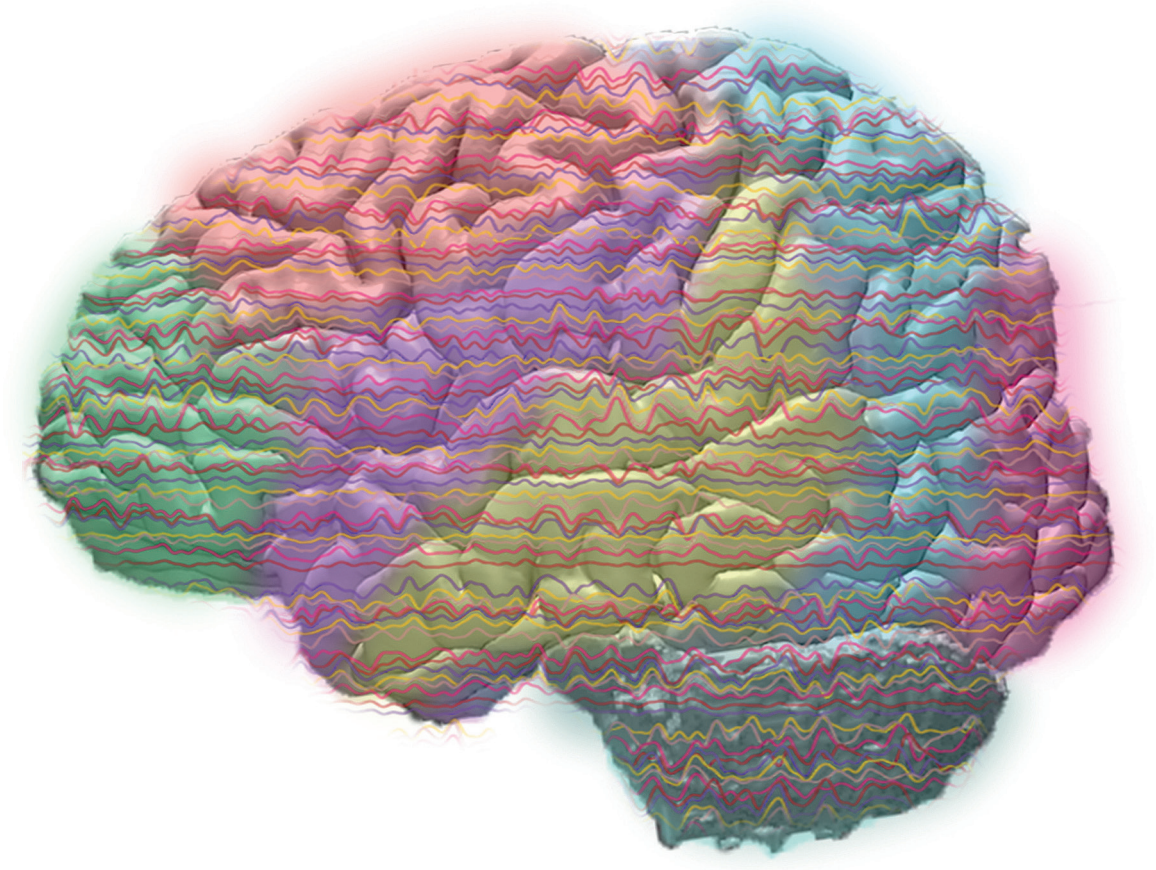
JYU DISSERTATIONS 459

Orsolya Beatrix Kolozsvári

Investigation of Brain Processes of Speech Perception and Production in Adults and Children Using MEG



UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF EDUCATION AND PSYCHOLOGY

JYU DISSERTATIONS 459

Orsolya Beatrix Kolozsvári

Investigation of Brain Processes of Speech Perception and Production in Adults and Children Using MEG

Esitetään Jyväskylän yliopiston kasvatustieteiden ja psykologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Historica-rakennuksen salissa H320
joulukuun 3. päivänä 2021 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Education and Psychology of the University of Jyväskylä,
in the building Historica, auditorium H320 on December 3, 2021 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2021

Editors
Noona Kiuru
Department of Psychology, University of Jyväskylä
Ville Korkiakangas
Open Science Centre, University of Jyväskylä

Cover picture by Orsolya Beatrix Kolozsvári.

ISBN 978-951-39-8927-9 (PDF)
URN:ISBN:978-951-39-8927-9
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-8927-9

ABSTRACT

Kolozsvári, Orsolya Beatrix
Investigation of brain processes of speech perception and production in adults and children using MEG
Jyväskylä: University of Jyväskylä, 2021, 62 p.
(JYU Dissertations
ISSN 2489-9003; 459)
ISBN 978-951-39-8927-9 (PDF)

Spoken language is a large part of our day-to-day life. During face-to-face communication, one needs to not only be able to follow and integrate the incoming auditory and visual information, but usually also to respond to and produce speech. While speech perception has been investigated rather thoroughly, from syllable-level perception to paragraph-long continuous tracking, there are still questions remaining regarding how the brain perceives speech. Speech production presents an even bigger challenge to investigate at brain level, as not only does it involve movement from articulators, which add muscle activity artifacts to the recorded signal, but in overt speech, the brain needs to process the perception of hearing ourselves speak. Brain-level studies often choose to use simplified stimuli, to ease the technical challenges involved in the measures, and thus comparison of the different linguistic units remains a rarity. This dissertation investigated brain-level correlates of speech perception at the syllable level in an audio-visual perception task, examined the developmental hemispheric differences in speech tracking of words and sentences, as indexed by coherence measures, in children and adults, and explored a method of analysis previously used in speech perception research to investigate speech production at the brain level, which could be used in future research. Study I investigated brain-correlates of audio-visual speech perception at syllable level, in particular the effect of congruence and familiarity of stimuli on brain responses in adults of two different language backgrounds (Finnish and Chinese). Congruence of stimuli, that is when what they saw matched with what they heard, was found to have an effect and differences in responses were found mainly in the right temporal areas. Study II examined the brain's ability to track the speech envelope of words and sentences in children and in adults, and correlated the findings with their auditory responses to syllables to investigate a possible overlap in development for the two processes. Furthermore, the brain indices were correlated with speech-related cognitive measures in children. The findings show an improvement with age in speech tracking, and that this increase is independent from the changes in the auditory response to simpler speech stimuli (syllables). Study III introduced a novel approach to measure speech tracking using an accelerometer during overt speech in adults while their brain activity was recorded.

Keywords: Speech tracking, Speech production, Coherence measure, Event-related responses, Magnetoencephalography

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Kolozsvári, Orsolya Beatrix
Aikuisten ja lasten puheen havaitsemiseen ja tuottamiseen liittyvien aivoprosessien tutkiminen MEG:llä
Jyväskylä: University of Jyväskylä, 2021, 62 p.
(JYU Dissertations
ISSN 2489-9003; 459)
ISBN 978-951-39-8927-9 (PDF)

Puhuttu kieli on tärkeä osa arkeamme. Kasvokkain tapahtuvassa viestinnässä on pystyttävä seuraamaan ja yhdistelemään vastaanotettua kuultua ja nähtyä informaatiota, mutta yleensä on myös pystyttävä reagoimaan puheeseen ja tuottamaan sitä. Vaikka puheen havaitsemista on tutkittu suhteellisen perusteellisesti, aina tavujen havaitsemisesta keskustelujen seuraamiseen asti, on yhä osittain epäselvää, kuinka aivot havaitsevat puheen. Puheen tuottamista on vieläkin haasteellisempaa tutkia aivotasolla, koska siihen ei liity pelkästään puhe-elinten liike, joka lisää lihastoiminnan häiriöitä tallennettuun signaaliin, vaan ääneen lausutussa puheessa aivojen tarvitsee prosessoida havainto kuulemastaan omasta puheesta. Aivotason tutkimuksissa käytetään usein yksinkertaistettuja ärsykkeitä aivotoiminnan mittaamiseen liittyvien teknisten haasteiden vuoksi, joten kielellisten yksiköiden välinen vertailu on harvinaista.

Tässä väitöstutkimuksessa selvitettiin puheen havaitsemisen aivotason korrelaatteja tavutasolla audiovisuaalisen havaintotehtävän avulla. Lasten ja aikuisten aivopuoliskojen toiminnan kehityksellisiä eroja puhuttujen sanojen ja virkkeiden seuraamisessa tutkittiin koherenssimittojen perusteella. Lisäksi perehdyttiin aiemmin puheen havaitsemisen tutkimuksessa käytetyn analyysimenetelmän soveltamiseen tutkittaessa puheen tuottamista aivotasolla; sitä voitaisi käyttää myös tulevissa tutkimuksissa.

Tutkimuksessa 1 tutkittiin audiovisuaalisen puheen havaitsemisen aivokorrelaatteja tavutasolla ja erityisesti ärsykkeiden kongruenssin ja tuttuuden vaikutusta aivovasteisiin aikuisilla, joiden kielelliset taustat ovat erilaiset (suomi ja kiina). Ärsykkeiden kongruenssilla – eli kun se, mitä tutkittavat näkivät, vastasi heidän kuulemaansa verrattuna siihen kun se, mitä nähtiin ei vastannut kuultua – havaittiin olevan merkitystä, ja vasteet erosivat pääasiassa oikeilla ohimoalueilla. Tutkimuksessa 2 tarkasteltiin lasten ja aikuisten aivojen kykyä seurata puheen voimakkuusvaihtelua sanoissa ja virkkeissä; kyky voimakkuusvaihtelun seuraamiseen korreloitiin tavujen aivovasteiden kanssa, jotta voitiin tutkia kahden prosessin kehityksen mahdollista päällekkäisyyttä. Lisäksi aivoindeksit korreloitiin kognitiivisten mittojen kanssa. Tulokset osoittavat, että puheen seuraaminen paranee iän mukana ja että tämä kehitys on riippumaton muutoksista aivovasteessa yksinkertaisempiin puheärsykkeisiin (tavuihin). Tutkimuksessa 3 otettiin käyttöön uusi puheen seuraamisen mittausmenetelmä, eli mitattiin aikuisten puhetta kiihtyvyysanturilla tallentaen samalla heidän aivotoimintaansa.

Asiasanat: Puheen seuraaminen aivoaktiivisuudessa, puheen tuottaminen, herätevaste, koherenssimitta, magnetoenkefalografia, MEG

Author

Orsolya Beatrix Kolozsvári
Department of Psychology
University of Jyväskylä
orsolya.b.kolozsvari@jyu.fi
https://orcid.org/0000-0002-1619-6314

Supervisors

Professor Jarmo A. Hämäläinen
Department of Psychology
University of Jyväskylä

Dr. Aude Noiray
Linguistics Department
University of Potsdam

Professor Paavo H. T. Leppänen
Department of Psychology
University of Jyväskylä

Reviewers

Professor Mathieu Bourguignon
Faculté des Sciences de la Motricité
Université Libre de Bruxelles

Professor Judit Gervain
Department of Developmental and Social Psychology
University of Padua

Opponent

Professor Judit Gervain
Department of Developmental and Social Psychology
University of Padua

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Jarmo Hämäläinen, for his unwavering support through my PhD studies. I cannot thank him enough for giving me the opportunity to be part of the PredictAble project and for being a top-notch supervisor. I would also like to thank Dr Aude Noiray, my second supervisor, for her help and support during the project and our shared manuscripts. I would also like to thank Professor Paavo Leppänen for his valuable comments, suggestions and support and for being Paavo. I wish to also thank my opponent Judit Gervain and my external reviewer Mathieu Bourguignon for their time and effort in reviewing my dissertation and for their kind comments and suggestions for improving my work.

Thank you to Tiina Parviainen for her support and advice on all things MEG and for running the Centre for Interdisciplinary Brain Research (CIBR) here at the University. Thanks also to Jan Wikgren, Piia Astikainen, Markku Penttonen, Miriam Nokia, and Kaisa Lohvansuu for the discussions and times shared in both the seminars and the social gatherings. Special thanks to Lea Nieminen for her help in our speech perception and production paradigm and her helpful contributions to the second manuscript. I also want to thank Jan Kujala for his contributions to and invaluable help with the third manuscript.

I wish to thank all my colleagues at the Department of Psychology for their support and for putting up with me during all these years. Thank you for making the journey enjoyable and fun. Special thanks to Weiyong Xu, Natalia Louleli, Najla Azaiez Zammit Chatti, Sam van Bijnen and Praghajieeth Santana Gopalan – it is a great honor to know you all and to have gained your friendship through these years. I cannot wait to see what you all do next. 谢谢, Ευχαριστώ, اركشً, Dank u, நன்றி.

Special thanks to Georgia, who was invaluable in helping with writing the articles and was an unending source of support. Thank you for listening, for the laughter and for having the best anecdotes :). I am also grateful to the PhD students (many of whom have already gained their own titles or are just about to) here at the University: Otto, Doris, Ilona, Laura, Elisa, Jari, Juho, Xueqiao, TianTian, Ana, Paola, Fatemeh, Lili, Eve, Ariane, Lisa, Daria, and Yixue. Thank you for the times shared together, the discussions and laughter over coffee, the support and for listening. I wish you all the best of luck, whatever you may do next. I am also very thankful to Chunhan Chiang and Anna-Maria Alexandrou for their friendship and support.

I must thank all the participants, and their families in case of the child participants, who took part in the measurements. I need to also thank Katja Koskialho, Sonja Tiri, Ainomaija Laitinen, Annamaria Vesterinen, Aino Sorsa, Maija Koskio, Hanna-Maija Lapinkero and Cherie Jenkins for their help with data collection in the MEG lab. Without you, this dissertation would not exist. Speaking of MEG, I need to thank Simo Monto for his help and patience with my emails about data analysis and Viki-Veikko Elomaa for his support in the lab and keeping my spirits up.

I want to thank Dr Ferenc Honbolygó, for telling me all those years ago that "There is a PhD opportunity you might be interested in" back in 2015. Thank you for your support and encouragement. I also want to thank the rest of the group - Gabi, Vera, Dávid, Ágoston, Andi, Ádám, Anett, Linda, Dénes and Valéria - for their continued support and friendship through the years and for always welcoming me whenever I visit home.

