, 2489.

Chapter 5

5 Does Speech Tracking Play a Role in Predicting Oncoming Speech? Evidence from Audio-visual Speech Perception

Linking Statement:

With Chapters 3 and 4, we investigated how both short-term bottom-up and top-down training with audio-tactile speech may benefit speech perception through enhanced neural speech-envelope tracking accuracy and intelligibility. Despite seeing evidence of enhanced tracking accuracy with audio-tactile speech, no benefits to intelligibility from audio-tactile training were seen, in contrast to audio-visual intelligibility benefits assessed in Chapter 2. To understand this further, this chapter returns to investigate audio-visual integration from a neural perspective. Through secondary data analysis of an intracranial audio-visual speech study, we attempt to understand the potential role of speech tracking in the prediction of oncoming speech.

Author Note: *This work was produced in collaboration Dr. Helen Nuttall, Prof. Christopher Plack, and Dr. Sana Hannan. This paper is currently under review from the wider collaborative team before planned submission to the Journal of Neuroscience for publication.*

Statement of Authorship

Chapter 5 (Paper Four):	Does Speech Tracking Play a Role in Predicting Oncoming Speech? Evidence from Audio-visual Speech Perception
Authors:	Brandon O'Hanlon, Sana Hannan, Christopher J Plack, and Helen E Nuttall
Publication status:	Published Publication has been accepted Publication has been submitted **Unpublished/Unsubmitted but in manuscript style**
Reference:	Not published.
Student/Principle Author:	Brandon Lee O'Hanlon
Contribution:	Theoretical conceptualization; statistical secondary data analysis; manuscript development; manuscript revisions based on supervisor feedback.

Principle Author Signature: Date: 14.04.2025

Through signing this statement, the Co-authors agree that:

(a) The student's contribution to the above papers is correct

(b) The student can incorporate this paper within the thesis

(c) The contribution of all co-authors for each paper equals 100% minus the contribution of the student.

Co-author Name:	Sana Hannan
Co-author Contributions:	Collaborator. Contributed theoretical knowledge and advice on electrocorticography methodology and analysis.

Co-author Signature: Date: 14/04/2025

Co-author Name:	Christopher J Plack
Co-author Contributions:	Co-supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 25/04/2025

Co-author Name:	Helen E Nuttall
Co-author Contributions:	Primary supervisor. Contributed to manuscript draft and final manuscript revisions.

Co-author Signature: Date: 28/04/2025

5.1 Abstract

The auditory cortex tracks speech by synchronising neural activations with fluctuations in the speech signal. Speech-tracking accuracy can be improved with speech-relevant visual information, such as through lipreading. Furthermore, enhancements in speech tracking are thought to be linked to enhancements in speech intelligibility. However, this is not always the case. For example, speech-relevant tactile information improves envelope tracking, but not intelligibility. It may be that speech tracking plays a role in the prediction of oncoming speech, much akin to visual benefit from lipreading, which may explain why tracking accuracy is enhanced with audio-visual speech but not necessarily for audio-tactile speech. This study aimed to investigate the potential role of tracking in speech prediction through secondary analysis of electroencephalography data (Karas *et al*., 2019). Intracranial electroencephalographic data were collected by Karas et al. from eight participants awaiting epilepsy treatment whilst participants perceived auditory or audio-visual speech. In some trials, the onset of the auditory stimulus preceded the onset of the visual information by either 60 or 100 ms (voice-leading condition). In other trials, the onset of visual information preceded the onset of the auditory stimulus by either 40 or 100 ms (mouth-leading condition). Karas *et al.* demonstrated that lipreading primed the posterior superior temporal gyrus (pSTG) to predict speech when stimuli were mouth-leading compared more accurately to voice-leading, indicated through suppressed broadband high-gamma activity for representations of incompatible phonemes. We hypothesised that if neural tracking in the pSTG also played a role in speech prediction, then visual benefit to tracking accuracy would be greater in the mouth-leading condition compared to the voice-leading condition. A generalised linear mixed-effects model found no significant difference in the visual benefit of speech tracking between mouth-leading and voice-leading conditions ($\beta = .002$, $t = .04$, *95% CI* = [-.07, .07], $p > .05$). This secondary data analysis suggests that speech tracking

enhancement may not be indicative of enhanced prediction of speech and therefore may have a different role in speech perception, such as assisting with continuous selective attention to a single speaker in multi-speaker environments.

5.2 Does Speech Tracking Play a Role in Predicting Oncoming Speech? Evidence from Audio-visual Speech Perception

Speech tracking refers to the auditory cortex's ability to phase lock neuronal activity to stimulus features present in speech, such as the speech envelope (Issa *et al.*, 2024). The purpose of this tracking in speech processing is unclear (Karunathilake *et al.*, 2023). In previous literature, there has been an assumed link between tracking accuracy and speech intelligibility (Vanthornhout *et al.*, 2018) with some evidence for tracking accuracy to be sufficient as a neuronal measure of intelligibility (Lesenfants *et al.*, 2019). Whilst some evidence has suggested that greater neural speech tracking accuracy accompanies greater speech intelligibility (Kong *et al*., 2015), other studies have found no intelligibility benefits associated with tracking accuracy enhancement (Kösem *et al.*, 2023), particularly in the audio-tactile domain (Riecke, *et al.*, 2019; O'Hanlon, *et al.*, in prep., 2025). This lack of understanding about the role of speech tracking and its relationship with speech intelligibility makes it difficult to assume that tracking can be used as an objective, neural measure of speech intelligibility or comprehension independently of behavioural measures. This uncertainty in purpose does complicate applications of speech tracking too, as with recent developments in dynamic neuro-steered hearing aids to provide improvements to hearing technologies through auditory attention decoding using speech tracking (see: Geirnaert *et al.*, 2021; Geirnaert *et al*., 2024; Straetmans, 2022). In this case, understanding the role of tracking in speech processing and perception may assist in the optimisation of these attentional decoding applications for dynamic hearing aids to further improve their application.

On the other hand, audio-visual speech studies indicate a link between tracking accuracy and speech intelligibility (Haider *et al.,* 2024; Golumbic *et al.*, 2013), in stark contrast to tracking evidence in the audio-tactile domain (Riecke, *et al.*, 2019; O'Hanlon, *et*

al., in prep., 2025). Audio-visual integration makes use of information from visual lipreading, in that viewing the lip movements of a speaker whilst listening to speech can improve speech intelligibility in noise (Sumby & Pollack, 1954; Maier *et al*., 2011). This integrative process is likely facilitated by established multisensory integration hubs in the medial geniculate body and lateral geniculate nucleus (Luo *et al.,* 2022; Meng & Schneider, 2022), as well as through corticollicular descending projections to the inferior colliculus from the auditory cortex (Hu & Dan, 2022). Lip movements related to specific speech phonemes are categorised into groups known as viseme categories (Massaro *et al*., 2012). An example of phonemes that share a viseme category are /b/ (‘ba’), /p/ (‘pa), and /m/ (‘ma’). Of note, no phoneme in the English language occupies a viseme category alone (Fisher, 1968). This means that it is impossible to distinguish between speech phonemes that share the same viseme category through lipreading alone (Van Engen *et al.*, 2022). Work by Karas *et al.* (2019) showed that when visual information preceded the auditory information of the speech, representations of incorrect phonemes were suppressed in the posterior superior temporal gyrus (pSTG). These suppressed phonemes were ones outside of the viseme categories determined from visual lipreading. An important distinction in Karas *et al.* is that this suppression of incongruent phonemes of alternate viseme categories only occurred when the speech was ‘mouth-leading’ (the onset of the visual information preceded the onset of the auditory information) and not when speech was presented as ‘voice-leading’ (auditory onset preceding the visual onset). This mouth-leading (or visual head-start) phenomenon has been evidenced extensively by introducing further stimulus onset asynchrony (SOA) to audio-visual speech (Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al*., 2017). When too much SOA is introduced, and the visual information precedes the auditory by too much time, visual benefit to speech intelligibility is lost (Chandrasekaran et al., 2009), despite the stimuli still presenting as mouth-leading. This window of integration varies in the literature but can be

estimated to be around 240 ms to 300 ms for phonemes and short word presentations (Navarra et al., 2005). In the case of Karas *et al.* (2019), mouth-leading stimuli were presented with SOAs of either 60 or 100 ms, landing within the window of integration for audio-visual word presentations.

Returning to neural speech tracking accuracy, audio-visual speech studies have shown links between tracking accuracy and intelligibility (Haider *et al.,* 2024; Golumbic *et al.*, 2013), compared to mixed findings in auditory and audio-tactile speech (Kösem *et al.,* 2023; Riecke, et al., 2019; O'Hanlon *et al*., *in prep., 2025*). This may be indicative of tracking playing a similar role in speech prediction as audio-visual speech integration by priming the auditory cortex to oncoming speech and supressing irrelevant phoneme representation (Zoefel & VanRullen, 2016). When this prediction is correct, processing of speech would likely be more accurate, potentially increasing speech intelligibility. However, neural speech tracking accuracy has not been investigated with regards to mouth-leading and voice-leading audio-visual stimuli. Using existing intracranial data from Karas *et al.* (2019), we conducted a secondary data analysis investigating potential differences in neural speech tracking accuracy between mouth-leading and voice-leading stimuli, providing insight into tracking's potential role in speech prediction akin to audio-visual integration.