Lastly, I am thankful to my parents, Éva and Lajos, for encouraging me and supporting me through the years - without you, I would literally not be here. Köszönöm a támogatást és ösztönzést – nélkületek szó szerint nem lennék itt. Péter, Zsolti, Andi, Szilvi, Bálint, Dávid, Toto, Mama és a család összes többi tagja – Köszönöm. Berci, Marci, Szofi, Héra, Kornél, Samu, Tuk, Nyt, Judy, Piipi <3

This dissertation has been supported by the European Union projects PredictAble (Marie Curie Innovative Training Network no. 641858), ChildBrain (Marie Curie Innovative Training Network no. 641652), the Academy of Finland (MultiLeTe #292466), the Nyyssönen Foundation, the Finnish Cultural Foundation Regional fund and the Department of Psychology, University of Jyväskylä.

Jyväskylä 01.11.2021
Orsi

LIST OF ORIGINAL PUBLICATIONS

I. Kolozsvári, O. B., Xu, W., Leppänen, P. H. T., & Hämäläinen, J. A. (2019). Top-down predictions of familiarity and congruency in audio-visual speech perception at neural level. *Frontiers in Human Neuroscience, 13*, 243. https://doi.org/10.3389/fnhum.2019.00243

II. Kolozsvári, O. B., Xu, W., Gerike, G., Parviainen, T., Nieminen, L., Noiray, A., & Hämäläinen, J. A. (2021). Coherence between brain activation and speech envelope at word and sentence levels showed age-related differences in low frequency bands. *Neurobiology of Language, 2*(2), 226-253. https://doi.org/10.1162/nol_a_00033

III. Kolozsvári, O. B., Xu, W., Gerike, G., Kujala, J., Parviainen, T., Leppänen, P. H. T., Noiray, A., Hämäläinen, J. A. (2021) Cortico-kinematic and cortico-acoustic coherence in adults during active speech production. Submitted manuscript.

Taking into account the instructions given and the comments made by the co-authors, the author of this thesis contributed to the original publications as follows: the author designed the experiments, collected the MEG data, conducted the analyses and wrote the manuscripts.

FIGURE

TABLE

CONTENTS

1 INTRODUCTION

Speech perception has been investigated quite thoroughly in the past decades, from comparisons of basic auditory responses to syllables to analysis of oscillatory brain activity to longer speech segments, yet a full account of speech perception has yet to be given. Speech production has been an even greater challenge to investigate at the brain level as it involves the movement of the articulators and thus produces artifacts from muscle activity. Further complicating the situation is that during overt speech, the speaker not only produces the sounds, but also hears themselves speak, and as such the brain needs to process not only the movements and actions of production, but involves active perception of the speech as well. Natural connected speech in particular has been a challenge to investigate – both on the perception and the production side - due to its technically challenging nature. To ease this, brain-level studies often opt to use simplified stimuli, focusing on specific aspects of speech perception, using single syllables or words in isolation and comparison of the different linguistic units has rarely been done.

This dissertation investigated brain-level correlates of speech perception at the syllable level in an audio-visual perception task, examined the developmental hemispheric differences in speech tracking of words and sentences, as indexed by coherence measures, in children and adults and explored a possible new method of analysis of speech production to be used in future research.

1.1 Speech perception at syllable level

During face-to-face communication, speech perception relies on one's ability to integrate the incoming auditory and visual information. Response to audio-visual speech has been found in the cortical areas in a sequential order, starting 100 ms from stimulus onset with activation in the sensory areas (Möttönen, Schürmann, & Sams, 2004; Sams et al., 1991) reflecting the processing of visual features in the occipital cortex and the analysis of acoustic features in the

temporal areas (Salmelin, 2007). This is followed by activation in the superior temporal sulcus (STS). This area has been suggested to be involved in perceiving and interpreting facial and body movements of the speaker (Iacoboni et al., 2001; Puce et al., 1998). Activation has been found at the inferior parietal cortex around 250 ms, which is suggested to connect the superior temporal sulcus to the inferior frontal lobe (Nishitani & Hari, 2002), where activation was stronger on the left hemisphere than the right (Campbell, 2008; Capek et al., 2004).

Furthermore, it has been suggested that seeing speech could also alter how and what is perceived, in a correlated or a complementary way (Campbell, 2008). If the speech is processed in a correlated way, successful processing depends on the similarities between the auditory and visual channels' temporal-spectral signature. In complementary mode, the visual input increases the information available about the speech, information which would not be available in auditory modality and which depends on the speaker's face's visibility.

Audio-visual speech perception involves the listeners' ability to integrate information presented in two modalities, i.e. the auditory speech signal and the visual information of the speaker's mouth moving (Nath and Beauchamp, 2011; Ross et al., 2007). In a realistic situation, these two match and the integration of the two informations give the final speech percept. Congruent orofacial articulatory movements have been found to significantly contribute to speech comprehension (Sumby & Pollack, 1954; van Wassenhove et al., 2005). Incongruent audio-visual stimuli on the other hand, has been found to lead to new audio-visual percepts (McGurk & MacDonald, 1976). Investigating the effects of this congruency gives insight into how the brain processes audio-visual speech specifically.

The match and mismatch between auditory and visual features of speech has been investigated focusing on the congruency of the incoming stimuli (Hein et al., 2007; Jones and Callan, 2003). In this case congruent (matching audio and visual signals) and incongruent (the auditory and visual information does not match) audio-visual pairs are presented and the brain responses to them are contrasted. The general assumption in such comparison is that congruency would effect perception of speech if the information coming in from the unimodal sources have been integrated (van Atteveldt et al., 2007). A broad network of brain areas have been shown to have differing brain responses to congruent and incongruent audio-visual speech stimuli, encompassing inferior parietal lobule, the precentral gyrus, the supramarginal gyrus, the superior frontal gyrus, Heschl's gyrus and the middle temporal gyrus (Callan et al., 2004; Jones & Callan, 2003; Miller & D' Esposito, 2005).

Studies exploring the temporal sequence of cortical responses to audio-visual speech found effects from early on, around the N1 and P2 time-windows (50-150 ms and 150-250 ms respectively) of event-related brain potentials (Baart et al., 2014; Stekelenburg & Vroomen, 2007). Visual information, such as lip-movement, could precede phonation by as much as several hundred milliseconds when producing individual syllables (Sketelenburg & Vroomen, 2007) or 2-syllable words (Baart et al., 2014), which allows predictions to be made by the

listener about features of the auditory signal. N1 and P2 components of ERPs have been shown to be diminished and become earlier when the auditory signal is accompanied by visual input (Besle et al., 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; van Wassenhoe et al., 2005).

Congruency was found to have an effect in the later ERP components as well (Arnal et al., 2009). Using matching and non-matching (or congruent and incongruent) audio-visual syllables, they found no effect of congruency in the M100 time-window (50-130 ms). When comparing the full time course of responses, they found the earliest difference between brain responses around 120 ms following voice onset, and peaks at 250, 370 and 460 ms, suggesting a manifold comparison between the incoming auditory signal and the top-down visual predictions.

Another way to study brain activity related to long-term audio-visual representations, besides direct comparison of brain responses to matching and mismatching audio-visual speech stimuli, is by examining suppressive/additive effects in audio-visual processing as suggested by the additive model (e.g., Besle et al., 2004). This is done by comparing the summed responses to separately presented auditory and visual stimuli to the responses to audio-visual stimuli. If there are no interactions between the modalities, and both features are processed separately, the amplitude of response to bimodal stimuli should be equal to the sum of amplitudes of the unimodal stimuli making it up. Difference between the two amplitudes would suggest an interaction of the sensory modalities. Such suppressive effects of audio-visual processing have been found in the auditory cortices on both hemispheres 150–200 ms and in the right superior temporal sulcus (STS) 250–600 ms after the onset of the stimulus, showing that sensory-specific and multisensory regions of the temporal cortices are involved in the processing of audio-visual speech (Möttönen et al., 2004).

It's also documented that an individual's language experience alters their speech perception, both at the behavioural (Iverson et al., 2003) and at the brain level (Kuhl, 2000; 2004), with participants with different language backgrounds responding to different acoustic features of the same speech stimuli. Long-term memory representations of speech sounds, or lack thereof, have been shown to have an effect on perception at the behavioural level, for example, Japanese speakers have been shown to have trouble differentiating between /r/-/l/ contrasts, as their perception appears to be more sensitive to the F2 or second fundamental frequency, while the determining differences between /r/ and /l/ sounds seem to be in F3 (Iverson et al., 2003). At brain level, language-specific analysis of phonetic-phonological aspects has been suggested to start as early as 100-200 ms after stimulus onset (Näätänen et al., 2007; Vihla et al., 2000). Such effects of familiarity, specifically, if the presented stimulus is part of the listener's phonology, have been demonstrated using the mismatch negativity (MMN). Mismatch negativity (MMN) or mismatch field (MMF) findings have suggested access to phonological categories (Näätänen et al., 2007; Vihla et al., 2000) and the processing of native and non-native phonemes (Näätänen et al., 1997, 2007) at this time-window. A way to investigate the long-term effects of exposure to

phonemes native to participants phonology, is to compare phonemes or syllables present in their own phonology and comparing them to sounds that do not appear in their phonology, e.g., comparing brain responses to phonemes that are aspirated vs those without aspiration when the native language of the participant does not have aspirated phonemes. Investigating how native language influences speech perception helps understanding how long-term representations affect perception and can show why learning a second language might be more challenging for an individual depending on their exposure to speech sounds during their early life.

1.2 Speech tracking in adults

Coherence analysis is a measure to study the synchrony between two signals in the frequency domain. Coherence values represent the consistency of phase difference between two signals at any given frequency. As such, it could be a useful tool to explore how the speech signal is tracked at the cortical level (cortico-acoustic coherence, CAC). CAC is used in both speech perception (Ghitza et al., 2013; Molinaro & Lizarazu, 2018; Poeppel, 2014; Poeppel & Asseneo, 2020) and speech production (Bourguignon et al 2020) studies, where the cortical oscillatory activity is compared to the envelope of speech recorded using a microphone.

Auditory information in speech comprises multiple timescales, from phonemes through syllables, words to phrases. Multi-time resolution models of speech processing (Ghitza, 2011; Ghitza & Greenberg, 2009; Poeppel, 2003) suggest that information in speech is processed and integrated in a hierarchical and interdependent way by neural entrainment or phase alignment in the oscillatory networks involved in the auditory cortices with distinct specialization in the left and right auditory areas. This realignment in the oscillations in the auditory system seems to synchronize the ongoing oscillations of large ensembles of neurons to the modulation rates in the stimulus. Low frequency brain oscillations synchronise to the rhythms of linguistic units (Ding et al., 2016; 2017) and higher frequencies appear more sensitive to syntactic and semantic features of speech (Ding et al., 2016). When listening to longer segments of continuous speech, the brain synchronizes these cortical rhythms to track the rhythm of the different linguistic units (Ding et al., 2017) and the linguistic information corresponding with the different timescales is then combined. Importantly, speech tracking has been shown to be increased when speech was intelligible versus unintelligible (Gross et al., 2013.; Molinaro & Lizarazu, 2018; Peelle et al., 2013), thus suggesting that it is affected by higher order processes in addition to acoustic features and their analysis.

Furthermore, processing of speech has been suggested to involve the auditory cortices bilaterally in multiple steps (Poeppel, 2003; Poeppel & Assaneo, 2020). Starting with a bilateral symmetry, the representation of the speech signal branches out depending on the integration window. The auditory areas on the left hemisphere have been suggested to sample information from shorter

integration windows (20-40 ms) (Giraud et al., 2007; Poeppel, 2003,), while the areas on the right hemisphere appear more sensitive to information from longer integration windows (150-200 ms) (Giraud et al., 2007; Poeppel, 2003).

1.3 Development of speech perception and tracking

Developmental changes have been observed in both general auditory processing and speech processing (Uhlhaas et al., 2010; Rióz-Lopez et al., 2020; Wunderlich et al., 2006). Maturation effects have been traditionally investigated using event-related potentials (and their magnetic equivalent ERFs) to short sounds. ERPs recorded to auditory stimuli in infancy and preschool age feature prominent P1 and N2 responses (Hämäläinen et al., 2011; Parvainen et al., 2011). With maturation, the latency and amplitude of these responses decreases and around ages of 8-9 years, the P1 and N2 responses become separated by two new components, an N1 and P2 response as children mature (Albrecht et al., 2000; Čeponien et al., 2002; Cunningham et al., 2000; Kraus et al., 1993; Ponton et al., 2000; Ponton et al., 2002; Takeshita et al., 2002). Furthermore, differences in responses between hemispheres have been found in auditory evoked potentials to speech sounds in the source structure of the response (Parviainen et al., 2011), where responses in the left hemisphere showed a superior-anterior direction of source current, while an opposite pattern was found in the right hemisphere.

Besides event-related responses, another method to investigate the maturation of the different hemispheres is looking at the neuronal oscillations. A hemispheric specialization for the different sizes of the temporal integration windows (25 and 160-300 ms) seems to be present from an early age (Telkemeyer et al., 2009; 2011), suggesting a general processing independent of the language environment. However, presently there are contradictory results in terms of studies investigating the hemispheric differences in speech tracking in children. On the one hand, already a few days after birth hemodynamic responses have revealed rapidly modulated stimuli (25 ms) to be processed bilaterally and processing of slowly modulated stimuli (160-300 ms) to be more right lateralized (Telkemeyer et al., 2009). Similar hemispheric specialization is observed at 6 months, although at 3 months left hemisphere dominance was found for both fast and slow modulations (Telkemeyer et al., 2011). Entrainment to the speech envelope in seven-month-old infants showed larger responses in the left compared to the right hemisphere in the theta band when listening to speech (Kalashnikova et al., 2018). This specialization was absent in younger children (4-7 years of age) (Riós-Lopez et al., 2020), albeit their analysis focused on the delta band. In children between 9 and 13 years of age, correlation was larger between brain responses and the amplitude envelope of sentences in the right than the left hemisphere (Abrams et al., 2008; 2009).

There are methodological differences to take into account when comparing findings that could cause such differences in findings. Some contrasted hemodynamic responses recorded with optical topography and event-related

potentials recorded with EEG to sounds with speech-like spectral structure (Telkemeyer et al., 2009; 2011), calculated the cortical entrainment to speech (Kalashnikova et al., 2018, Riós-Lopez et al., 2020), or used cross-correlation between the speech envelope and cortical activity (Abrams et al., 2008; 2009). Furthermore, the studies were done in different language environments, such as English (Abrams et al., 2008; 2009; Kalashnikova et al., 2018), Basque (Riós-Lopez et al., 2020), and German (Telkemeyer et al., 2009; 2011; although the stimuli used were speech-like sounds in this case) which could have had some effect on the findings.

While ERP responses have been compared to different linguistic units (syllables, words and sentences) in adults (Bonte et al., 2006), differences between speech tracking of words and sentences have not yet been investigated, nor have basic auditory responses been compared to an index of speech tracking at longer stimulus lengths. Furthermore, studies have yet to investigate the developmental changes of these processes via comparison of adult and child groups.

The processing of speech envelope has been shown to be linked to segmentation of speech into elements of syllables or phonemes (Poeppel, 2014), and it has been suggested that development of speech tracking might interact with co-developing cognitive and language-related abilities. Children initially process larger linguistic units, such as words, before developing representations for smaller linguistic units, such as syllables or phonemes (for review: Vihman, 2017). Reading acquisition has been shown to put emphasis on phonemes (Brennan et al., 2013; Popescu & Noiray, 2019; Ziegler & Goswami, 2005) thus suggesting that starting to learn to read would effect speech segmentation abilities. Phoneme-grapheme mapping has also been shown to effect speech processing at both behavioural (Seidenberg & Tannenhaus, 1979; Ziegler & Ferrand, 1998) and brain level (Bonte et al., 2017; Pattamadilok et al., 2004).

1.4 Speech production

Speaking involves the complex operation and coordination of neural networks underlying cognitive, linguistic and motor processes and speakers monitor, evaluate and correct their speech while anticipating listener responses (Hickok 2012; Huode & Chang, 2015). The production of linguistic sound sequences (or speech) requires the activation of articulators or speech organs by the cortical systems. These are highly specialized organs, and can be divided into passive articulators, which remain still during production and active articulators, moving relative to the passive articulators. They are species- and task-specific, allowing for a broad range of phonetic capabilities (Liberman & Whalen, 2000). These intricate articulatory systems are supported by just as complex and specialized cortical systems (Scott, 2005).

Theories of speech production (Hickok, 2012; Indefrey, 2011; Tourville & Guenther, 2011; for review, see Huode & Chang, 2015; Munding et al., 2016) suggest a feedback monitoring system, which monitors the speech output to help

correct errors during production (Hickok, 2012; Huode & Chang, 2015; Tourville & Guenther, 2011). The planned speech sound maps, representing the expected speech output which could be a phoneme, syllable, or even short syllable sequence (Guenther & Vladusich, 2012), are compared to the actual production in the auditory and somatosensory areas. These areas send the resulting comparison error for correction of the production in the ventral premotor cortex and posterior inferior frontal gyrus (Hickok, 2012; Huode & Chang, 2015; Tourville & Guenther, 2011). In parallel, a feedforward system is assumed (Tourville & Guenther, 2011), starting in the premotor and inferior frontal cortex where frequently encountered speech sounds are mapped. These regions then provide input to the bilateral ventral motor cortex to activate the speech articulators and produce speech (Simonyan, 2014; Simonyan & Horwitz, 2011; Tourville & Guenther, 2011).

Similar to speech perception (Bourguignon et al., 2013a; Ding et al., 2016; Gross et al., 2013; Molinaro et al., 2016; Peelle et al., 2013; Zion Golumbic et al., 2013), speech production can also be investigated using coherence analysis although only a few studies have currently done so (e.g., Alexandrou et al., 2018; Bourguignon et al., 2020; Ruspantini et al., 2012,). It poses a challenge to disentangle the different aspects of processing, as unlike speech perception, where the listener would not be moving, production involves the movement of a complex network of articulators and thus would produce movement-related artifacts in the brain recordings when using electroencephalography (EEG) and/or magnetoencephalography (MEG). Furthermore, the execution of overt speech involves contribution from an extended network of cortical areas, such as the somatosensory and motor areas.

Different methods have been established to try and investigate the coupling between speech signal from overt production and brain oscillations. In general, cortico-muscular coherence (CMC) reflects synchronization between electromyographic signals from isometric contractions in the muscles and EEG or MEG activity (Conway et al., 1995; Kilner et al., 2000; van Vliet et al., 2018; Yang et al., 2016). CMC is thought to originate from the primary motor cortex contralateral of the contracted muscle during movements (Fletcher & Wennekers, 2016; Gross et al., 2000; Larsen et al., 2017; Maezawa, et al., 2016), and has been utilized in studying, for example, upper (Liu et al., 2019) and lower limb (Gwin & Ferris, 2012) control. In speech production, it has been used to study speech articulators (Maezawa et al., 2016; Ruspantini et al., 2012). When used to study speech production, corticomuscular coherence looks at the synchrony between the muscle activity of speech articulators and the speaker's brain responses (Maezawa, et al., 2016; Ruspantini et al., 2012). In speech production, CMC has been found to match the rate of articulatory movement, showing highest coherence in the frequency of the individual preferred rate of articulation (2-3 Hz), with largest values found over the sensory motor cortex (Ruspantini et al., 2012).

Another approach to investigate speech production looks at the relationship between the speech envelope of the production (usually recorded

with a microphone) and the brain activity (Bourguignon et al., 2020). Bourguignon and colleagues (2020) found that the sensory-motor cortex appears to monitor auditory and proprioceptive feedback during speech production, and the auditory cortex tracks the envelope of speech at 2-4 Hz, while the parietal operculum showed higher coherence in the 4-8 Hz frequency band. Their results also suggests that coherence in 4-8 Hz receives contribution from motor encoding in preparation of overt speech.

Cortico-kinematic coherence (CKC), meanwhile, reflects the relationship between brain activity and kinematics of movement. Accelerometers are measurement devices recording movement in 3 dimensions, and thus are useful to measure kinematic activity. CKC has been studied in voluntary (Bourguignon et al., 2011; 2012), passive (Piitulainen et al., 2013; 2015) and observed (Bourguignon et al., 2013b) movements. This method has been used to show that, for example, both active and passive repetitive finger movements lead to similar CKC peaking at the movement frequency (Marty et al., 2015a; 2015b; Piitulainen et al., 2015). CKC has been found to reflect proprioceptive feedback to the primary sensory-motor cortex in the case of finger and limb movements (Piitulainen et al., 2013). It has also been used in speech perception (Bourguignon et al., 2013a) to investigate cortical tracking during listening to a person speaking. Importantly, when placed near the vocal cords, accelerometer has been proven to measure vocal cord activity, and that it is especially sensitive to the fundamental frequency (F0) in the voice (Hillman et al., 2006; Lindström et al., 2009; 2010; Orlikoff, 1995). Unlike voice recording using a microphone, accelerometers are not sensitive to environmental noise, and thus provide a clearer measure of speech specific features (Lindström, et al., 2009). Accelerometers have been used successfully in studies investigating the cortical coupling during speech perception (Bourguignon, et al., 2013a), and have the potential to be used in a similar manner when examining speech production.

Due to the technical difficulties inherent in measuring the cortical correlates of overt speech production, most studies examined adults. As such, the development of this process during childhood remains unexplored as of yet. More importantly, comparison of findings from typically developing children and children with speech impediment or production problems could shine light onto their brain-level differences, thus improving understanding of these speech disorders and their cortical correlates.

1.5 Aims of the research

The aims of this dissertation were to study different aspects of the brain-level speech processes from three perspectives:

- to investigate how audio-visual speech perception at syllable-level is affected by the congruency of the visual and audio cues in adults, using MEG

- to investigate the differences in speech tracking of different linguistic units (syllables, words and sentences) between children and adults during speech perception using MEG
- to study the usability of accelerometer to measure speech tracking during overt speech production in adults and combine it with MEG measures

The aim of **Study I** was to investigate the effects of congruency and familiarity of a syllable stimulus on audio-visual speech perception in two groups (native Chinese and native Finnish speakers,) of adult participants. One group had long-term exposure to all stimuli, while for the other group some of the stimuli were not part of their phonology and thus would have no long-term representation for them. In terms of congruency, previous studies found significant differences between brain responses to congruent and incongruent audio-visual speech (Arnal et al., 2009; Callan et al., 2004; Hein et al., 2007; Jones & Callan, 2003; Miller & D'Esposito, 2005; van Atteveldt et al., 2007). We expected both groups to have significantly larger response to incongruent stimuli from 150 ms. Furthermore, long-term memory representation of speech could also affect speech perception. Therefore, we expected to find different brain responses to familiar and unfamiliar stimuli (here, familiarity refers to whether the syllable is part of the listener's native phonology) with larger responses to unfamiliar syllables from an early (100-150 ms) time-window following stimulus-onset if the processing of audio-visual stimuli is facilitated by long-term memory representations. Two groups were compared in their perception of syllables, Finnish and Mandarin Chinese native speakers, and the basis of comparison was whether the syllables were aspirated or not. Finnish does not contain aspirated sounds, while Mandarin Chinese would have both aspirated and unaspirated forms of a syllable - e.g. both /pa/ and /pha/, while in Finnish, only /pa/ would be common and /pha/ would be not present.

In **Study II**, the aim was to investigate the developmental and hemispheric differences in speech tracking of different linguistic units (syllables, words and sentences) between children and adults. We expected to find hemispheric differences in both groups and in both delta and theta frequency bands, specifically larger coherence values in the right than the left hemisphere (Giraud et al., 2007; Luo & Poeppel, 2007; Poeppel & Assaneo 2020). Furthermore, to investigate how the maturation of speech tracking, as reflected in the coherence values, is linked to the maturation of onset responses to syllables, we compared the coherence values for words and sentences and the N1m response amplitudes to syllables. Evoked responses to sounds have been documented to shift from preschool to school age and into adulthood (Albrecht et al., 2000; Ponton et al., 2000). If the N1m amplitude shares underlying maturational mechanisms with coherence values representing speech tracking, similar developmental effects are expected for both measures. Finally, we also investigated the relationship between speech tracking and phonological skills, as the tracking of the speech envelope has been linked to segmentation into syllable and phoneme level elements (Poeppel, 2014) and also to speech intelligibility (Peelle et al., 2013). In

case of Finnish children, they start their schooling at the age of 7 years, thus investigating children's ability to track the speech envelope when they're aged before (e.g. around age 5) and after (older than 7) using a correlational approach could also shine light onto the developmental changes which could be affected by the children starting to learn to read.

The aim of **Study III** was to establish whether it is possible to measure brain-level speech tracking during active speech production of sentences. We compared coherence values using CAC and CKC where participants spoke sentences out loud. The primary question was whether CKC measures coherence above chance level in speech production. A secondary goal was to investigate whether coherence calculated using the speech signal recorded by an accelerometer near the vocal cords (CKC) is comparable to the coherence calculated using the signal recorded using a microphone (CAC). Thirdly, we wanted to assess if these coherence measures would demonstrate the feedback and feedforward system of speech production as proposed by the DIVA model (Tourville & Guenther, 2011). If CAC and CKC both reflect only auditory feedback, we expect no difference between the coherence values. If they also measure either somatosensory feedback or motor cortex activity, we expect coherence to be found in more extended brain regions. Furthermore, we wanted to explore whether the activation patterns reflect similar activity or are actually sensitive to different features of speech production. Finally, we compared hemispheric dominance for both methods, to evaluate whether hemispheric differences previously reported in speech perception are also present during speech production.

2 METHODS

2.1 Participants

In **Study I**, two groups of participants were recruited: adult native Finnish speakers (N=12) and adult native Chinese speakers (N=12), who were studying in Jyväskylä at the time. In **Study II**, two age groups participated: typically developing children (N=34, age range 4.7-9.3 years) and young adults (N=19, age range 20.3-35.2). The adults were studying at the University of Jyväskylä, Finland. All participants were Finnish native speakers. In **Study III**, data from the adult group of **Study II** were used.

All participants (and the parents of the children in **Study II**) gave written informed consent before the experiment. Adult participants were recruited through mailing lists at the University of Jyväskylä. The children participants of the Study II were recruited through the National Registry of Finland. A child-friendly consent form was created for the child participants, and the youngest participants gave also spoken consent after debrief by the research assistants. Participants had the option of stopping the experiment during measurements and to withdraw from the study at any time.

All studies were approved by the ethical committee of the University of Jyväskylä.

2.2 Stimuli and task

2.2.1 Study I

In order to study the effect of familiarity and congruence of speech, audio-visual syllable stimuli were created. Video recordings of the syllables /pa/, /pha/, /ta/, /tha/ and /fa/ were used. To keep participants' attention, the syllable /fa/ was used for a cover task.

Stimuli consisted of video recordings of a male native Mandarin Chinese speaker saying the syllables. Videos were cropped to the mouth area of the speaker. The stimuli could be visual only, where the video was presented without an audio track; audio only, where a still image of the speaker was presented with the audio track; and audio-visual where both video and an audio track was presented at the same time. The audio track could be congruent, where what they saw corresponded to what they heard, or incongruent, where the audio did not match the video. (See Table 1 for description of the stimuli and Figure 1 for a schematic representation of a trial)

For the Finnish participants, "pa" and "ta" were considered familiar stimuli, because they are part of their native phonology while for Chinese participants all four syllables were familiar.

TABLE 1 Descriptions of stimuli used in Study I

Modality	*Target*	*Familiar / Unaspirated*		*Unfamiliar / Aspirated*	
Audio	fa A	pa A	ta A	pha A	tha A
Visual	fa V	pa V	ta V	pha V	tha V
AV congruent	fa V / fa A	pa V / pa A	ta V / ta A	pha V / pha A	tha V / tha A
AV incongruent		pa V / tha A	ta V / pha A	pha V / ta A	tha V / pa V

During the measurement, participants watched short videos of a speaker saying the syllables. When not presented with the videos, participants were asked to focus on a fixation cross which was in the same place as the lips of the speaker in the videos. Videos were presented on a black background. Visual stimuli were presented on a projection screen and sounds were delivered through insert earphones (Lo-Fi auditory stimulation system, Elekta MEGIN Triux) at approximately 70 dB sound pressure level.