The following primary hypothesis was tested:

(i): Presenting audio-visual speech with mouth-leading onsets (the visual information preceding the auditory information) will increase visual benefit to speech tracking accuracy more than when audio-visual information is presented with voice-leading onsets (the auditory information preceding the visual information).

5.3 Materials and Methods

5.3.1 Participants

The Karas *et al.* (2019) dataset included eight participants (five female, M_{age} = 36 years) who were awaiting neurosurgery for refractory epilepsy. Participants were implanted with intracranial electrodes to record ECoG data. From these eight participants, data from all 128-channels were collected during the task. The data were then mapped onto a single brain atlas, and electrodes from this atlas were selected that met anatomical (being placed over the posterior superior temporal gyrus) and functional (had a significant broadband high frequency activity response to auditory-only speech) criteria. This atlas consisted of 28 electrode channels averaged across all eight participants.

5.3.2 Sample Size Calculations

Post hoc power and sample size calculations were computed using *pwrss* (Bulus, 2023: https://pwrss.shinyapps.io/index/), a shinyapp and r package used to conduct sample size calculations, for a repeated measures ANOVA. Intracranial neural measures are of high quality compared to scalp measures such as electroencephalography due to higher spatial and temporal resolutions (Todaro *et al.*, 2019), which is useful for speech tracking decoding. Moreover, signal quality is significantly improved in ECoG recordings, with reduced artifacts (Ball *et al.*, 2009; see also Kanth & Ray, 2020). This makes neural decoding via stimulus reconstruction more valid, as the decoder is trained on data that more accurately represent activity in the auditory pathway (Crosse *et al.*, 2016). Thus, we determined sample sizes for both assumed medium and assumed large effect sizes to achieve a power of .8. For a large effect size of $\eta 2 = .26$, a total sample size of eight was needed. Therefore, we assumed this study's sample of eight to be sufficiently powered for our planned analysis with a large effect size. Again, this justification was made with the expectation of intracranial measures to

provide clean neural data for tracking analyses, as it records data at a significantly higher signal-to-noise ratio than non-invasive methods (Parvizi, & Kastner, 2018).

5.3.3 Experimental Design

To address our hypothesis, secondary data analysis was conducted on an open access Electrocorticography (ECoG) dataset collected by Karas *et al.* (2019). Of this dataset, only the neural intracranial data was analysed and not the behavioural experimental data. The ECoG data were collected during a word listening task, in which participants listened to single English spoken words. These were presented in video format, with some videos showing a blank screen during listening (auditory) and other videos showing the speaker's lips moving as they spoke (audio-visual). With audio-visual trials, some stimuli were presented as voice-leading (the onset of the auditory information played before the onset of the visual lip movements in the video played) and some as mouth-leading (vice-versa). With auditory only trials, where no visual information was present, the stimuli were presented auditory onsets as their audio-visual counterparts. For example, the mouth-leading audio-visual presentation of 'drive' had a visual onset of 170 ms and an auditory onset of 230 ms. For the auditory only counterpart, 'drive' was presented with a matching auditory onset of 230 ms and with no visual information present in the video (blank screen).

A single (leading stimulus; voice-leading, or mouth-leading) within factor design was used to evaluate this data and test our hypothesis. For the dependent variable, visual benefit to tracking accuracy (VbRz) was used. Ethical approval was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (approval reference: FST-2023-4163-SR-1, project ID: 4163). The analysis code can be found on the Open Science Framework (OSF) at: https://osf.io/autv6/.

5.3.4 Electrocorticography

Neural activity was recorded intracranially during listening in the original study by Karas et al. (2019) using a 128-channel Cerebus Data Acquisition electrocorticography system (Blackrock Microsystems, Salt Lake City, UT). Electrodes were placed on both the left and right posterior superior temporal gyri (pSTG), with an inversed electrode facing the skull as a reference. These were placed away from detected epileptogenic zones.

5.3.5 Stimuli

There were two mouth-leading exemplars (visual information preceding the auditory information by either 40 or 100 ms) and two voice-leading (auditory preceding visual by either 60 or 100 ms) exemplars, giving rise to four different stimuli presented during neural recordings. These were the words: ‘drive’, ‘known’, ‘meant’, and ‘last’. Of these, voice-leading trials used ‘meant’ (40 ms offset) and ‘known’ (100 ms offset) whereas mouth-leading trials used ‘drive’ (60 ms offset) and ‘last’ (100 ms offset). These stimuli were presented at 30 frames per second (30 Hz monitor refresh rate) with an auditory bandwidth of 22.05 kHz (44.1 kHz sampling).

5.3.6 Procedure

In the original procedure, participants were presented with auditory only, visual only, and audio-visual stimuli – all without noise. Subjects only responded to catch trials, which were always an audio-visual stimulus of the word ‘press’. No behavioural measures were taken during neural recordings. For this secondary data analysis, only the auditory only and audio-visual trials were analysed for neural tracking accuracy.

5.3.7 Calculation of Broadband High-frequency Activity (BHA)

The open dataset provided by Karas *et al*. presents broadband high-frequency activity (BHA) values for each participant and their selected atlas electrodes during word listening,

reflecting high-gamma neuronal activations in the pSTG. BHA was obtained via conversion of raw neural data into frequency and phase domains using wavelet transformation, power transforming into signal change percentage, and averaging over frequencies between 75 and 150 Hz (see Karas *et al.*, 2019).

5.3.8 Speech tracking accuracy (Rz)

Speech tracking accuracy (Rz) was obtained using the stimulus reconstruction method via the multivariate Temporal Response Function toolbox in MatLab (Crosse *et al.*, 2016). This method of reconstruction uses a backwards approach with a decoder for the neural data. For cross-validation, the method of 'leave-one-trial-out' was chosen (see Riecke *et al.,* 2019). The reconstructed speech was then correlated with the original speech signal using Pearson's R. This correlational value was used as Rz. As the intracranial dataset involved high-gamma BHA values rather than raw neural output with low delta and theta information, the full stimulus signal (downsampled to match the neural sampling rate) was used for the correlation between original and reconstructed speech signal. Whilst typically a stimulus feature such as the speech-envelope is used for backwards reconstruction models (see Crosse *et al.*, 2016), this intracranial BHA data does provide clear representations of the listened speech directly from the pSTG as evidenced by the original authors. Therefore, the full stimulus signal was deemed adequate for providing Rz measurements. Reconstruction outputs were averaged across all leave-one-trial-out validations to provide a final Rz value for each condition.

5.3.9 Visual benefit to speech tracking accuracy (VbRz)

The Rz measures from each leading stimulus condition for auditory only and audio-visual data were used to calculate visual benefit to speech tracking accuracy (VbRz). Here, the auditory only condition was used as a baseline as we wanted to evaluate the addition of visual information comparatively to auditory only speech. There was a separate auditory only

baseline for both mouth-leading and voice-leading audio-visual conditions. These baselines, whilst presenting no visual information, matched the auditory onsets for each stimulus to preserve presentation timings (see Experimental Design for onset examples). The formula for VbRz was adapted from previous literature (Yuan *et al.*, 2021) to account for the tracking boundaries of -1 to 1 instead of percentage boundaries of 0 to 100. VbRz for the audio-visual mouth-leading (AVML) condition used the auditory onset-matched baseline (AOML) condition as its baseline in the following equation:

$$VbRz = (AVML – AOML)/(1-AOML)$$

VbRz for the audio-visual voice-leading (AVVL) condition used the auditory onset-matched baseline (AOVL) condition as its baseline in the following equation:

$$VbRz = (AVAL – AOVL)/(1-AOVL)$$

5.3.10 Statistical analyses

A generalised linear mixed effect regression model (GLMER) will be used to test the hypothesis. Stimulus was used as the fixed effect, and participant IDs and the word stimulus were added as random factors:

$$VbRz \sim Leading\_Stimulus + (1|ID) + (1|stimuli))$$

Based on our hypothesis, we expected to find a significant increase in visual benefit to speech tracking accuracy from baseline when audio-visual information was presented in the mouth-leading condition compared to when audio-visual information was presented in the voice-leading condition.

5.4 Results

Table 1 shows the mean tracking accuracies for mouth-leading and voice-leading conditions when speech was audio-visual and auditory only, whilst the visual benefit to tracking accuracy for mouth-leading and voice-leading conditions is displayed in Figure 1. The planned GLMER analysis for testing hypothesis (i) was conducted. There was no significant difference in visual benefit to speech tracking accuracy between mouth-leading audio-visual speech and auditory only speech (β = .002, t = .04, *95% CI* = [-.07, .07], $p >$ .05). This finding does not support our hypothesis, suggesting that speech tracking accuracy is not affected by the removal of visual head-starts and may not play a role in oncoming speech prediction.