During a trial, a blank screen was shown for 500 ms, then a fixation cross for 550 ms. This was followed by a still image of the speaker for 500 ms and then the stimuli, which was 1800 ms long. To keep participants' attention, as a cover task they were asked to indicate seeing and/or hearing the target syllable /fa/ via button press.

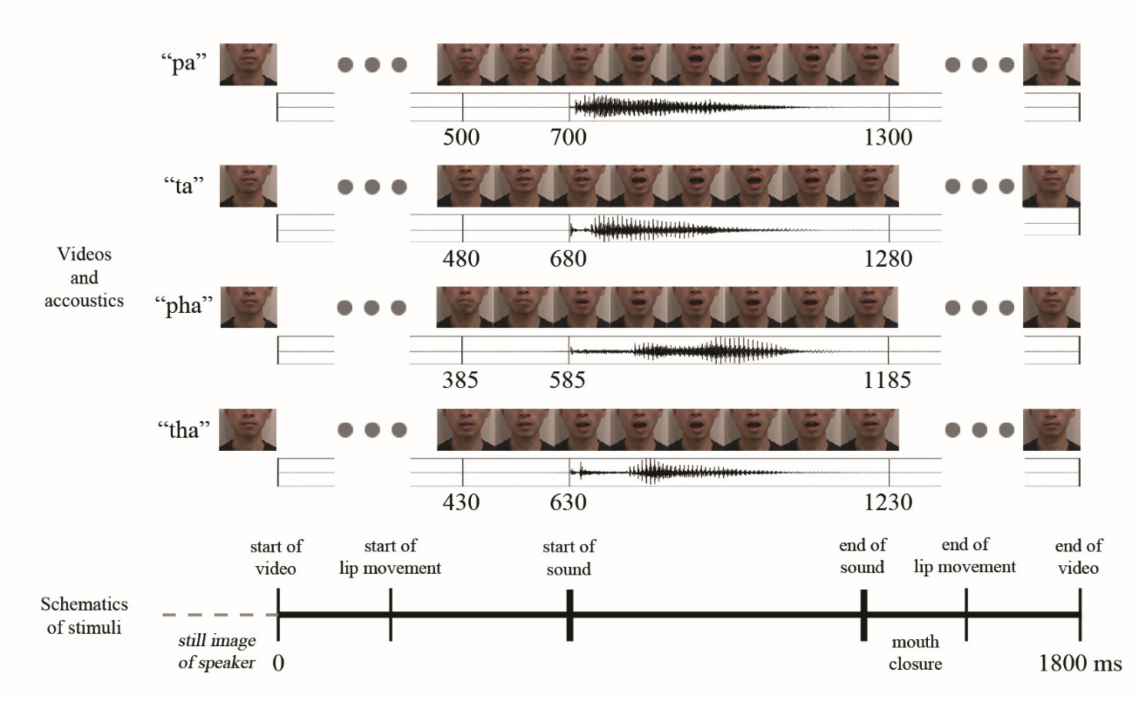


FIGURE 1 Schematic representation of the audio-visual speech perception task used in Study I.

Before the actual experiment, eight practice trials were shown to get the participants familiar with the task. During the measurement 220 stimuli were shown - the cover task had 20 stimuli with randomized modality, and 50 stimuli of each type of modality (audio only, visual only, congruent audio-visual and incongruent audio-visual). Presentation was in pseudo-random order and the same stimuli was never shown in repetition. The experiment was presented in two blocks (110 stimuli each), with a brief break in-between, the duration decided by the participant.

2.2.2 Studies II & III

Stimuli for the speech perception and production paradigm were of three different linguistic units: syllables, words and sentences. Syllables were consonant-vowel pairs, where the consonants could be /p/, /t/ or /k/, while the vowel was /a/ in all cases. Eighteen words beginning with each syllable were selected (54 different words altogether), each word 2-3 syllable long common, everyday nouns. Each sentence was created starting with one of the words (54 sentence altogether), composed of 3-4 words and consisting of a noun followed by the verb 'to be' in the present tense. Sentences were simple, easy to repeat by most speakers. All stimuli were produced by a female, native Finnish speaker as separate tokens. Sampling rate of all auditory stimuli was 44 kHz, recorded in a professional recording studio. The individual segments were cut from the continuous recording using Praat (Boersma & Weenick, 2018).

During MEG recording, a trial began with a fixation cross in the middle of the screen for 500 ms. An exclamation mark appeared for 1000 ms to alert

participants of the coming sound. This was followed by silence with the fixation cross on the screen for 750 ms. The auditory stimuli was then presented through insert earphones, the fixation cross remaining on the screen during the presentation. Next the fixation cross stayed on screen for 750 ms in silence.Then a still image of a parrot appeared on the screen for 1250-4250 ms (duration dependent on the stimuli presented previously) to cue the participants to repeat out loud the stimulus heard previously (See Figure 2).

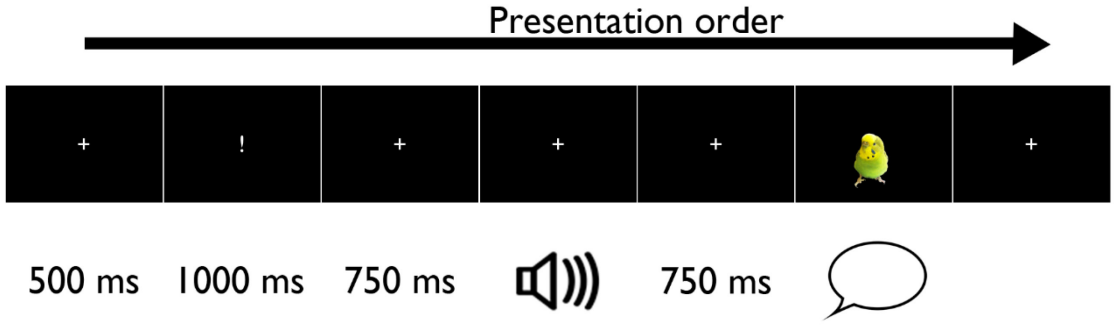


FIGURE 2 Schematic representation of one trial of the experimental paradigm used in Studies II and III. In Study II, data analysis was focused on the time-window during the stimuli presentation, indicated by the picture of the loudspeaker. In Study III, data analysis was focused on the time-window of production, represented here by the speech bubble, when participants were presented with a parrot on the screen.

Participants' task was first to listen to a sound, which could be a syllable, word or a sentence and to repeat what they heard once a visual cue appeared on the screen (a parrot). All visual stimuli were presented on a black background with white standard characters in Times New Roman font and in a font size of 64. The bird stimuli were on average 9 x 15 cm on the projection screen.

A child-friendly narrative was created to keep children's attention and to stimulate their motivation to complete the task. The task was framed as a story where participants were teaching 3 parrots to "speak". Participants were to listen to syllables, words and sentences, then wait for the parrot to 'listen' (when they appeared on the screen) and to keep eye-contact with the parrot to make sure it's paying attention to minimize movement-related artefacts in the recording. Once the parrot appeared, participants were instructed to repeat what they heard at a normal speaking loudness, to be able to record the production during the task as clearly as possible. This ensured children (and adults) were fully engaged in the task. To give the impression that the parrots indeed 'learnt' the stimuli, at the end of every third block, they 'repeated' some of the heard sounds. This was done by raising the pitch of the original stimuli, giving the illusion the parrot made the sounds.

After instructions, participants were given 6 practice trials, with 2 examples of each of stimuli in random order. During the actual measurement 162 stimuli (54 of each type) were presented in 9 blocks, with breaks between each block. The complete task took approximately 30 minutes to complete, including instruction and practice.

2.3 Behavioural measures

2.3.1 Behavioural tasks

In **Study I**, the reaction time and accuracy of responses were collected to the cover task. In **Studies II** and **III**, the accuracy of repetition and the length of repetition was measured from the audio recordings.

2.3.2 Cognitive assessments

No cognitive assessments were collected for the audio-visual speech perception study (**Study I**). For the speech tracking and production study (**Studies II** and **III**) a battery of behavioural tests was conducted assessing children's general cognitive abilities with an emphasis on language-related skills.

To measure participants' visuo-spatial reasoning and vocabulary the WPPSI (Wechsler, 2003a) was used for younger (5-7 years) children, the WISC-IV (Wechsler, 2003b) was used for school age (8-9 years) children and the WAIS-IV (Wechsler, 2008) was used for adults. To measure working memory the digit span subtest was used from the WISC-IV and WAIS-IV in school-age children and adults, respectively. To measure the motor development of the child participants, the oro-motor subtests were used from the Neuropsychological test battery I (NEPSY; Korkman et al., 1998), and the visuo-motor task was used from NEPSY II (Korkman et al., 2008). Phonological processing was measured using the subtest from NEPSY II. Speed of lexical retrieval was measured using the Rapid automatized naming (RAN; Denckla & Rudel, 1976) Object and Letters subtests. Children's memory for sentences was tested using the Sentence Repetition subtest from NEPSY II. Reading skills in older children were measured using the word reading task from the Lukilasse test battery (Häyrinen et al., 1999), a pseudoword reading task adapted to Finnish from TOWRE (Torgesen et al., 1999) and a pseudoword text reading task (Eklund et al., 2015).

2.4 MEG data acquisition

Magnetoencephalography data were collected using Elekta Neuromag TRIUX system (Elekta AB, Stockholm, Sweden) using 102 magnetometer sensors and 102 gradiometer sensor pairs. Participants sat under the MEG helmet in a 68° sitting position, in a magnetically shielded, sound-attenuated room. The experiments were done at the Centre for Interdisciplinary Brain Research at the University of Jyväskylä, Finland.

Head position in relation to the sensor array was continually tracked using 5 digitized head position indicator (HPI) coils, three of which taped to the forehead and one behind each ear. Three anatomic landmark point were used to define the head coordinate system: the left and right preaurical point and the

nasion. The HPI coil positions, the anatomic landmarks and the shape of the participant's head (using >100 points distributed evenly over the scalp) were recorded using the Polhemus tracking systems (Polhemus, Colchester, VT, United States).

Eye-movements during the experiment was tracked using one or two pair(s) of electrodes attached either diagonally near the participants' eyes, lightly below the left eye and slightly above the right eye (**Study I**), or horizontally near the eyes and above and below the right eye (**Studies II** and **III**), and an additional ground electrode was attached to the collarbone.

Stimuli were presented using Presentation software (version 18.1, Neurobehavioral Systems, Inc., Albany, CA, United States) running on a Microsoft Windows computer using a Sound Blaster Audigy RX sound card and NVIDIA Quadro K5200 video card. Stimuli were projected from outside of the measurement room onto a mirror, reflecting the image onto a rear projection screen. Participants were sitting 1 m from the projection screen.

Magnetoencephalography data were collected with a sampling rate of 1000 Hz and an online filter of 0.1-330 Hz. Data were preprocessed using the temporal extension of the signal space separation (tSSS) method with buffers of 30s (Taulu and Kajola, 2005; Taulu et al., 2005) in MaxFilter (Elekta AB) 3.0 (**Study I**) and MaxFilter 2.2 (**Studies II** and **III**) to remove external interference and correct for head movements. Bad channels were automatically excluded and reconstructed by the MaxFilter program. Head position was estimated in 200 ms time-windows and 10 ms steps for movement compensation.

For **Studies II** and **III**, an accelerometer was attached to the right side of the throat to measure vocal cord activity, added as three miscellaneous channels to the brain data. Two pairs of electromyography (EMG) electrodes were attached at the base of the throat to record tongue and vocal cord activity and to right side of the top and bottom lips of the participants to record muscle activity from the lips during production - data from these were not used in any further analysis since the EMG picked up activity from multiple muscles and would have been less specific than the signal from the accelerometer. An optical microphone was also used to record the speech production, recording the sound both as an mp3 file with 44.1 kHz sampling rate and as a miscellaneous channel with the MEG data with the same sampling rate of 1000 Hz as the brain data was recorded.

2.5 MRI data acquisition

For the adults in **all three Studies**, the fsaverage brain template from Freesurfer (RRID: SCR_001847, Martinos Center for Biomedical Imaging, Charlestown, MA, United States) was used.

In **Study II**, the child participants' structural magnetic resonance images (MRI) were acquired in Synlab Jyväskylä. T1-weighted 3D-SE images were collected on a GE 1.5 T MRI scanner (GoldSeal Signa HDxt) with a standard head coil. Scanning parameters were as follows: TR/TE = 540/10 ms with sagittal

orientation, matrix size = 256 x 256, a flip angle of 90 degrees and slice thickness of 1.2 mm. The source reconstruction was based on their own T1 MRIs.

Co-registration was done between the digitized head points and the MRI images for the children, and the brain template with 3-parameter scaling when using the template MRI in case of adults.

2.6 Data analyses

In **Study I**, reaction times and accuracy of responses to the cover task were compared based on modality of the target stimuli. In **Study III**, the accuracy of sentence repetition was assessed using the voice recordings from the optical microphone and only data from correct repetitions were used in the analysis.

MEG data were pre-processed in MNE Python (0.16.2) (Gramfort et al., 2013). Independent component analysis (ICA) using fastICA algorithm (Hyvärinen & Oja, 2000) was applied to remove eye blinks, horizontal eye movements and cardiac artifacts.

In **Study I**, data were low-pass filtered at 35 Hz using a zero-phase FIR filter with a bandwidth of 8.8 Hz. Continuous MEG recording was epoched into 200 ms before and 1800 ms after the onset of the video stimuli in the audio-visual condition. The data were baseline corrected using the 200 ms preceding the onset of the video. The epochs were shortened to 200 ms before and 1000 ms after the onset of sound in the stimuli. Following visual inspection, movement artifacts and electronic jump artifacts were removed and 96.50% of trials were used in the analysis.

Event-related fields were obtained by averaging trials for different conditions separately. Data were resampled to 250 Hz to shorten computation time in the statistical analysis. Five time-windows were defined based on previous literature looking at AV speech perception: 75-125, 150-200, 200-300, 300-400, and 400-600 ms.

Sensor level data analysis was done in FieldTrip toolbox (downloaded 20th of October, 2016; Oostenveld et al., 2011) for MATLAB R20160b (The MathWorks Inc., Natick, MA, 2000) while source-level analyses were run in MNE Python.

At sensor level planar gradiometer data were transformed into combined planar gradients. Combined planar gradients give the vector sum of the two orthogonal sensors at each position. Permutation tests with spatial and temporal clustering based on t-test statistics were carried out on individual averaged ERFs (Maris & Oostenveld, 2007).

Source level analysis was done using the minimum-norm estimate (MNE) method on the event-related fields of the both magnetometers and gradiometers (Hämäläinen & Ilmoniemi, 1994). Depth-weighted L2-minimum-norm estimate (wMNE) was calculated for 4098 current dipoles with free orientation, distributed on the cortical surface of each hemisphere. Dynamic statistical parametric mapping (dSPM) (Dale et al., 2000) was used to noise-normalize the inverse solution, and cluster-based permutation statistics were run on the source waveforms.

In **Study II**, following preprocessing the continuous data was epoched to 100 ms before and 1000 ms after the onset of sound in case of the syllable stimuli, and 100 ms before the onset of and 100 ms after the end of sound in the case of words and sentences. Following visual inspection of trials, 97.82 % of trials were kept of the epochs for children and 99.22% of trials were kept for adults. Data were then low-pass filtered at 45 Hz and baseline corrected using the 100 ms preceding the onset of the stimuli.

Coherence was calculated between the MEG signal and the speech signal for words and sentences, and the event-related fields evoked by the syllable stimuli were examined to investigate possible associations between the two measures.

To calculate the coherence between the envelope of speech signal, the speech stimuli were downsampled to 1000 Hz from 44.1 kHz and then calculated the absolute Hilbert envelope for each stimulus separately. This envelope was then appended to the MEG data set as a 307th channel.

To avoid effects from the onset evoked response, the first 250ms of brain activity were removed. Frequency analysis of the data was done to calculate the cross and power spectra of the trials between 1 and 45 Hz, using a multitaper frequency transformation method, where the maximum trail length was rounded up to the next power of 2, with 3 Hz smoothing. Coherence analysis was then run between the sound envelope and the MEG data.

At sensor level, channels were grouped together by hemispheres, and in the statistical analysis, data from magnetometers were used, averaged together based on hemispheres and separated into two frequency bands (delta: 1-3.5 Hz, theta: 4.5-8 Hz).

To check whether the coherence values were significant at the individual level, 1000 permutations of random coherence were calculated by randomly assigning the sound envelopes to the brain activity of another stimuli and calculating the coherence for them. Then these permuted values were compared to the original coherence value, and was taken as significant if the value was larger than 95% of the permuted values. Each participant had at least one channel with a significant value.

Source level analysis was done using dynamic imaging of coherent sources (DICS; Gross et al, 2001) between 1 and 8 Hz for every 0.5 Hz. The resulting coherence values were then averaged together into the same two frequency bands as at sensor level. Coherence values were then extracted based on the Desikan-Killiany Atlas (Desikan et al., 2006). Two regions of interest were defined: temporal area (including the superior temporal, transverse gyrus and band of superior temporal sulcus areas) and inferior-frontal area (including the pars opercularis, pars orbitalis, pars triangularis and precentral areas).

Event-related fields to syllables were also investigated. Trials were averaged together for each participant separately and the global mean field power was calculated for each group separately. This GMFP was used to identify the time-window of auditory response by automatically finding the peak near 100 ms, using a time-window of +/- 25 ms for each hemisphere and group. Squared values from the temporal channels were averaged together for the

hemispheres separately and then correlated with the coherence values from the respective hemispheres.

The topography of the averages were also visually inspected to confirm they had the correct N1 response pattern or its equivalent in children, since it's been shown, that N1m and P1m could likely occur at similar time-windows, with opposing current directions and likely reflect distinct processes.

Source analysis of the event-related fields were calculated using the minimum-norm estimate (MNE) method (Hämäläianen & Ilmoniemi, 1994), and the power of the source was used in the statistical analyses. Source power waveforms were extracted based on the Desikan-Killiany Atlas (Desikan et al., 2006) and a region of interest was defined in the temporal area (including the superior temporal, transverse gyrus, bank of superior temporal sulcus, postcentral and supramarginal areas) using the same time-windows as the sensor level analysis.

In **Study III**, following preprocessing, the start and end of production were identified for each trial using the signal from the accelerometer and the microphone. The continuous data recording was epoched to 100 ms before start and 100 ms after end of production. Following visual inspection of the data, and the assessment of the production using the voice recordings from the measurements, 99.32% of sentences were used in the data analysis (53.63 out of 54 sentences on average).

In case of the signal from both the accelerometer and microphone, the analysis focused on the F0 speech envelope. The signals were high pass filtered at 50 Hz and low pass filtered at 300 Hz. Frequency analysis was done on each channel to compute the power spectra of each. For accelerometers, the average was calculated for the three channels (for the accelerometer signals) and trials (both accelerometer and the microphone). The F0 peak was detected automatically from the averaged power spectra for the accelerometer and the microphone signals. The original signal was filtered +/- 25 Hz around the F0 peak. The Hilbert envelope was extracted for all channels. Then for the three accelerometer channels, the Euclidian norm was calculated and this was used in the following coherence analyses.

Brain data was filtered between 1 and 45 Hz, and trials were epoched to 100 ms before the start of production and 100 ms following end of production. At sensor level, data was downsampled to 200 Hz sampling rate to reduce the computational load for permutation calculations, while at source level the sampling rate was kept at 1000 Hz. The first 350 ms of the epochs (100 ms preceding the start of production + first 250 ms of the production) were removed to avoid contamination in the coherence calculations from evoked responses.

Coherence was calculated between both the envelopes from the accelerometer (cortico-kinematic coherence, CKC) and from the microphone (cortico-acoustic coherence, CAC) and the brain responses.

At sensor level, the power and cross-spectra of the trails was calculated using FieldTrip's frequency analysis function (ft_freqanalysis) with multitaper frequency transformation method, rounding up the maximum trial length to the

next power of 2 (cfg.pad = nextpow2), between 1 and 45 Hz and with a 3 Hz smoothing. This was followed by coherence analysis using FieldTrip's ft_connectivityanalysis function, to calculate coherence between the speech envelope and the brain data.

The same permutation tests were run for both CKC and CAC values as **Study II**. To confirm the findings from the permutation tests, a theoretical 95% confidence limit (Halliday et al., 1995) was also calculated using the formula 1-(1-0.95)^1/(L-1), where L represents the number of independent FFT-windows, which in this case was the number of trials included in the analysis for each participant. These values were used as a cut-off point to check that the coherence values were above chance level at the individual level.

At source level coherence analysis was done for frequencies between 1.5 and 10 Hz with a resolution of 0.5 Hz to investigate which frequencies show highest coherence between the speech envelope of speech production and brain activity.

Regions-of-interests were used for the analyses defined by visual inspection of the regions using the Desikan-Killiany atlas (Desikan et al., 2006) with largest coherence values regardless of hemisphere, resulting in 5 regions: temporal, motor, supplementary motor area, inferior frontal area and somatosensory area (See Figure 3).

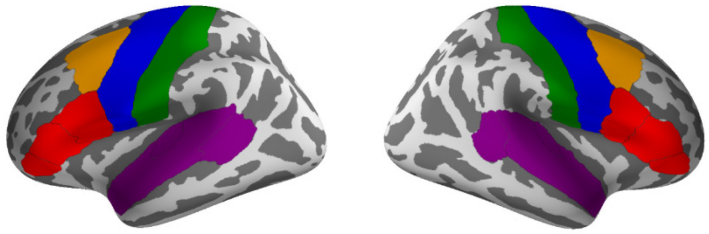


FIGURE 3 Regions of interests used in Studies II and III. Temporal area – purple, Inferior-frontal area – red, Motor area – green, Somatosensory area – blue, Supplementary motor area – yellow. In Study II, only the temporal area and inferior-frontal areas were examined.

2.7 Statistical analyses

In **Study I**, the main interest was in investigating how familiarity, in this case whether the syllable seen was part of the participant's phonology, and congruence affect the brain responses to syllables. Effect of congruency and familiarity of the stimuli on the brain responses were examined using permutation tests, focusing on the interaction of Stimulus by Native language by comparing the difference waves between the two groups. If no significant results

were found, Stimulus main effects were investigated. To investigate effect of congruence, the brain responses to congruent and incongruent audio-visual stimuli were compared. To investigate effects of familiarity of the stimuli, responses to congruent unaspirated and congruent aspirated stimuli were compared.

As a secondary question, we wanted to investigate how the modality of the stimuli affected reaction time and accuracy of responses in the cover task. To do so, accuracy and response times were compared using Target type (Audio only, Visual only, Audio-Visual) by Native language (Finnish, Chinese) in a repeated mixed ANOVA.

In **Study II**, the interest was to investigate the developmental differences in the brain's ability to track the speech envelope. To do this, a Type (word, sentence) x Hemisphere (left, right) x Group (children, adults) repeated measures mixed ANOVA was used to inspect possible age, hemisphere and stimulus type effects on the coherence values at both sensor and source levels. Significant interactions were further examined using independent samples t-tests, and paired samples t-tests where groups were part of the interaction. Further, Pearson correlations coefficients were calculated between the coherence values at source level and children's age in years rounded to months, and also between coherence values at source level and performance on three behavioural tests (RAN: objects subtest, NEPSY: Phonological processing subtest, Sentence repetition subtest).

Similarly, a Hemisphere (left, right) x Group (children, adults) repeated measures mixed ANOVA was used to compare the brain responses around N1m to syllables. Pearson correlation coefficients were calculated to inspect the relationship between the average amplitudes of the auditory responses and the coherence values.

Alpha level was 0.05 and false discovery rate correction for multiple comparisons was calculated for each analysis.

In **Study III**, the main interest was in verifying whether the signal envelope from speech recorded using accelerometer can be used in coherence measures looking at overt speech production. To compare coherence values at the sensor level, permutation tests based on t-test statistics were carried out on the CKC and CAC values (Maris & Oostenveld, 2007). Bonferroni correction of the p-value was used for correction for multiple comparisons. At source-level, T-tests were used between coherence measured using a microphone and an accelerometer over 8196 source points to see if and where the values differ between the two measurement methods. Then, after extraction of regions in MNE, Method (Accelerometer, Microphone) x Hemisphere (left, right) repeated measures ANOVA was used to investigate the relationship between the coherence values found using the two methods and the two brain hemispheres in five Regions of interests (ROI).

Alpha level was 0.05 and false discovery rate correction for multiple comparisons was calculated for the ANOVA results.

3 RESULTS

3.1 Study I – Audio-visual speech perception in adults

At sensor level, significant congruency effects were found only in the fourth (300-400ms) time-window, significant differences were found in the left frontal area ($p = 0.037$) and in the right temporal area ($p = 0.005$) across groups, showing larger responses to incongruent stimuli than to congruent stimuli. No significant familiarity effects were found in any of the time-windows.

At source level, similarly, significant congruency effect was found only in the fourth (300-400 ms) time-window. The difference was found to be significant ($p= 0.039$) encompassing the right temporal-parietal and medial areas across groups, showing larger responses to incongruent than congruent stimuli. No significant familiarity effect were found in any of the time-windows. Figure 4 shows both sensor and source level topographic maps and dynamic statistical parametric maps and waveforms.

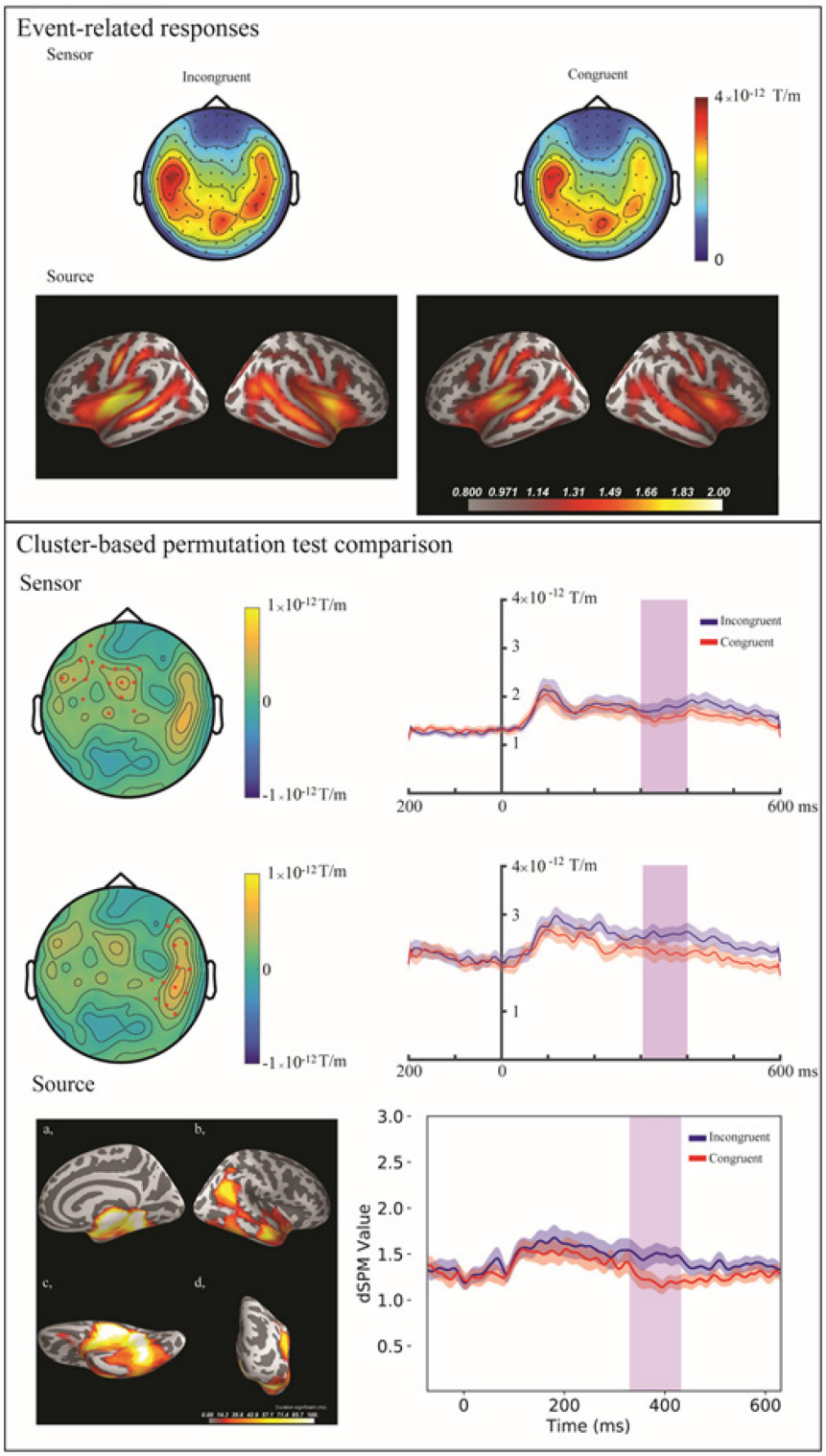


FIGURE 4 Sensor and source level results from Study I. Top: Sensor and source level topography of event-related responses to the incongruent and congruent stimuli averaged together. Bottom: The results of the cluster-based permutation comparisons at sensor and source level. On the left: the topography of significant differences found. On the right: the waveforms of the responses to incongruent and congruent stimuli in the clusters. Purple boxes highlight the time-window of the difference.