5.5 Discussion

The role of neural speech tracking accuracy remains unclear in the literature, with mixed evidence of tracking accuracy being intrinsically linked with speech intelligibility (Riecke *et al.,* 2019; Kösem *et al.,* 2023; O'Hanlon *et al.*, *in prep., 2025*). Audio-visual integration remains consistent in its enhancement of both tracking and intelligibility and has been shown to play a role in incoming speech prediction by narrowing down representations of phonemes in the pSTG through speech-relevant visual head-starts (Karas *et al*., 2019). This may provide insight into the role of tracking as contributing to the auditory cortex's ability to predict oncoming speech. To test this, a secondary data analysis of intracranial data collected by Karas *et al.* (2019) was conducted for both mouth-leading and voice-leading stimuli. Results indicated no significant difference in visual benefit to tracking accuracy between mouth-leading and voice-leading conditions. Therefore, no evidence was found for neural tracking enhancements relating to enhanced ability to predict oncoming speech.

As such, it is important to look towards alternative purposes of speech tracking. If the function does not aid in prediction of oncoming speech, it may play a role in attentional decoding when in a listening environment with multiple sources of auditory information (see: Geirnaert *et al.*, 2021; Geirnaert *et al*., 2024; Straetmans, 2022). This may also explain why visual benefit to tracking accuracy was rather low in this analysis compared to other audio-visual tracking research, as there was no difficult listening condition that may have required a further boost to attentional processes to decode the speech. Alternatively, tracking accuracy may be more complex than a singular measure derived from neural activity. There is evidence to suggest that delta and theta frequency bands track speech differentially (Bröhl, & Kayser, 2020), with delta tracking providing benefit to speech clarity and theta to speech comprehension (Etard & Reichenbach, 2019). It may be that tracking across all frequency bands plays a role in a multitude of different speech processing functions. This would help to

explain the inconsistency between tracking and intelligibility measurements across various studies, as tracking accuracies may be reflecting enhancements in different speech processing mechanisms relevant to the listening environment participants experience.

5.5.1 Study limitations

This secondary data analysis was conducted on a relatively small sample of eight patients undergoing neurosurgery. Whilst this sample size was adequate for our post hoc power calculation (assuming a large effect size), it is important to consider the varied amount of intracranial data that was available per each participant for stimulus reconstruction. The open dataset contained pSTG-only recordings from a total of 28 electrodes across eight participants. These electrodes were selected based on anatomical and functional criteria between participants. However, the number of electrodes present for each participant did vary between them. Seven of the participants had three to six electrodes present in the data. One participant, however, only had one electrode present. This does mean that some participants had more intracranial data to train the decoder on for stimulus reconstruction than others, with one participant having considerably less available. It is difficult to ascertain whether one electrode of data was enough to decode tracking accuracy with validity, though the study did present many trials to participants and so even a single electrode may have provided sufficient datapoints. Future research should provide as much neural activity as possible for more valid reconstructions. Furthermore, the original methodology was not designed with speech tracking analyses in mind. The experiment presented single word trials to participants, with no active task to conduct during word listening except for catch trials. These catch trials were not included in the dataset and this analysis. Whilst these single word trials still prompted predictive processing of oncoming speech, as evidenced by the narrowing of phoneme representation in the original study, they may have benefited less from audio-visual integration due to the simplistic nature of the trials presented. Moreover, the lack of difficulty

added to trials by introducing noise to the speech signal would further bring intelligibility towards ceiling. As a result, the visual benefit to tracking accuracy in both conditions may not have seen significant difference due to the ease of listening. In our secondary analysis, we also used broadband high-frequency activity values instead of low-frequency information typically used in speech tracking studies that utilise non-invasive neural methodologies (EEG, MEG). These values were also only derived from the pSTG and not from other speech-relevant nonprimary regions of the auditory cortex, like the planum temporale or the planum polare (Poeppel *et al.*, 2012), which may have been playing a larger role in cortical tracking of speech (de Heer *et al.*, 2017). This does make it more difficult to compare our intracranial analyses with these scalp-based non-invasive studies, as our tracking accuracies reflect phase-locking in the pSTG only that is more relevant to high-gamma signal activity than low-frequency information such as the speech-envelope.

5.5.2 Conclusion

In conclusion, no evidence in this secondary data analysis was provided to support the hypothesis that neural speech tracking accuracy is directly involved in the auditory cortices ability to predict oncoming speech by narrowing down potential phoneme representations, akin to audio-visual integration. This indicates that tracking plays a different role in speech perception, such as through attentional decoding, or that speech tracking assists with speech prediction outside of the pSTG in nonprimary regions of the auditory cortex. Future research should aim to clarify the role of neural speech tracking in speech processing. Understanding this role is crucial for the continued development of future neuro-steered hearing technologies that seek to provide dynamic algorithms using neural tracking accuracy to provide enhancement to speech intelligibility.

Tables and Figures

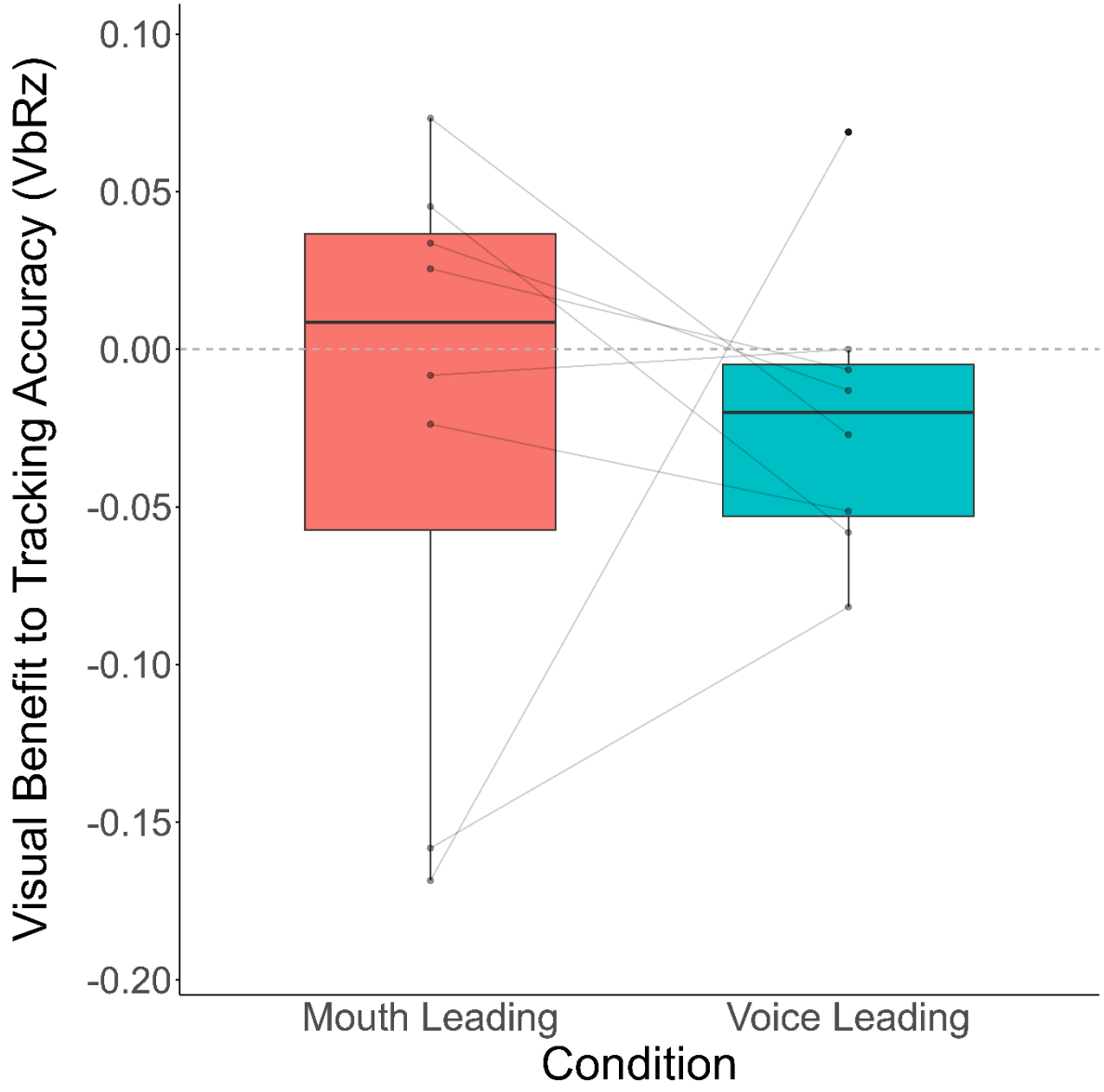
Table 1.

Means and Standard Deviations (SD) of neural speech tracking accuracy (Rz) for auditory only and audio-visual stimuli, when presented with mouth-leading sensory onsets and voice-leading sensory onsets.

	Mouth-leading		Voice-leading	
	Mean	SD	Mean	SD
Auditory Only	.28	.08	.34	.09
Audio-visual	.27	.04	.33	.11

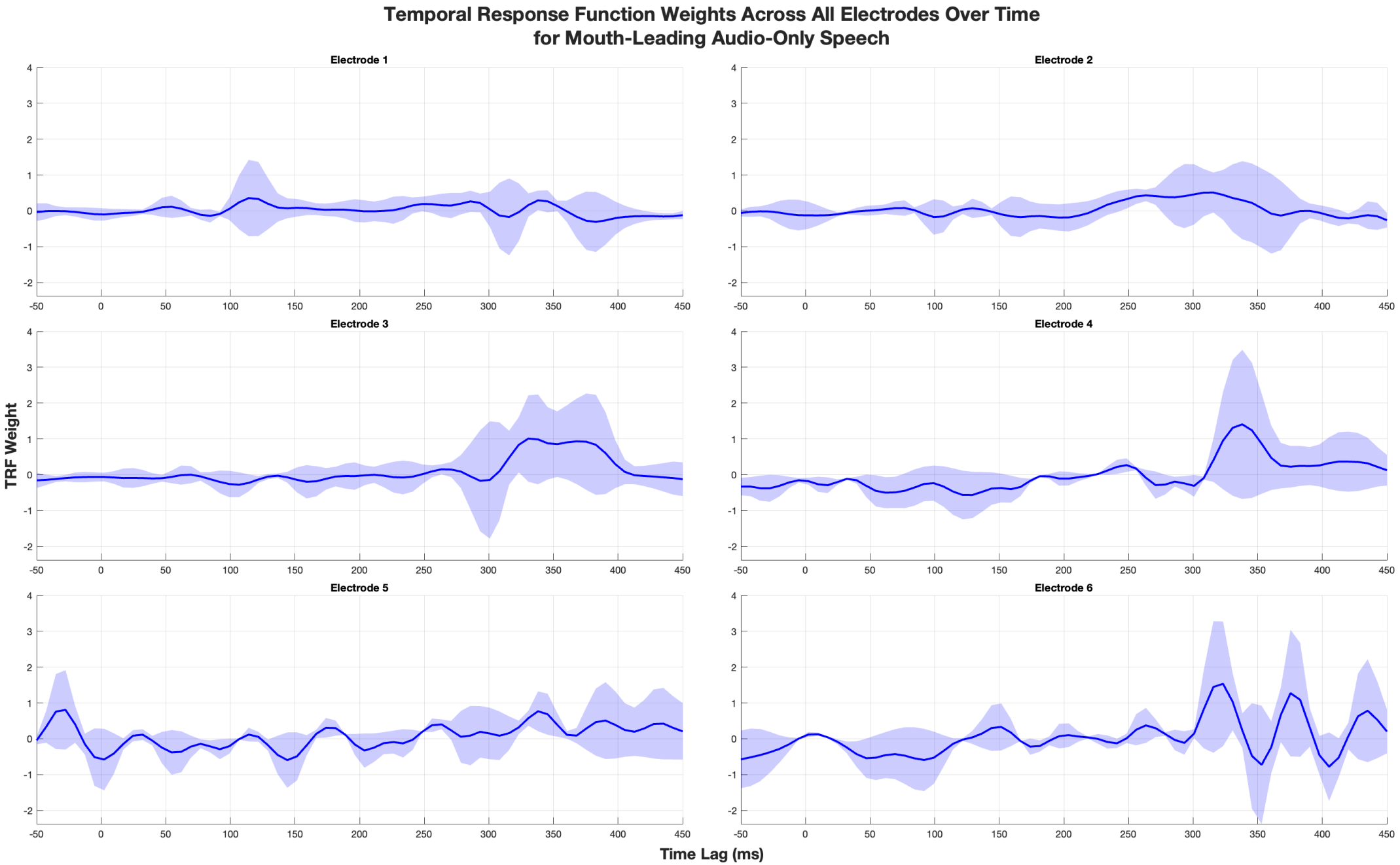
Figure 1.

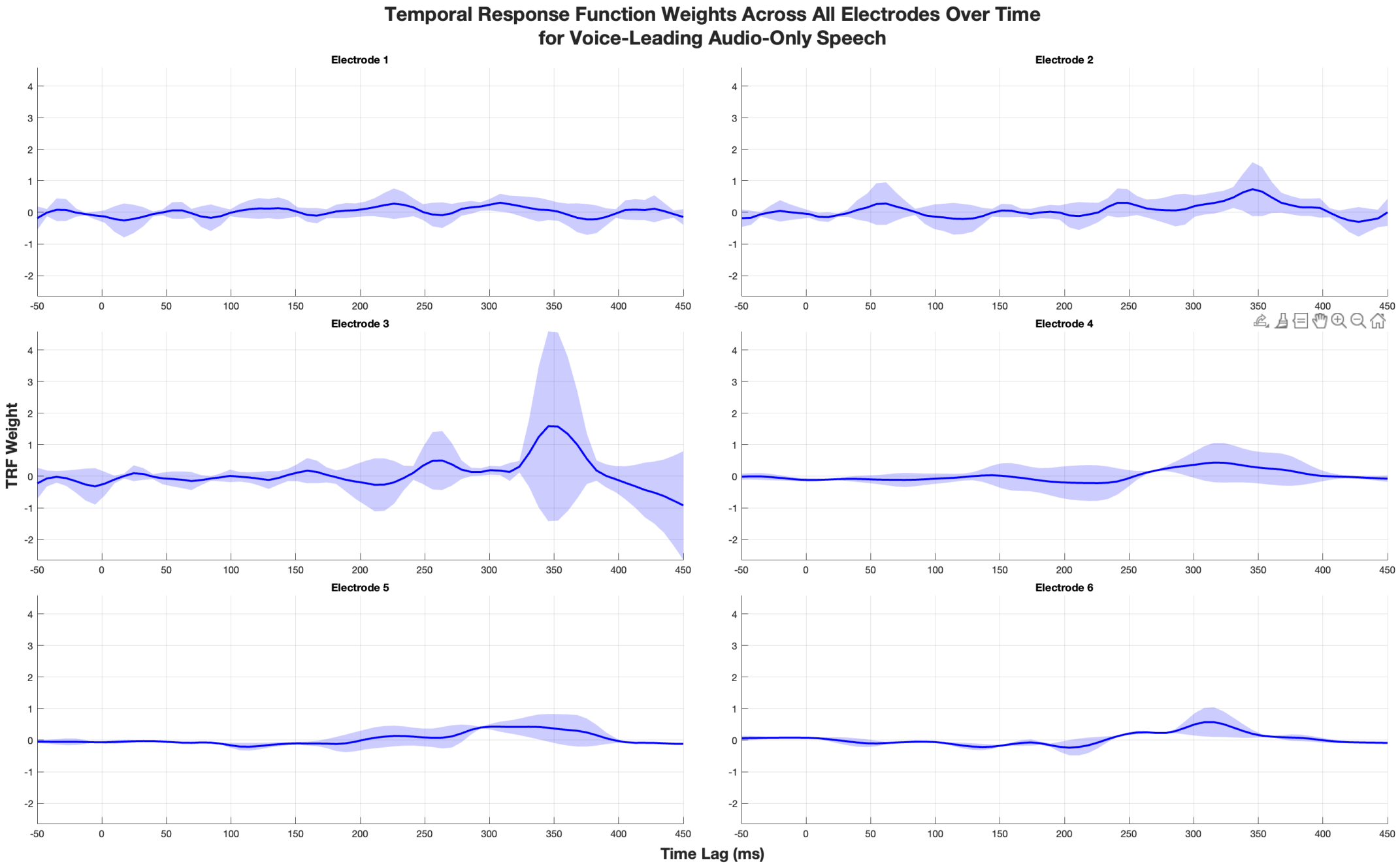
Boxplots showing the median and interquartile ranges for visual benefit to tracking accuracy (VbRz) for both mouth-leading and voice-leading stimuli.