At the behavioural level, accuracy was high for both groups (Chinese, 98.35%, Finnish 97.88%) and no significant interaction or main effect was found after the 3 (Target type: Audio only, Visual only, Audio-visual) by 2 (Group: Finnish, Chinese) repeated measures mixed ANOVA. Reaction times showed a Target type main effect ($F(1.954, 42.985) = 6.338$, $p=0.004$, partial eta squared = 0.224). Post hoc tests revealed that audio only targets had longer reaction times than visual only stimuli ($t(23) = 2.943$, $p = 0.007$) or audio-visual stimuli ($t(23) = 3.518$, $p= 0.002$).

3.2 Study II – Speech tracking in children and adults

At the behaviour level, correct repetition was on average 88.41% for children and 97.86% for adults.

Adults had larger coherence values than children in both delta and theta bands, and participants had larger coherence values to words than sentences, as indicated by the Group (children and adults) and Type of stimulus (words and sentences) main effects. In the delta band, adults had larger coherence values in the left hemisphere compared to children, as well as larger coherence values in the left hemisphere than the right, as the Hemisphere by Group interaction was also significant (Hemisphere x Group interaction) (see Table 2 and Figure 5 for significant effects and interactions).

In terms of evoked responses, no significant differences were found between the groups or hemispheres, nor any significant correlation between the evoked responses or the coherence values in the corresponding hemispheres in either frequency bands.

TABLE 2 Significant main effects and interactions from the sensor level ANOVA of Study II

Frequency band		*df*	*F value*	*p value*	*partial eta squared*
Delta	Type	1,51	227.754	<0.001	0.817
	Hemisphere x Group	1,51	5.822	0.019	0.102
Theta	Type	1,51	239.307	0.000	0.836
	Group	1,51	14.089	0.000	0.216

At source level, when looking at coherence between brain responses and the stimulus envelope of words and sentences, a significant main effects of Stimulus type and Group were found in both delta and theta frequencies and in both Regions of Interest. Larger values for words than sentences were found, and adults had larger coherence values than children. Furthermore, in the temporal region, adults had significantly larger coherence values for both words and sentences compared to children, and also the adult group itself had larger

coherence values for words than sentences (Table 3 and Figure 5 for significant effects and interactions). When looking at the evoked responses, no significant differences between groups or hemispheres were found, nor any significant correlations between the evoked responses and the coherence values.

TABLE 3 Significant main effects and interactions from the source level ANOVA of Study II

Frequency band	*ROI*			*df*	*F value*	*p value*	*partial eta squared*
Delta	Temporal	Type x Group		1.51	6.519	0.014	0.113
	Inferior-Frontal	Type		1,51	9.143	0.004	0.152
		Group		1,51	13.476	0.001	0.209
Theta	Temporal	Type		1,51	44.799	0.000	0.468
		Group		1,51	6.849	0.012	0.118
	Inferior-Frontal	Type		1,51	50.638	0.000	0.498
		Group		1,51	14.688	0.000	0.224

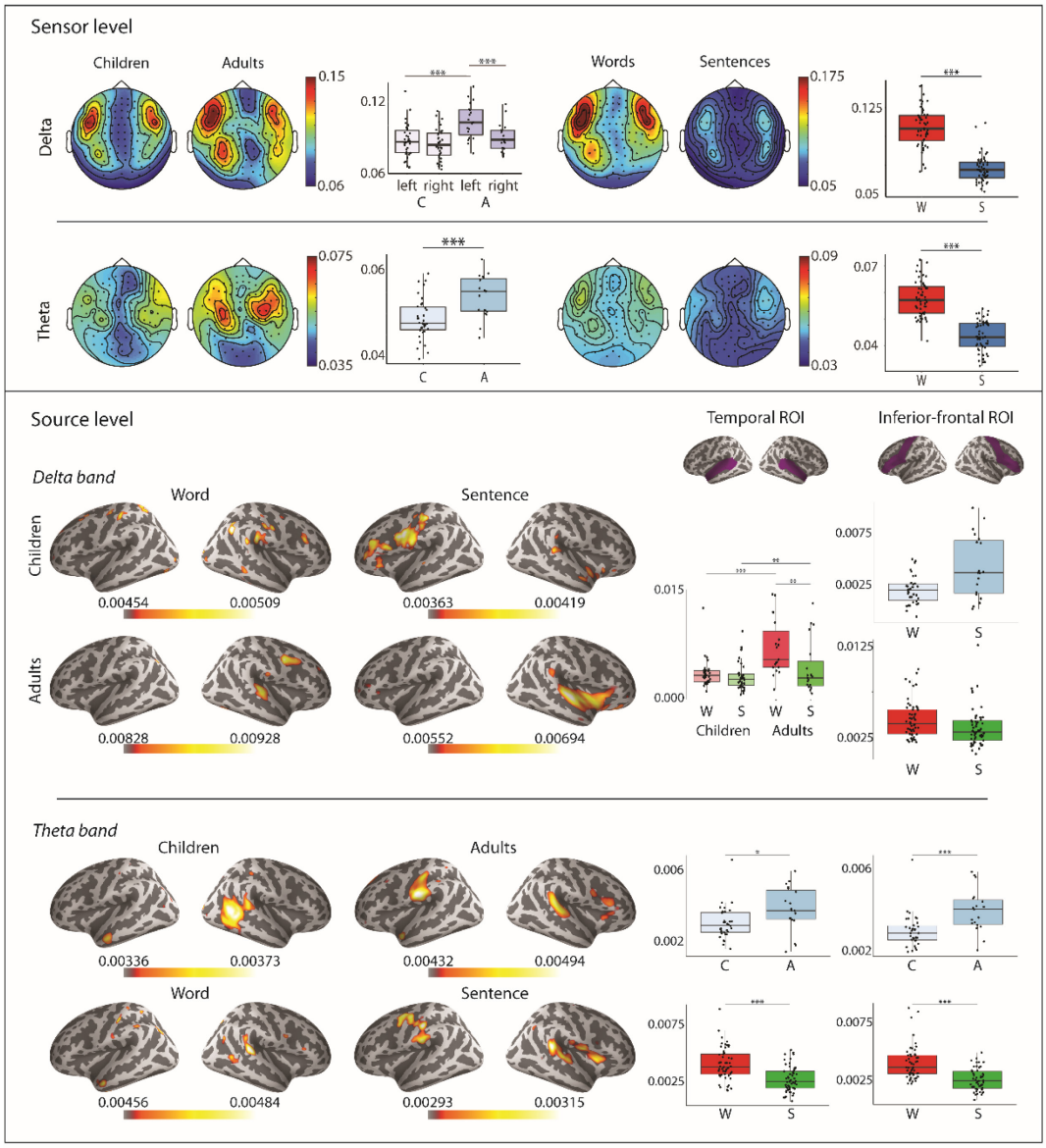


FIGURE 5 Sensor and source level results from Study II. Top: Sensor level results of significant Group by Hemisphere, Group and Type differences in the delta and theta frequency bands. The box plots show the distribution of values. Bottom: Source level results of significant Group by Type, Group and Type differences in the delta and theta frequency bands. On the right, the Regions of interest are highlighted in purple and the box plots show the distribution of values.

Finally, to investigate the relationship between performance on behavioural measures and the brain's ability to track the speech envelope as represented by the coherence values, the coherence values and the behavioural scores for Rapid Automatized Naming (RAN), phonological processing and sentence repetition were correlated. A negative correlation was found between the coherence values and participants' RAN scores only when not controlled for age. There was no correlation between coherence values in either frequency bands or regions of interests and performance on the phonological processing or the repetition of sentences tasks.

3.3 Study III – Speech production in adults – comparison of CKC and CAC

At the behavioural level, correct production was on average 99.32% (53.63 out of 54 on average).

At sensor level, the coherence calculated between the signal from the accelerometer and brain activity (CKC) and coherence between the speech signal recorded with microphone and brain activity (CAC) both peaked near 4.5-5 Hz. While there was a large individual variability in the pattern to the coherence, there were higher values in the temporal and parietal areas in the grand average (Figure 6).

The t-test based permutations revealed significant differences between CKC and CAC for all frequencies investigated, however after correcting for multiple comparisons, the difference was only significant at 1.5, 7 and 7.5 Hz (Figure 6).

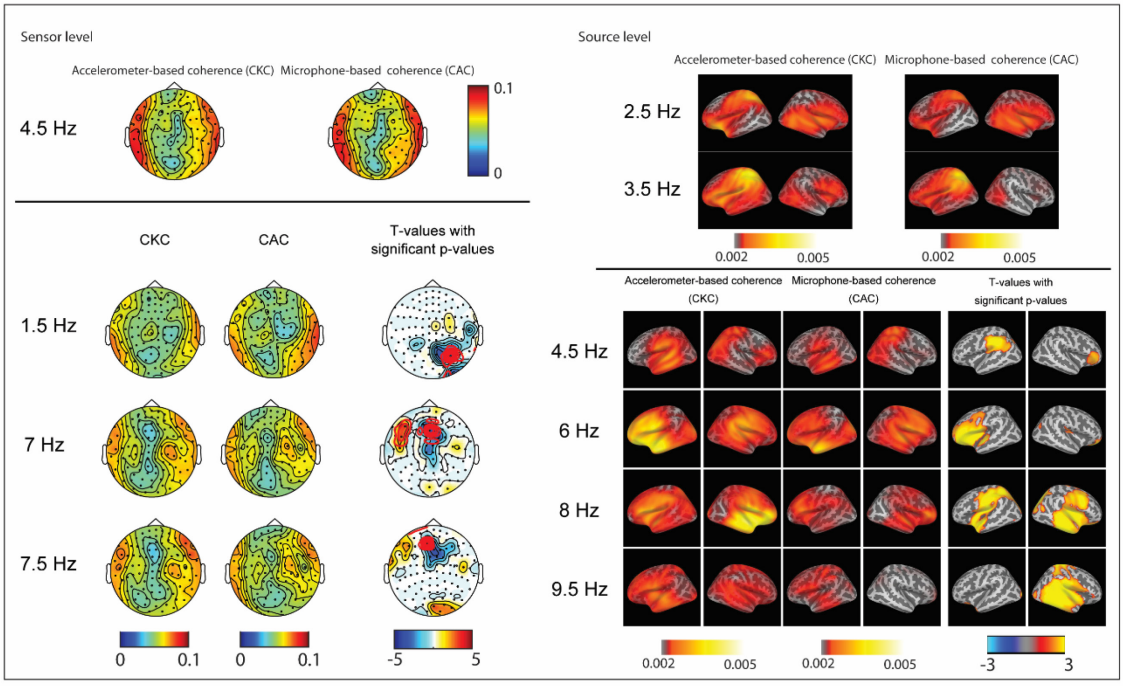


FIGURE 6 Sensor and source level results from Study III. Top: Topography of the group average coherence found in the 4.5 Hz frequency at sensor level (left) and 2.5 and 3.5 Hz frequency at source level (right). Warmer colours represent higher coherence values. Bottom: Sensor-level plots of grand average coherence values and significant differences found using permutation statistics at 1.5, 7 and 7.5 Hz (left). Source plots of 4.5, 6, 8 and 9.5 Hz frequencies, and the t-value distributions after comparing CKC and CAC at group level (right). In the grand average plots warmer colours represent higher coherence values. In the plots from statistical comparisons, orange colour represents significant t-test result where CKC was larger, blue colour represents significant t-test result where CAC was larger.

At source-level, both CKC and CAC had peaks for both at around 2.5-3.53 Hz, as well as 6, 8 and 9.5 Hz. Coherence values appeared larger for CKC than CAC with larger distribution of difference found at 4.5, 6, 8 and 9.5 Hz.

No significant Hemisphere main effect was found after FDR correction in any of the regions of interest.

4 DISCUSSION

4.1 Audio-visual speech perception – syllable level

Study I looked at audio-visual speech perception in 2 groups of participants with differing phonology (Chinese and Finnish), to see whether perception of simple speech stimuli such as syllables is affected by the familiarity of the sounds, i.e. if they are part of the listeners phonology, and how this interacts with the congruence of the presented stimuli, i.e. there is a match or mismatch in what they see and what they hear. Chinese participants would have long term-representations of all the syllables shown (/pa/,/ta/,/pha/ and /tha/) while Finnish participants would only have such representations for two syllables (/pa/ and /ta/), making the other two unfamiliar. Despite expectations, significant differences were found only in the congruency comparisons across the groups in the 300-400 ms time-window at both sensor and source level following stimuli presentation. Familiarity of the stimuli did not produce significantly different brain responses in any of the time-windows.

The significant difference between brain responses to matching and mismatching audio-visual speech stimuli in the 300-400 ms time-window was present bilaterally at sensor level and in the right hemisphere at source level, indicating that participants detected the incongruency in the stimuli. The time-window matches earlier findings (Arnal et al., 2009, Baart et al., 2014), but the effect was found in different brain areas from earlier findings, where they found a more left lateralized emphasis when using more complex combinations of stimuli (Arnal et al., 2009), and left front-temporal distribution when comparing speech sound-symbol combinations (Xu et al., 2019). The difference in localization of the congruency effect shows how the task and stimuli used affects the response to congruent and incongruent stimuli. In **Study I**, the contrast was four syllables with specific combinations to create the incongruent stimuli (the auditory track was always with the different consonant and aspiration), while Arnal and colleagues (2009) used six syllables with multiple combinations, thus making the distinction between congruent and incongruent stimuli harder. The

direction of difference, i.e. larger activation for incongruent audio-visual stimuli also follows some findings of earlier studies (Arnal et al., 2009; Xu et al., 2019).

The lack of early congruency effects in **Study I** could be attributed to the contrast used in our study. Early responses have been indicated to be sensitive to difference in modalities and differences were related to suppression effects when brain response to visual only stimuli was subtracted from brain response to audio-visual stimuli and was then contrasted with brain response audio only stimuli (Stekelenburg & Vroomen, 2007) or brain response to audio-visual stimuli was compared to brain response to unimodal (audio only and visual only) stimuli (van Wassenhove et al., 2005). Unlike these studies, **in Study I**, the comparison was between brain responses to congruent and incongruent stimuli, and as such, significant differences in the responses wouldn't necessarily be found in the early time-windows following stimulus onset.

Familiarity of stimuli also did not result in significant differences in brain responses in Finnish participants. It's important to note brain responses to the familiar and unfamiliar stimuli were contrasted directly, where the number of presentation of stimuli was equalized, unlike earlier studies, where the language-specific difference were examined using an auditory oddball paradigm (Näätänen et al., 1997; Winkler et al., 1999). It is also possible that due to the active attention to the syllables, unlike in mismatch paradigms, where stimuli are observed passively, diminished the differences between the responses to familiar and unfamiliar stimuli. Finally, in the current study, the difference in familiarity was in whether the syllables were aspirated or unaspirated. Perhaps the use of a better contrast of syllables, for example using syllables that are completely new to the listeners might have increased the differences in responses.

4.2 Speech tracking during speech perception – word and sentence level

Study II investigated possible differences between children and adults in their overall brain activity and specifically in their left and right auditory cortex activity while they were listening to various speech units (from syllables to sentences). In particular, cortical level tracking of words and sentences was examined and whether hemispheric differences are present in coherence measures across development. Furthermore, how the ability to track the speech envelope is related to the processing of shorter stimuli, such as syllables at both sensor and source level was investigated.

An overall improvement with age in the brain's ability to track speech was found, reflected in an increase in coherence values in both the delta and theta bands and the topography of coherence showed a clear pattern of auditory cortex activity at sensor level. Increase in coherence values here is assumed to reflect increased precision in tracking speech between childhood and adulthood. Although, based on the lack of correlation between age and coherence values

within the child group, this relationship is not linear. Continued exposure to speech could hone the auditory system's bottom-up pathway (Kuhl, 2000; Ponton et al., 2000) making speech tracking more precise. At the same time, this continuous exposure could also improve the brain's speech perception abilities in a top-down manner as well (Kuhl, 2000), as the input could shape the long-term memory representations thus affecting the precision of speech tracking.

Furthermore, at sensor level, there was an interaction between age and hemisphere in the delta band. It is important to note, however, that despite expectations, no consistent hemispheric differences were found. At sensor level, adults had larger coherence values than children, and in adults coherence values were larger in the left than right hemisphere, but none of the hemispheric comparisons remained significant at source level after correction for false discovery rate. The sensor level findings were surprising in themselves, as the Asymmetric Sampling in Time (AST, Poeppel, 2003) would suggest that right auditory areas would sample information from longer (150-200 ms) integration windows, and thus would have larger oscillatory activity in the theta band in this hemisphere (Luo & Poeppel, 2007). Interestingly, the difference, while not significant after FDR correction, was in the expected direction at source level. A potential reason for this difference between the sensor and source level results could be the channels selected at sensor level or brain regions inspected at source level. At sensor level, all channels were used in the comparison, while source level comparisons were restricted to specific regions of interest, with a focus on temporal regions in particular.

The amplitudes of evoked responses likely represent a more general maturation of the auditory and speech perception system (Ponton et al., 2000; 2002), while coherence is possibly related to the comprehension of speech (Luo & Poeppel, 2007; Peelle et al., 2013). However, it could be assumed that the general maturation underpins the development of speech comprehension. Development of the coherence values and the evoked responses were compared in two ways. First, Global Mean Field Power (GMFP) was used to define the time-window of the evoked response, which were then compared between children and adults and also hemispheres, and no significant differences were found. The time-window based on GMFP is assumed to include both P1m and N1m responses, which likely represent different processes. P1 response has been shown to shift during development to earlier latencies and decrease in amplitude, while the N1 response emerges around early school years (Albrecht et al., 2000; Ponton et al., 2000). P1m of younger children and N1m of older children and adults show similarities in timing (Parviainen et al., 2019), thus complicating purely GMFP-based interpretations. As a next step, the spatial patterns and timings of responses were checked in the left and right hemisphere for each individual. No correlation was found between the first prominent evoked field (now classified as N1m or P1m specifically) and the coherence values in the delta or the theta frequencies. This suggests that the two do not share robust developmental mechanisms and reflect different speech perception processes, as ERFs have been shown to be sensitive to the physical features of sounds

(Näätänen & Picton, 1986; Näätänen et al., 1997) and speech tracking appears to be linked more to attention and speech intelligibility (Peelle et al., 2013).

Finally, participants' coherence values were significantly larger for words than sentences, which was opposite of expectations in as much as the assumption was that longer speech envelope might be easier to synchronize to. However, the significant difference was most likely due to the way coherence was calculated in general rather than difference between the linguistic units – i.e. shorter trial times resulted in higher coherence. This highlights the fact that coherence measures need to be used with careful considerations of trial lengths and comparison between different levels of speech need to be done taking this difference into account.

To summarize, the brain's ability to track speech, even at word and sentence level, appears to increase with age, indicated by the significantly larger coherence values that adults had compared to children, however it did not show consistently significant differences between the two hemispheres, which was surprising. Furthermore, this ability seems to develop independently from auditory responses to syllables, as no correlation was found between the ERPs amplitude and the coherence values nor did it interact with cognitive and language-related abilities.

4.3 Speech tracking during overt speech production

Study III looked at speech tracking during overt speech production in adults. Coherence between speech envelopes recorded with a microphone (CAC) and with an accelerometer (CKC) was calculated and compared. Coherence peaked around 4.5-5 Hz at sensor level for both methods, while multiple peaks were found at source level depending on the cortical region, starting around 5 Hz. Topography at sensor level showed coherence near the temporal and parietal channels, while at source level, coherence was present near the auditory and sensory-motor areas. Furthermore, significant differences were found between the two methods at group level, suggesting that accelerometer-based coherence taps into different aspects of speech track during production than microphone-based coherence. Inspection of individual participants' coherence patterns (following the peaks found at both sensor and source level) showed rather diverse patterns.

The peak around 3-5 Hz in the auditory and sensory-motor areas is consistent with earlier studies looking at speaking rate peaks around 4 Hz (Poeppel, 2003; Poeppel & Assaneo, 2020), while the higher frequency peaks in the source level could reflect synchronization during production to the phonemic rate (Poeppel, 2003; Poeppel & Assaneo, 2020) or alternatively, the first harmonic of frequency of syllable production rate (Ruspantini et al., 2012). At source level, a further peaks were found at 6 and 8 Hz, however more specific experimental manipulations are needed to examine the functional role of this higher frequency peak, which the current study was not designed to do.

While the frequency of the peak coherence matches with the Asymmetric Sampling in Time theory (AST, Poeppel, 2003), no significant hemispheric differences were found. However, it is important to keep in mind that the AST was built for speech perception processes, while in **Study III**, speech production was investigated. As such, differences from the expected response patterns based on AST are not unexpected, since both feedback and feedforward systems (DIVA, Tourville & Guenther, 2011) are engaged during speech production.

In the DIVA model (Tourville & Guenther, 2011) the feedforward system is proposed to include the premotor and inferior frontal regions and the motor areas, while the feedback system involves the auditory and somatosensory areas. Both CKC and CAC were found in these areas, peaking at similar frequencies. Coherence values were overall larger for the accelerometer-based CKC at source level.

In general, the results of this study show that the speech signal recorded using accelerometer during overt speech results in above chance level coherence, and is comparable to measures using a microphone as the distribution of coherence showed similarities. The results suggest that both CKC and CAC are useful methods to measure coherence between the speech envelope and brain responses, and thus could be used in future to successfully investigate overt speech production. The two methods combined could be used for more thorough artifact reduction in the future, since CKC is more susceptible to pick up movement or body artifact from the speaker, while the microphone would pick up more environmental noise.

4.4 General discussion

The aims of this dissertation were to investigate speech perception at syllable, word and sentence level, to examine developmental effects in the brain's ability to track different units of speech, and to examine a new method of investigating overt speech production in adults.

In **Studies I** and **II**, evoked responses to syllables were examined, although in different ways. In **Study I**, the focus was to investigate differences in brain responses to congruent and incongruent audio-visual stimuli while in **Study II** the responses to syllables were used in comparison with coherence values to see if the two share any developmental mechanisms. In **Study II**, the event-related fields in response to audio-only syllables showed that the responses were focused in the temporal areas, but no significant difference was found between the hemispheres in either sensor or source level. The timing and location of this response is in line with previous studies, where they found activation in the sensory areas around 100 ms following onset of stimulus (Möttönen, Schürmann & Sams, 2004; Sams et al., 1991), which has been suggested to reflect processing of acoustic features in this brain area (Salmelin, 2007). In **Study I**, visual inspection of the responses in a later time-window (300-400 ms) show that in case of con-

gruent audio-visual stimuli, there is a more left lateralized response in the temporal areas both at sensor and source level, although no statistical comparison was done between the responses in the hemispheres. This later time-window has been suggested to reflect processing of higher-order analysis than general acoustic features, such as congruence, i.e. whether the incoming audio and visual information are matching or mismatching (Arnal et al., 2009).

In **Studies II** and **III** coherence measures were used to investigate speech tracking during speech perception and overt speech production, looking at the synchrony between the envelope of the speech signal and cortical activity. Overall, coherence values were larger during perception than production at sensor level in the delta band, suggesting that participants' were better able to track speech when listening than when they were actively producing it, although in the theta band the values were in the same range. During perception, the coherence values have been found to be generally larger in the delta (1-3.5 Hz) band than theta (4.5-8 Hz), while during production, the opposite was found, with larger coherence values in higher frequencies.

The DIVA model (Tourville & Guenther, 2011) postulates that during speech production there are two systems working in parallel to ensure successful and correct speech, a feedforward and a feedback system. Both systems begin in the inferior frontal gyrus, where representations of the expected speech output are mapped. From here, the two systems diverge, with the feedforward system engaging the bilateral ventral motor cortex to activate the speech articulators and produce speech (Simonyan, 2014; Simonyan & Horwitz, 2011; Tourville & Guenther, 2011). Meanwhile, the feedback system compares the produced speech to the expected output in the auditory and somatosensory areas and sends any encountered errors for correction in the ventral premotor cortex and the posterior frontal gyrus (Hickok, 2012; Tourville & Guenther, 2011). Important to note that this model maps the sequence of speech production in the temporal dimension, and does not look at how these might be represented in the frequency domain. Nonetheless, the results of **Study III** show activation in all areas involved in both the feedforward and feedback systems, with activity focused in the frontal and temporal areas in lower frequencies (3-5 Hz) and a more widespread coherence pattern in the higher (7-10 Hz) frequencies. In the future, how coherence in different frequencies are related to each other should be investigated, however this was not the focus of **Study III**. Furthermore, in **Study II**, where speech perception of longer (words and sentences) speech stimuli was investigated with coherence measures, the temporal areas showed increased values in both the delta and theta bands, suggesting the engagement of the feedback system at least. In the scope of **Study II**, the motor areas were not examined directly, but the grand averages suggest less engagement in those areas. This makes sense, as during passive perception, the listener does not produce the sounds themselves and thus would not need to correct production. Overall, the patterns of coherence were similar between perception and production in the temporal and parietal areas, while coherence was also found in the motor and somatosensory areas for production.

The Asymmetric Sampling in Time theory (Poeppel, 2003) assumes a symmetric representation at the start of speech perception, but then the two hemispheres diverge and "prefer" to extract information from different integration windows. The theory suggests that the left auditory areas extract from short (20-50 ms) temporal integration windows and the right hemisphere extracts information from longer (150-250 ms) integration windows, which correspond with gamma band and theta band respectively in cortical oscillations. As such, hemispheric differences are expected to found in the theta and gamma bands during speech perception. In **Study II**, no hemispheric differences were found in the theta band (4.5-8 Hz) in the temporal areas, while the gamma band was not investigated at all. Interestingly, a significant hemispheric difference was found at sensor level in the delta (1-3.5 Hz) band, but with a larger coherence in the left than right hemisphere, which is opposite of the expectations based on the AST theory. Important to remember, however, that at sensor level, activity from the whole hemisphere were compared instead of focusing on the temporal areas. In source space, the right temporal areas indeed had larger coherence than the left in the delta band, which agrees with the AST theory, but this result did not remain significant following FDR correction. In **Study III**, when investigating speech production, the temporal areas showed a tendency for larger coherence in the left than the right areas in the lower (3-3.5 Hz) frequencies, as did the inferior-frontal, motor and somatosensory areas at 3.5 Hz. But none of these differences were robust enough to be significant following FDR correction.

The dual stream of speech processing theory (Hickok & Poeppel, 2015) suggests that speech perception involves two streams of processing in the brain, a ventral stream and a dorsal stream. The ventral stream is tasked with mapping sound to meaning and thus is involved in speech perception more, while the dorsal stream is responsible for mapping sound to action and is more involved in preparation for speech production. The ventral stream is suggested to encompass the superior temporal gyrus and sulcus bilaterally, while the dorsal stream involves the posterior planum temporale region and the posterior frontal lobe. During speech perception, an activation starts in the dorsal superior temporal gyrus and sulcus, activating the phonological representations of speech sounds mapped there. The information is then sent in two directions. In the ventral stream, the information is processed in the middle temporal gyrus and the posterior-inferior-temporal sulcus where correspondences between phonological and conceptual information are stored, followed by activity in the anterior-middle-temporal gyrus and inferior-temporal sulcus. In the dorsal stream, the information is sent to the Sylvian parietal-temporal junction and the articulatory network is involved in the inferior-frontal gyrus, the premotor and anterior insula on the left hemisphere. Similar to the DIVA model, this theory represents speech perception as a temporal sequence of activations, while both **Studies II** and **III** focused on coherence between the speech envelope and brain activity and thus was looking at the frequency domain rather than the temporal domain. Nonetheless, the findings in **Studies II** and **III** follow the suggestions of the Dual stream model, as we found coherence in the temporal areas, as well as

the frontal, premotor and motor areas, although in our comparisons that hemispheric differences were not present as strongly, with significant difference found only at sensor level during perception, and in the delta band in particular.