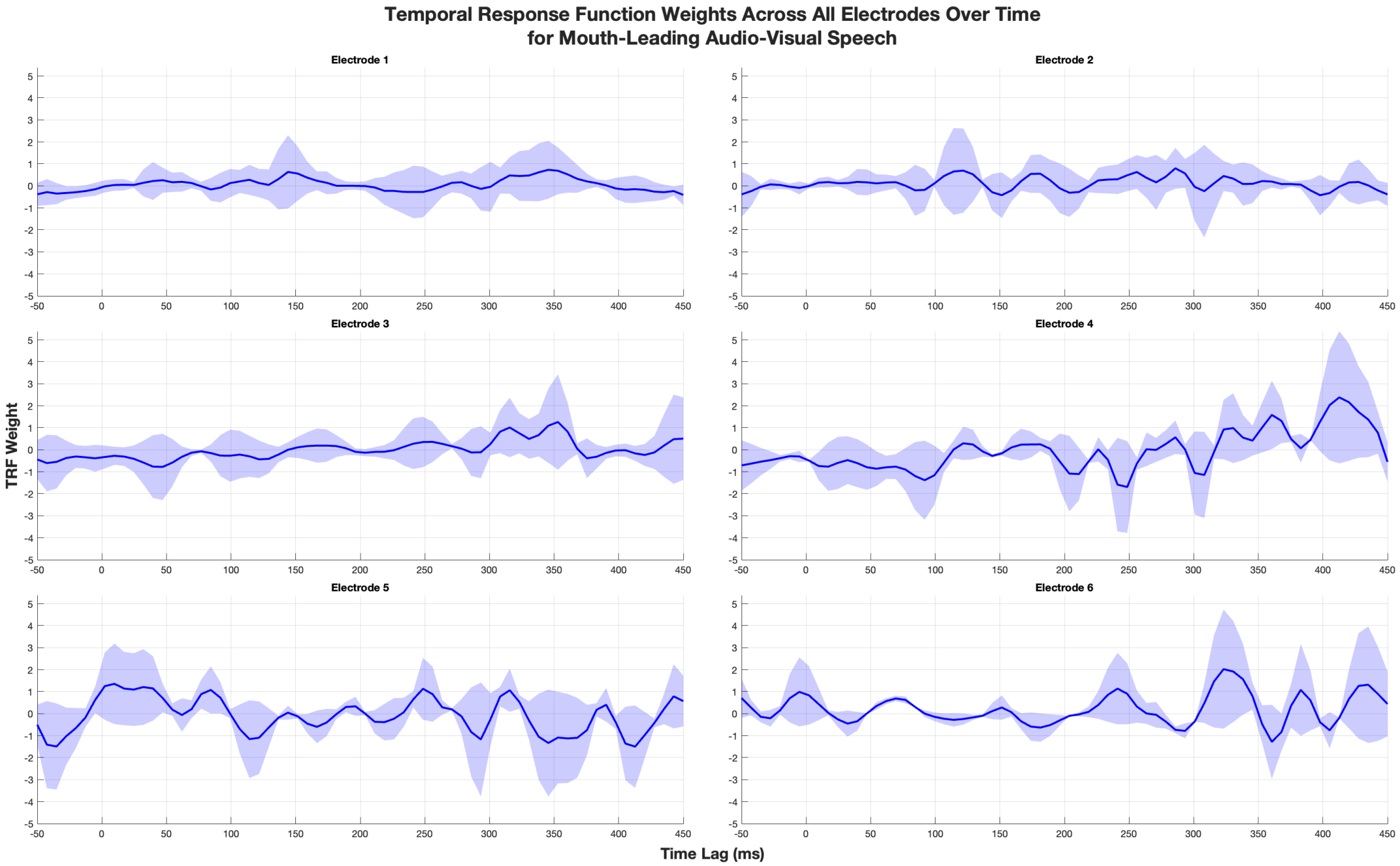


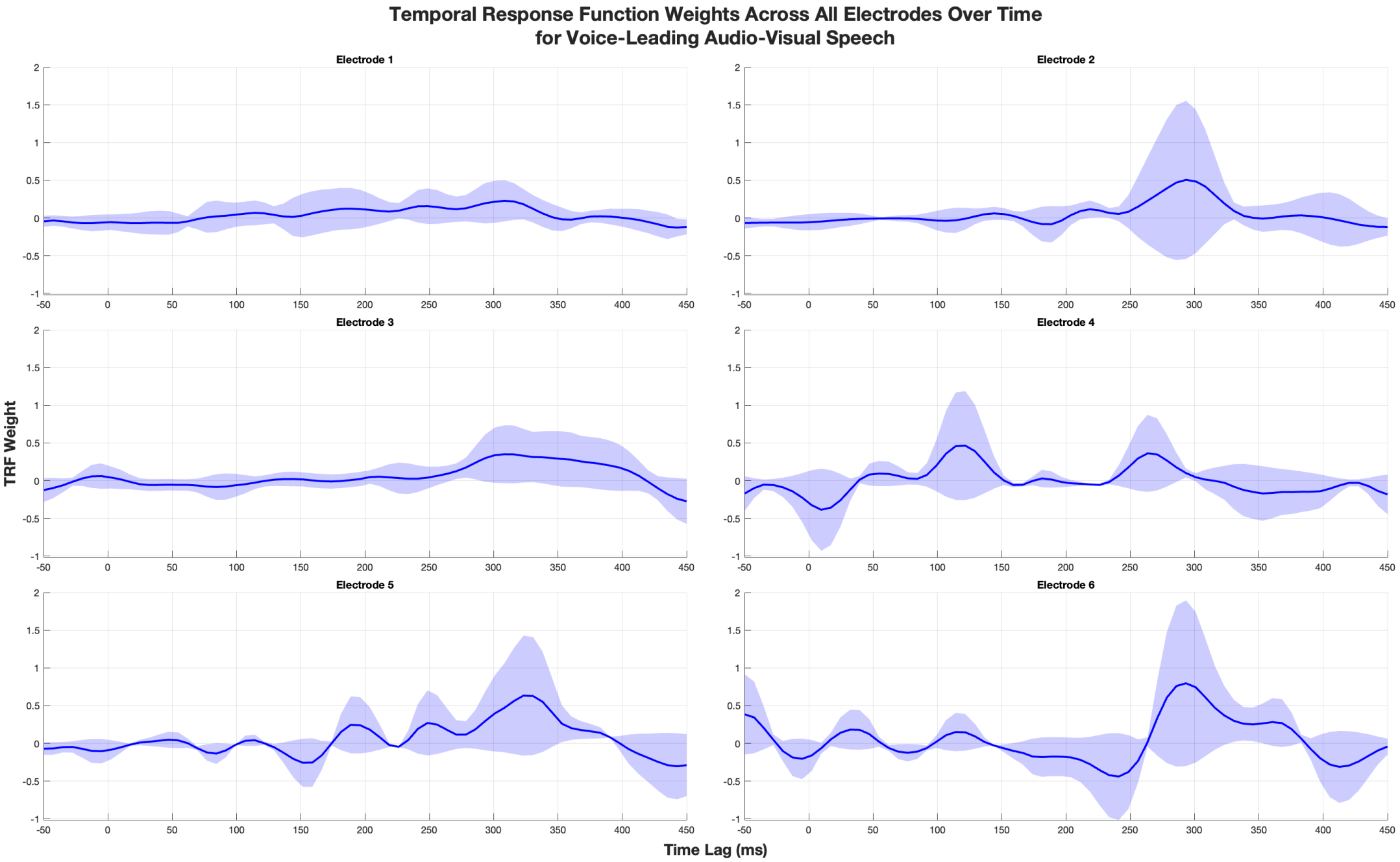
Supplementary Materials A

The following figures show the temporal response function (TRF) weights across all recorded iEEG electrodes for each leading condition (mouth-leading, and voice-leading), and stimulation type (audio-only, and audio-visual). In each figure, the solid blue line represents the grand average (mean) of TRF weightings across all participants, with variance displayed as one standard deviation away from the mean.


Temporal Response Function Weights Across All Electrodes Over Time
for Mouth-Leading Audio-Only Speech
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for Voice-Leading Audio-Only Speech
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for Mouth-Leading Audio-Visual Speech
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
TRF Weight
Time Lag (ms)


Temporal Response Function Weights Across All Electrodes Over Time
for Voice-Leading Audio-Visual Speech
Electrode 1
Electrode 2
Electrode 3
Electrode 4
Electrode 5
Electrode 6
TRF Weight
Time Lag (ms)

References

Ball, T., Kern, M., Mutschler, I., Aertsen, A., & Schulze-Bonhage, A. (2009). Signal quality of simultaneously recorded invasive and non-invasive EEG. *NeuroImage, 46*(3), 708-716.

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *Neuroimage, 233*, 117958.

Bulus, M. (2023). *Pwrss: Statistical power and sample size calculation tools*. R Package Version 0.3, 1.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in Integrative Neuroscience, 4*, 1316.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience, 37*(27), 6539-6557.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience, 39*(29), 5750-5759.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*(4), 796-804.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B. T., ... & Bertrand, A. (2021). EEG-based auditory attention decoding: Towards neuro-steered hearing devices. *arXiv preprint*. Accessed at: https://arxiv.org/abs/2008.04569

Geirnaert, S., Zink, R., Francart, T., & Bertrand, A. (2024). Fast, accurate, unsupervised, and time-adaptive EEG-based auditory attention decoding for neuro-steered hearing devices. *In Brain-Computer Interface Research: A State-of-the-Art Summary 11* (pp. 29-40). Cham: Springer Nature Switzerland.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, *33*(4), 1417-1426.

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of *Cognitive Neuroscience, 36*(1), 128-142.

Hu, F., & Dan, Y. (2022). An inferior-superior colliculus circuit controls auditory cue-directed visual spatial attention. *Neuron, 110*(1), 109-119.

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., & Lizarazu, M. (2024). On the Speech Envelope in the Cortical Tracking of Speech. *NeuroImage*, 120675.

Kanth, S. T., & Ray, S. (2020). Electrocorticogram (ECoG) is highly informative in primate visual cortex. *Journal of Neuroscience, 40*(12), 2430-2444.

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *Elife, 8*, e48116.

Karunathilake, I. D., Kulasingham, J. P., & Simon, J. Z. (2023). Neural tracking measures of speech intelligibility: Manipulating intelligibility while keeping acoustics unchanged. *Proceedings of the National Academy of Sciences, 120*(49), e2309166120.

Kong, Y. Y., Somarowthu, A., & Ding, N. (2015). Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. *Journal of the Association for Research in Otolaryngology, 16*, 783-796.

Kösem, A., Dai, B., McQueen, J. M., & Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage, 272*, 120040.

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., & Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations. *Hearing Research, 380*, 1-9.

Luo, B., Li, J., Liu, J., Li, F., Gu, M., Xiao, H., ... & Xiao, Z. (2022). Frequency-dependent plasticity in the temporal association cortex originates from the primary auditory cortex, and is modified by the secondary auditory cortex and the medial geniculate body. *Journal of Neuroscience, 42*(26), 5254-5267.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 245.

Massaro, D. W., Cohen, M. M., Tabain, M., & Beskow, J. (2012). Animated speech: Research progress and applications In Clark RB, Perrier J, P, & Vatikiotis-Bateson E (Eds.), Audiovisual Speech Processing (pp. 246–272). *Cambridge: Cambridge University*.

Meng, Q., & Schneider, K. A. (2022). A specialized channel for encoding auditory transients in the magnocellular division of the human medial geniculate nucleus. *Neuroreport, 33*(15), 663-668.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507.

O'Hanlon, B. L., Riecke, L., Hausfeld, L., Usherwood, B., Plack, C. J., & Nuttall, H. E. (in prep., 2025). *Effects of Short-Term Audio-Tactile Training on Cortical Speech-Envelope Tracking and Speech Intelligibility*. [Unpublished Manuscript]. Department of Psychology, Lancaster University.

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience, 21*(4), 474-483.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience, 32*(41), 14125-14131.

Ren, Y., Yang, W., Nakahashi, K., Takahashi, S., & Wu, J. (2017). Audiovisual integration delayed by stimulus onset asynchrony between auditory and visual stimuli in older adults. *Perception, 46*(2), 205-218.

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*, 116134.

Schwartz, J. L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10*(7), e1003743.

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2022). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering, 18*(6), 066054.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Todaro, C., Marzetti, L., Valdés Sosa, P. A., Valdés-Hernandez, P. A., & Pizzella, V. (2019). Mapping brain activity with electrocorticography: resolution properties and robustness of inverse solutions. *Brain Topography, 32*, 583-598.

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America, 152*(6), 3216-3225.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, *19*(2), 181-191.

Yuan, Y., Meyers, K., Borges, K., Lleo, Y., Fiorentino, K. A., & Oh, Y. (2021). Effects of visual speech envelope on audiovisual speech perception in multitalker listening environments. *Journal of Speech, Language, and Hearing Research, 64*(7), 2845-2853.

Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage, 124*, 16-23.

Chapter 6

6 General Discussion and Conclusions

In this thesis, the benefits of audio-visual and audio-tactile integration for speech perception have been explored in the context of both behavioural outcomes through speech intelligibility, and neuronal outcomes through EEG- and ECoG-based cortical speech tracking accuracy. In Chapter 2, the general behavioural benefits of audio-visual speech were reassessed, finding expected restoration of intelligibility in noise compared to auditory only speech. Interestingly, however, this benefit was limited to specific viseme categories and was diminished in comparison to previous audio-visual literature. Unexpectedly, stimulus onset asynchrony manipulation could not identify the upper limit of integration for the specific speech phonemes used in this experiment. This may relate to the study being delivered online, due to stimulus-dependent effects (Ma *et al.*, 2009), or even due to the window of integration for audio-visual stimuli having shifted or expanded post COVID-19, though this is only speculative. Despite seeing this diminished and stimulus-dependent benefit with audio-visual speech perception, audio-visual integration remains a robust means of enhancing speech intelligibility in noise. This thesis then aimed to investigate other potential multisensory integration pathways that may be of interest to the improvement of speech perception in difficult listening conditions.

To that end, Chapter 3 investigated the potential benefits of audio-tactile integration. Previous audio-tactile integration research has demonstrated neural enhancement of cortical speech-envelope tracking (Riecke *et al.*, 2019). This neural enhancement was not accompanied by any behavioural increases in speech intelligibility to speech-in-noise. This goes against the assumption that speech intelligibility and neural speech tracking accuracy are

intrinsically linked. One explanation for this could be that the neural tracking enhancement is a pre-cursor to behavioural benefit, as we do not naturally experience audio-tactile speech-envelope integration in real-world environments like we are exposed to audio-visual speech integration through visual lipreading. Behavioural benefit may arise alongside this enhancement of neural tracking through training. To that end, short-term audio-tactile speech training was given to participants in Chapter 3 whilst investigating the effects of training on speech intelligibility and neural tracking accuracy. As expected, without training, participants showed an enhancement in tracking accuracy when audio-tactile speech was present over auditory only speech, but no differences in speech intelligibility were observed. Post-training, however, congruent audio-tactile speech training did not confer any intelligibility benefits relative to incongruent training, and no further neural tracking enhancements were observed. This is with exception to the pseudo-trained group (active control), who saw significant enhancement post-training for auditory-only sentences. Again, this was not linked to increased intelligibility, further providing evidence against the link between the two measures and failing to demonstrate audio-tactile training to enhance speech perception relative to unimodal processing. With these findings from Chapter 3, three possible explanations or the findings remain particularly interesting. One is that audio-tactile integration is simply insufficient to benefit processing of complex speech sentences and speech-envelope derived decoding. Another possible explanation is that the type of training used in Chapter 3, which primarily integrated bottom-up processes for the tactile information stream, may not have been sufficient to induce significant audio-tactile training effects in the short-term. From this, it was suggested that a focused, top-down training paradigm that utilised selective attention to the tactile stimulation may have provided missing intelligibility benefit, which was investigated in Chapter 4. With Chapter 4, where top-down audio-tactile training was assessed, preliminary findings show no post-training benefit to neural tracking accuracy or

speech intelligibility, despite training focusing on the top-down modulations of the tactile element during integration. Whilst general auditory benefit to speech intelligibility and neural tracking was high, even without sleep consolidation, the lack of tactile benefit seen post-training does suggest that audio-tactile integration was insufficient for enhancing speech perception in both neural and behavioural output. As this is a preliminary analysis, conclusions regarding the insufficiency of audio-tactile integration and training may not remain the same when a final, fully-powered sample is obtained and analysed. Effect sizes for the tactile benefit to speech intelligibility (TbSI) were comparable (r^2 = .17) to that of the original priori sample size calculation (small to medium; r^2 = .17), whereas tactile benefit to neural speech tracking accuracy (TbRz) was much lower (r^2 = .03). This may indicate that, whilst results for TbSI may not change with a full sample tested, there could still be a significant increase or decrease seen in either training group and either condition for TbRz when all 60 participants required for power are tested and analysed.

The third and final possible explanation as to why there was no link between tracking accuracy and intelligibility seen is that there is different potential role of neural tracking in the auditory cortex to that of speech segmentation and understanding. Here, Chapter 5 utilised pre-existing intracranial audio-visual data to investigate if this role of neural tracking was related to the prediction of oncoming speech, as well as to further knowledge of audio-visual integration in the posterior superior temporal gyrus (pSTG). This data presented audio-visual and auditory only speech to participants, with some audio-visual trials presenting as mouth-leading (the onset of the visual information preceded the onset of the auditory information) or voice-leading (vice-versa). In the original dataset, conclusions were drawn through these leading conditions that audio-visual speech assists with the prediction of oncoming speech by suppressing incorrect phoneme expression in the pSTG, which was present in mouth-leading but not voice-leading conditions. Chapter 5 hypothesised that, if

neural tracking also played a role in the prediction of oncoming speech, then visual benefit to tracking accuracy would be higher in the mouth-leading conditions versus voice-leading. Findings of this chapter, however, were nonsignificant, showing no changes to tracking accuracy between both leading conditions. This may, in part, be due to large participant variance in the data. As such, no further insight into audio-visual speech processing were uncovered with this analysis, and the role of neural speech tracking in the brain may not be relevant to the prediction of oncoming speech.

6.1 Audio-visual Speech Integration and its Benefit to Speech Perception

Audio-visual integration was examined in Chapters 2 and 5 with varying results. Chapter 2 investigated audio-visual speech benefit to intelligibility of phonemes in noise, specifically with chosen phonemes that belonged to visually distinct viseme categories. Here, visual benefit was present as expected, with audio-visual speech increasing speech-in-noise intelligibility compared to auditory only speech. However, this increase in audio-visual benefit was not as large as in previous literature. This may be indicative of experimental design differences, as phonemes selected for discrimination in this experiment were of different viseme categories. This would be further supported by exploratory analyses in this chapter, which highlighted differences in intelligibility between each of the three phonemes (Ba, Fa, and Ka) in noise. This experiment also presented audio-visual speech at varying levels of stimulus onset asynchrony, with the onset of the visual information preceding the onset of the auditory information by six different levels between 200 and 266.6 ms. This was to identify a window of integration for this set of stimuli, wherein the maximum level of visual onset asynchrony before visual benefit to speech intelligibility or processing speed was lost could be identified. Here, no window of integration could be determined using speech discrimination accuracy or response time measurements. This finding is also intriguing, as it suggests that the window of integration for audio-visual stimuli has expanded compared to

previous phoneme discrimination research (Navarra *et al*., 2005), although other research had found varying windows of integration with short-words and other speech stimuli (Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al.,* 2017). Relative to aging research, a wider window of integration typically indicates poorer performance, as there is a longer period in which an erroneous visual cue in the environment may be misinterpreted as speech-relevant (Ganesh *et a*l., 2018; Sekiyama *et al.*, 2014; see also development disorder differences with: Smith & Bennetto, 2007; Megnin-Viggars & Goswami, 2013; Michalek *et al*., 2014; Noel *et al*., 2018). Therefore, this result may be indicative of poorer integrative ability across the participants tested. Future research should readdress this conflicting window of integration period from previous literature to gain further insight into the mechanisms at play in current audio-visual environments.

Chapter 5, on the other hand, investigated audio-visual integration in an entirely neural perspective. This chapter presented a secondary data analysis of previous intracranial audio-visual work, where the pSTG was identified as suppressing incorrect phoneme representation to assist with accurate prediction of oncoming speech. Here, we applied neural speech tracking accuracy analyses to the intracranial data to examine if speech tracking played a role in speech prediction as well. Findings of this chapter demonstrated no supporting evidence to suggest that neural speech tracking of audio-visual speech played a role in assisting with speech prediction in the pSTG. As another interesting note, there was also no visual benefit to tracking observed at baseline in this experiment, going against previous tracking work showing enhancements in cortical tracking accuracy with audio-visual speech presentation (Haider *et al.,* 2024; Golumbic *et al.*, 2013). In speculation, this may suggest that audio-visual speech is not tracked at the specific location of the pSTG (Kubanek *et al.*, 2013). Given this work's intracranial nature, future research may wish to reinvestigate this paradigm using whole-scalp non-invasive methodologies, such as EEG, to

provide further insight into how audio-visual speech is represented across cortical areas other than the pSTG.

6.2 Audio-tactile Integration and its Benefit to Neural Speech Tracking but Not Speech Intelligibility

Audio-tactile integration in speech perception was investigated with non-invasive electroencephalography work in Chapter 3 and Chapter 4. Both chapters presented both audio-tactile and auditory-only sentences-in-noise during a speech discrimination task whilst recording cortical oscillations for use in neural tracking accuracy analysis. Participants in the target audio-tactile training groups experienced either a bottom-up (Chapter 3) or top-down (Chapter 4) audio-tactile training paradigm, before speech intelligibility was reassessed following training with further speech-in-noise discrimination. Here, Chapter 3 captured an enhancement of neural speech tracking accuracy pre-training for audio-tactile versus auditory only speech sentences, reflecting a more accurate representation of the speech cortically when speech-relevant tactile stimulation was provided. This, as expected, was not accompanied by any benefit to speech intelligibility with increases in performance in the speech-in-noise discrimination task.

Interestingly, this finding was not present in pre-training data from Chapter 4, despite the experiment using the same pre-training task and experimental stimuli. This may be due to Chapter 4 presenting an underpowered preliminary analysis, which with further data collection may see pre-training enhancement to neural tracking accuracy, as the current effect size for TbRz remains much lower than that of the sample size calculation (r^2: .03 < .17). Alternatively, this finding may be relevant in explaining why audio-tactile speech did not see further enhancement in tracking accuracy post-training in either study. One key pre-training difference between Chapters 3 and 4 was that tactile familiarisation was presented to both training groups in Chapter 4 prior to beginning the pre-training task. As such, any potential

novelty or surprise at receiving tactile stimulation for participants in Chapter 3 would not have been present in the pre-training task for participants in Chapter 4. If the enhancement seen in neural speech tracking accuracy for audio-tactile speech in Chapter 3 was due to erroneous processing of a surprising or new stimuli, rather than reflective of a speech processing benefit, then this would further explain why no intelligibility increases were seen alongside tracking accuracy enhancement. Future audio-tactile speech work may wish to investigate how unfamiliar and familiar tactile stimulation may affect neural speech tracking accuracy measures to ensure that no erroneous processing is captured by stimulus reconstruction.

Furthermore, post-training data in both experiments highlighted no further tactile-relevant increases in either speech intelligibility or neural tracking accuracy. Arguably, this evidence suggests that audio-tactile integration is – at least after a state of short-term bottom-up and limited single session top-down training – insufficient to drive any meaningful benefit to the processing of speech. With regards neural tracking, it may be that the speech-envelope tracking benefit missing post-training was driven by oscillatory activity outside of the frequency ranges used for stimulus reconstruction, or deeper in subcortical regions not represented by descending, top-down oscillatory activity from the auditory cortex. For example, the sensorimotor cortex has been seen to benefit from beta-band oscillatory activity (~13 – 30 Hz; Morillon *et al.,* 2019), which mostly lies outside our optimised frequency range for neural tracking analysis of 0.5 – 15 Hz. Speculatively, the initial audio-tactile speech benefit to tracking may have been influenced by low beta activity (13 – 15 Hz) captured from the sensorimotor regions during tactile stimulation, reflective in the audio-tactile enhancement of tracking accuracy. Post-training, however, audio-tactile integration processes may utilise faster beta frequencies outside of our captured range, masking potential enhancements to tracking accuracy in our data. Future research may wish to investigate

potential beta-band influence on audio-tactile integration, within the sensorimotor cortex, to better understand how speech-relevant tactile information is represented in the wider processing network.

6.3 Effectiveness of Top-Down versus Bottom-up Training Paradigms

Training paradigms were utilised to train audio-tactile speech integration for experiments described in Chapters 3 and 4. Whilst both chapters investigated short-term training effects to speech intelligibility through speech-in-noise discrimination, as well as neural speech-envelope tracking accuracy during listening, the training paradigms employed in both studies differed in their targeted direction of auditory pathway involved in speech perception. Here, Chapter 3 utilised what was argued to be a bottom-up, sensory driven paradigm for tactile integration with auditory processing. On the other hand, Chapter 4 presented participants in the audio-tactile training group with a top-down training paradigm. This training paradigm aimed to utilise selective attentional processes specifically with the tactile element, by presenting both congruent and incongruent tactile stimulation to speech both with and without noise. Here, participants were asked to actively identify which tactile stimulation was congruent to the listened speech, with incorrect responses in a trial resulting the sentence being added back into the training pool for reassessment.

Despite neither training method evidencing benefit to audio-tactile speech intelligibility or speech tracking accuracy, there is evidence that top-down training was more effective at improving general intelligibility of speech-in-noise for auditory-only sentences. In Chapter 3, training was spread out across three separate days, with time for sleep consolidation between each. Furthermore, a total of 90 unique sentences were presented during training. In contrast, Chapter 4 presented top-down training during a single session only. This training took place immediately after the pre-training task and before the post-training task, with no sleep consolidation before retesting. The experiment also only allocated

60 unique sentences for training, although these sentences were repeated if participants responded to training trials incorrectly. Despite these drawbacks, top-down training on average still saw an approximately 7% increase in speech intelligibility performance for the audio-tactile training group post-training for auditory only sentences, though the benefit was not specific to the audio-tactile group alone. This is a larger average increase than what was seen in the audio-tactile training group after four sessions (5%), and comparable to increases seen in the pseudo-trained group for that chapter (6 – 7%). This does help speculate as to if further sleep consolidation with top-down audio-tactile training would induce greater training benefits to intelligibility and tracking accuracy (Drouin *et al.*, 2023).

6.4 What is the Role of Neural Speech Tracking in Speech Processing?

Throughout Chapters 3, 4, and 5, the neural measurement of neural speech tracking accuracy has been investigated in relation to the intelligibility of audio-visual, audio-tactile, and auditory only speech. In Chapter 3, enhanced neural tracking accuracy was observed following initial introduction to audio-tactile speech through envelope-shaped tactile stimulation at baseline. This was not accompanied by any benefits to speech intelligibility, in line with previous audio-tactile and auditory only work (Riecke *et al.*, 2019; Kösem *et al.*, 2023). Even after short-term bottom-up training with audio-tactile stimuli, no further intelligibility increases relative to tactile stimulation were found. In contrast, neural tracking accuracy in the pseudo-trained group increased for auditory-only stimuli, still with no relationship to changes in intelligibility. This lack of intrinsic link was further supported in Chapter 4, where top-down audio-tactile training remained insufficient in inducing tactile-relevant intelligibility benefits. Although, experimental findings in Chapter 4 also failed to capture any enhancement of neural tracking accuracy with audio-tactile speech pre-training as well. Taken together, findings from both audio-tactile chapters indicate that the role of

neural speech tracking in speech processing is not related directly to the intelligibility of speech, nor does it play a role in assisting speech segmentation.

This raises the question as to what speech tracking in the auditory cortex is utilised for. Chapter 5 attempted to support the hypothesis that tracking plays a role in the prediction of oncoming speech signals, much akin to how visual lipreading primes the pSTG to predictions of oncoming speech through the suppression of incorrect phonemes, likely processed as viseme categories. However, results from this preliminary analysis were also unexpected, with no support found for neural tracking in the pSTG playing an assisting role in speech prediction. Potentially, tracking may assist with speech prediction in a separate auditory processing region than the pSTG, such as through integration hubs on the lateral geniculate nucleus or the inferior colliculus, which were not represented in the intracranial dataset used. As such, it may be beneficial for future research to investigate neural tracking accuracy using whole-scalp methods like EEG and MEG to pick up oscillatory activity relevant to wider corticofugal modulation of subcortical integration regions, or through direct intracranial measures recorded at these subcortical areas of interest. In all, these chapters indicate that the role of neural speech tracking requires further consideration in future speech processing work.

6.5 Limitations

Throughout all the presented experimental chapters, limitations are present which must be considered when evaluating the contributions of this thesis to wider auditory research. Firstly, Chapter 2 is an online behavioural experiment, conducted partly during the COVID-19 pandemic and partly after. As such, it is difficult to associate the findings of audio-visual benefit to typical listening environments that we find ourselves in, as they may be more applicative to contexts where less face-to-face social interaction was present during speech listening (Brown *et al.,* 2021). Moreover, this experiment produces only behavioural

outcomes. As such, it is difficult to use findings from this chapter to pinpoint specific mechanisms for audio-visual integration in the auditory pathway and surrounding networks. Likewise, Chapter 5 presents a neural perspective only to audio-visual speech integration through neural speech tracking accuracy measurements in the pSTG. It may be of benefit for future research to investigate non-invasive methodologies when examining the neural tracking accuracy of mouth- and voice-leading audio-visual speech, combined with further behavioural testing akin to Chapters 3 and 4, to provide a more holistic insight into audio-visual integration during speech processing.

With regards to the audio-tactile training studies in Chapters 3 and 4, these are limited in scope by being short-term or single-session training experiments. Long-term training with audio-tactile speech may see benefits in speech intelligibility and neural representations that were not apparent in the presented work. In the case of the top-down training paradigm, short-term top-down audio-tactile training may still be effective with the introduction of sleep consolidation into the experimental design. Furthermore, the optimisation parameters used when reconstructing speech features in Chapters 3, 4, and 5 do highlight further limitations in this thesis. This thesis provides a basis for analysis the tracking accuracy of shorter sentences using EEG without training the decoder to the seams of each sentence (which typically leads to erroneous output: Crosse *et al.*, 2016). Here, by training decoders to longer segments of continuous speech (10-minute story) and removing initial ERP responses to the onset of each stimulus, the shorter sentences can then be joined into longer chunks of data for reconstruction analysis with successful results. Despite this successful demonstration of utilising reconstruction decoder models for shorter sentences in EEG, reconstruction is still made more valid with more task-relevant neural data available (Destoky *et al.*, 2019). Presenting more trials per condition with longer segments of continuous speech would result in more accurate measures of neural speech tracking accuracy. This is of note for Chapter 5,

where intracranial data was limited to broadband high-gamma activity from only a few selected electrodes per participant. This in turn may have led to the large participant variability seen in the analysis. Future work looking to utilise reconstruction methods should be aware of the challenges faced in this thesis when designing neural speech tracking based experiments.

6.6 Future Directions

Based on the findings of this thesis, future research may wish to investigate further the neural mechanisms underlying both audio-visual and audio-tactile speech integration. It would also be beneficial for more work to be done investigating the true role of neural speech tracking in speech processing. From Chapter 2, the benefits of audio-visual speech to intelligibility were seen to be variable dependent on the phoneme, and thus potentially the viseme category, presented. It would be beneficial to research in audio-visual speech to understand how visemes are processed in multisensory integration hubs like the inferior colliculus (Shore, 2005; Balmer & Trussell, 2021), and to investigate if there are more stimulus-dependent effects of other phonemes and viseme categories. From Chapters 3 and 4, where both top-down and bottom-up short-term audio-tactile training appears to be ineffective at providing enhanced benefit to speech intelligibility in noise, future research may wish to consider how sleep consolidation could facilitate short-term top-down training effects (Drouin *et al.*, 2023) or investigate longer-term audio-tactile training. Finally, it is crucial that an understanding of the role of neural speech tracking in speech perception and processing is strengthened. Following Chapter 5, a non-invasive examination of the tracking accuracy differences between mouth-leading and voice-leading audio-visual speech will provide clarity to the potential role of tracking in speech prediction outside of the pSTG, as well as strengthen our understanding of audio-visual speech integration. Although, it may be that tracking plays an entirely different role altogether, which should be further investigated.

In example, attentional decoding might be the key to understanding how neural tracking is utilised. This understanding is also crucial in assisting with the continued development of dynamic, neuro-steered hearing aids, which aim to utilise neural tracking accuracy during real-time listening to tailor hearing aid outputs to a user's needs (see: Geirnaert *et al*., 2024; Straetmans, 2022).

6.7 Conclusion

In conclusion, this thesis demonstrates the importance of non-auditory sensory cues in assisting with speech processing when listening is difficult. The well documented benefit of audio-visual speech integration has been reassessed using visually-distinct phoneme selection following the COVID-19 pandemic, highlighting key differences in benefit to speech intelligibility and to the window of integration of audio-visual speech. Moreover, audio-tactile speech has been shown to be insufficient at providing speech intelligibility or neural speech tracking accuracy enhancement with both bottom-up and top-down training paradigms, indicating a need to revisit the types of tactile stimulation that may be provided to assist with speech perception in difficult listening conditions. Finally, through work in both the audio-visual and audio-tactile domains, evidence has been found to support the notion that speech intelligibility and neural tracking accuracy are not intrinsically linked, nor does speech tracking in the posterior superior temporal gyrus indicate a role in assisting with predictions of oncoming speech. Future work should prioritise the role of speech tracking in the auditory pathway, as this will provide further insight into how multisensory integration is represented in the brain during listening and aid with the continued development of neuro-steered hearing aids.

References

Balmer, T. S., & Trussell, L. O. (2021). Trigeminal contributions to the dorsal cochlear nucleus in mouse. *Frontiers in Neuroscience, 15*, 715954.

Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications, 6*(1), 49.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in Integrative Neuroscience, 4*, 1316.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., ... & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage*, *184*, 201-213.

Drouin, J. R., Zysk, V. A., Myers, E. B., & Theodore, R. M. (2023). Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech. *Journal of Speech, Language, and Hearing Research, 66*(2), 720-734.

Drouin, J. R., Zysk, V. A., Myers, E. B., & Theodore, R. M. (2023). Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech. *Journal of Speech, Language, and Hearing Research, 66*(2), 720-734.

Ganesh, A. C., Berthommier, F., & Schwartz, J. L. (2018). Audiovisual binding for speech perception in noise and in aging. *Language Learning, 68*, 193-220.

Geirnaert, S., Zink, R., Francart, T., & Bertrand, A. (2024). Fast, accurate, unsupervised, and time-adaptive EEG-based auditory attention decoding for neuro-steered hearing devices. *In Brain-Computer Interface Research: A State-of-the-Art Summary 11* (pp. 29-40). Cham: Springer Nature Switzerland.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, *33*(4), 1417-1426.

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of *Cognitive Neuroscience, 36*(1), 128-142.

Kösem, A., Dai, B., McQueen, J. M., & Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage, 272*, 120040.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PLoS One*, *8*(1), e53398.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One, 4*(3), e4638.

Megnin-Viggars, O., & Goswami, U. (2013). Audiovisual perception of noise vocoded speech in dyslexic and non-dyslexic adults: the role of low-frequency visual modulations. *Brain and Language, 124*(2), 165-173.

Michalek, A. M., Watson, S. M., Ash, I., Ringleb, S., & Raymer, A. (2014). Effects of noise and audiovisual cues on speech processing in adults with and without ADHD. *International Journal of Audiology, 53*(3), 145-152.

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience & Biobehavioral Reviews, 107*, 136-142.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507.

Noel, J. P., Stevenson, R. A., & Wallace, M. T. (2018). Atypical audiovisual temporal function in autism and schizophrenia: similar phenotype, different cause. European *Journal of Neuroscience, 47*(10), 1230-1241.

Ren, Y., Yang, W., Nakahashi, K., Takahashi, S., & Wu, J. (2017). Audiovisual integration delayed by stimulus onset asynchrony between auditory and visual stimuli in older adults. *Perception, 46*(2), 205-218.

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*, 116134.

Schwartz, J. L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10*(7), e1003743.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology, 5*, 323.

Shore, S. E. (2005). Multisensory integration in the dorsal cochlear nucleus: unit responses to acoustic and trigeminal ganglion stimulation. *European Journal of Neuroscience, 21*(12), 3334-3348.

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry, 48*(8), 813-821.

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2022). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering, 18*(6), 066054.