Overall, this dissertation gives insight into speech perception of different linguistic units, showcases developmental changes in speech tracking and finally shows that speech tracking can be investigated during overt speech production and introduces using a new method to do so.

4.5 Limitations

In **Study I**, the task and stimuli used could have influenced the differences found. We used a passive cross-modal task, which could have affected our null results in terms of the familiarity comparison. Previous studies investigating familiarity of speech stimuli used auditory oddball paradigms (Nätäänen et al., 1997, Winkler et al., 1999) to examine participants' ability to detect unfamiliar phonemes. In our case, presenting each type of stimuli in equal numbers allowed us to examine obligatory responses without the influence of other processes. However, the fact that we did not find significant difference in brain responses from a passive task suggests the need for an identification task that engages the long-term representations of the phonemes. Furthermore, our comparison focused on the evoked brain activity following the presented stimuli. It is possible that the effect of familiarity of the syllable might not be phase-locked to the stimuli, in which case the difference in brain responses might not be detectable with comparison of evoked responses. The contrasts used perhaps also had an effect and aspiration of a phoneme might not be enough of a difference to cause robust enough differences in brain responses.

Similarly, in **Studies II** and **III**, choice of stimuli could have affected the coherence values found. In **Study II**, we used the envelope of words and sentences to calculate coherence between the speech signal and brain activity. It is possible that word level stimuli were affected by their shorter length compared to sentences in the estimations in the lower frequencies, therefore coherence calculated to words should be considered with caution. Furthermore, speech tracking studies have been using longer speech segments of up to multiple seconds (Bourguignon et al., 2013a; Molinaro et al., 2016; Riós-Lopez et al., 2020). Using words and sentences in isolation allowed us to investigate differences in coherence to different linguistic units and thus highlight the brain's ability to track speech at shorter time-frames as well. In **Study III**, using single sentences to investigate speech tracking during overt speech production could have affected the findings. Perhaps using longer segments of speech could better accentuate the differences between using the speech envelope from an accelerometer and a microphone to calculate coherence. Nevertheless, even with shorter segments of sentences, we were able to find coherence values above chance level. However, using longer sentences might have introduced confounds on the findings due to the increased load on the short term or working memory

of the participants. Furthermore, as participants were required to say out loud what they have just recently heard, longer sentences or paragraphs would have made the task much harder, especially on the younger participants.

The accuracy of source localization may have limited the interpretations on the brain areas in the studies. In all three studies, in the case of the adult participants' data, the FsAverage brain template was used for source localization due to limited resources available. A three-parameter scaling was used to fit the template with the individual digitized head points, thus minimizing mismatch of brain size and shape between the template brain and the individual brain data. While MEG might not be the most optimal choice for accurate localization of cortical activity, compared to for example fMRI, the primary goal of the research here was not to localize brain activity within millimetre accuracy. Furthermore, MEG is insensitive to tissue conductivity, which has been a challenge when using EEG. Overall, MEG could yield good enough source localization while maintaining a relatively high temporal resolution of up to milliseconds of activity. In **Study II**, in the case of the child participants, individual structural MRIs were collected to improve the accuracy of source localization for that group.

In **Study II**, the number of child participants affected the power of the study to detect more subtle variations related to development and hemispheric processing of speech stimuli in relation to comparison involving the evoked responses as the early responses were only detectable in about half the group.

4.6 Future directions

The results presented in this dissertation focused on brain activity recorded from adults (**Study I, II** and **III**) and typically developing children between the ages of 5 and 9 years (**Study II**). The results provide a good basis to further examine the development of speech perception and production.

Making use of the paradigm created for **Studies II** and **III**, and collecting brain data from more age groups, (e.g. school age children, adolescence, older adults and elders) could give a better view of how speech tracking develops through the lifetime. Our findings showed that with age the brain's ability to track speech increased, indexed by the increase in coherence values, however the lack of correlation between coherence values and the children's age suggests that this increase is most likely not linear. A further study could investigate the ages in between the groups in **Study II** to map out the exact trajectory of changes. Furthermore, it would be worthwhile to examine whether the precision of speech tracking persists with age. Finally, examining speech production in children with the methods used in **Study III** could give us a better understanding of how speech tracking changes during development from early childhood into adulthood.

An interesting finding of **Studies II** and **III** was that coherence values were drastically smaller when calculated in the source space than at sensor level. This was consistent for both adult and child data, and was present for both

perception- and production-related brain data. Future research could investigate how to improve the calculation of coherence values at source level.

In **Study III**, as a secondary goal, the feedforward and feedback systems of speech production were investigated. However, the analysis focused on comparing the two methods (microphone- versus accelerometer-based coherence) at source-level. Connectivity analysis or cross-frequency coupling could reveal more about the sequential relationships between the regions that show coherence and whether the coherence values in the different frequencies are related, as the DIVA model would suggest (Tourville & Guenther, 2011) or are actually independent of each other.

Finally, future research could investigate speech tracking during perception and production in participants with speech production problems by utilising the paradigm used in **Studies II** and **III** with slight modifications to fit the group better. Especially examining cortical tracking during production could provide valuable information about if and how children with speech production problems differ from typically developing children in terms of the feedback and feedforward systems.

YHTEENVETO (SUMMARY)

Aikuisten ja lasten puheen havaitsemiseen ja tuottamiseen liittyvien aivoprosessien tutkiminen MEG:llä

Kasvokkain tapahtuvassa viestinnässä on pystyttävä seuraamaan ja yhdistelemään vastaanotettavaa kuultua ja nähtyä informaatiota, ja yleensä on lisäksi pystyttävä reagoimaan puheeseen ja tuottamaan sitä. Puheen havaitsemista on tutkittu melko laajasti aina tavujen havaitsemisesta keskustelun seuraamiseen asti, mutta on edelleen osittain epäselvää, kuinka aivot havaitsevat puheen. Puheen tuottamista on vieläkin haasteellisempaa tutkia aivotasolla, koska siihen ei liity pelkästään puhe-elinten liike, joka lisää lihastoiminnan häiriöitä tallennettuun signaaliin, vaan puheessa aivojen tarvitsee prosessoida havainto kuulemastaan omasta puheesta. Aivotason tutkimuksissa käytetään usein yksinkertaistettuja ärsykkeitä aivotoiminnan mittaamiseen liittyvien teknisten haasteiden vuoksi, ja sinänsä kielellisiä yksiköitä vertaillaan harvoin.

Tämän väitöstutkimuksen tavoitteena oli selvittää puheen havaitsemista aivoissa tavu-, sana- ja virketasolla, tutkia kehitykseen liittyviä vaikutuksia aivojen kyvyssä seurata erilaisia puheen yksiköitä, ja soveltaa aiemmin puheen havaitsemisen tutkimuksessa käytettyä menetelmää aikuisten puhutun kielen tuottamisen tutkimisessa.

Tutkimuksessa 1 selvitettiin, kuinka yhteensopivien ja -sopimattomien audiovisuaalisten puheärsykkeiden aivovasteet eroavat toisistaan ja kuinka kielellinen kokemus saattaa vaikuttaa näihin eroihin. Tutkimuksen kaksi ryhmää (kiinalainen ja suomalainen) osallistuivat audiovisuaaliseen MEG-kokeeseen, jossa he katsoivat, kun syntyperäinen kiinan puhuja lausui tavuja. Tavut valittiin siten, että ne kaikki kuuluivat kiinalaisten osallistujien fonologiaan, mutta suomalaisille osallistujille vain puolet niistä oli fonologiansa puolesta tuttuja. Kongruenssin vaikutus havaittiin 300–400 millisekunnin aikaikkunalla vasemmalla frontaali- ja oikealla temporaalialueilla sensoritasolla sekä oikealla temporaali-parietaalialueilla lähdetasolla. Tuttuus ei aiheuttanut merkittäviä eroja aivovasteisiin.

Tutkimuksessa 2 syvennyttiin aikuisten ja lasten puheen seuraamiseen sana- ja virketasolla sekä verrattiin seuraamiskykyä, jota koherenssiarvot osoittivat, auditiivisiin tavujen vasteisiin. Tarkoituksena oli selvittää, onko näillä kortikaalisilla vasteilla yhteisiä kehitysmekanismeja. Koherenssiarvoja myös korreloitiin käyttäytymismittojen pisteisiin, jotta nähtäisi, heijastuuko kortikaalisen seuraamisen ero käyttäytymiseen. Aikuisten koherenssiarvot olivat suurempia kuin lasten. Se osoittaa havaitun puheen kortikaalisen seuraamisen lisääntyvän iän mukana ja viittaa siihen, että aivojen kyky seurata puhesignaalin voimakkuusvaihtelua paranee. Korrelaatiota ei havaittu sanojen ja virkkeiden koherenssiarvojen ja tavujen tapahtumasidonnaisten kenttien amplitudin välillä, minkä perusteella vasteilla ei näytä olevan vahvoja yhteisiä kehitysprosesseja. Ikään liittyvät korrelaatiot koherenssiarvojen ja käyttäytymismittojen pisteiden välillä eivät myöskään olleet merkitseviä.

Tutkimuksessa 3 selvitettiin aikuisten puheen seuraamista ääneen lausutun puheen tuottamisen aikana ja pyrittiin validoimaan kiihtyvyysanturin käyttö puheen tallentamisessa vertailemalla koherenssiarvoja, jotka oli laskettu mikrofonin puhesignaalin (kortikoakustinen koherenssi, CAC) ja kiihtyvyysanturin avulla (kortikokinemaattinen koherenssi, CKC). Kaulalle kiinnitetty kiihtyvyysanturi on vähemmän herkkä ulkopuoliselle melulle, joten se voisi tuottaa selkeämmän puheen voimakkuusvaihtelun koherenssiarvojen laskemista varten kuin mikrofoni. Havaitsimme, että CKC on verrattavissa CAC:iin, mutta se on herkkä puheen eri aspekteille ja tuottaa suuremman koherenssin matalammilla taajuuksilla sekä fokaalimman topografian.

Tutkimus toi esille puheen havaitsemisen samankaltaisuutta kaikissa kielellisissä yksiköissä: temporaalialueiden aktivointia tavujen tapahtumasidonnaisten vasteiden yhteydessä sekä sanoihin ja virkkeisiin liittyvissä koherenssiarvoissa. Lisäksi osoitimme, että puheen seuraamisella passiivisen kuuntelun aikana ja ääneen lausutun tuottamisen aikana on yhteisiä vasteita temporaalisilla ja frontaalisilla alueilla, kun taas tuottamisessa ovat mukana myös motoriset ja somatosensoriset alueet.

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.

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2009). Abnormal cortical processing of the syllable rate of speech in poor readers. *Journal of Neuroscience, 29*(24), 7686–7693.

Albrecht, R., Suchodoletz, W., & Uwer, R. (2000). The development of auditory evoked dipole source activity from childhood to adulthood. *Clinical Neurophysiology, 111*(12), 2268–2276.

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018). Cortical tracking of global and local variations of speech rhythm during connected natural speech perception. *Journal of Cognitive Neuroscience, 30*(11), 1704–1719.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience, 29*(43), 13445-13453.

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53*, 115-121.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20*(8), 2225–2234.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program] (Version 6.0.37). Retrieved from http://www.praat.org/

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex, 16*(1), 115-123.

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific Reports, 7*(1), 1-11.

Bourguignon, M., De Tiège, X., de Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., Goldman, S., Hari, R., & Jousmäki, V. (2013a). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping, 34*(2), 314–326.

Bourguignon, M., De Tiège, X., de Beeck, M. O., Pirotte, B., Van Bogaert, P., Goldman, S., Hari, R., & Jousmäki, V. (2011). Functional motor-cortex mapping using corticokinematic coherence. *NeuroImage, 55*(4), 1475–1479.

Bourguignon, M., De Tiège, X., de Beeck, M. O., Van Bogaert, P., Goldman, S., Jousmäki, V., & Hari, R. (2013b). Primary motor cortex and cerebellum are coupled with the kinematics of observed hand movements. *NeuroImage, 66*, 500–507.

Bourguignon, M., Jousmäki, V., Op de Beeck, M., Van Bogaert, P., Goldman, S., & De Tiège, X. (2012). Neuronal network coherent with hand kinematics during fast repetitive hand movements. *NeuroImage, 59*(2), 1684–1691.

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

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems: Learning to Read Reorganizes Language Network. *Human Brain Mapping, 34*(12), 3354–3368.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16*(5), 805–816.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1001–1010.

Capek, C. M., Bavelier, D., Corina, D., Newman, A. J., Jezzard, P., & Neville, H. J. (2004). The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 Tesla. *Cognitive Brain Research, 20*(2), 111–119.

Čeponien, R., Rinne, T., & Näätänen, R. (2002). Maturation of cortical sound processing as indexed by event-related potentials. *Clinical Neurophysiology, 113*(6), 870-882.

Conway, B. A., Halliday, D. M., Farmer, S. F., Shahani, U., Maas, P., Weir, A. I., & Rosenberg, J. R. (1995). Synchronization between motor cortex and spinal motoneuronal pool during the performance of a maintained motor task in man. *The Journal of Physiology, 489*(3), 917–924.

Cunningham, J., Nicol, T., Zecker, S., & Kraus, N. (2000). Speech-evoked neurophysiologic responses in children with learning problems: development and behavioral correlates of perception. *Ear and Hearing, 21*(6), 554-568.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R.Belliveau, JLewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron, 26*(1), 55-67.

Denckla, M. B., & Rudel, R. G. (1976). Naming of object-drawings by dyslexic and other learning disabled children. *Brain and Language, 3*(1), 1–15.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968-980.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11*, 481.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158-164.

Eklund, K., Torppa, M., Aro, M., Leppänen, P. H. T., & Lyytinen, H. (2015). Literacy skill development of children with familial risk for dyslexia through grades 2, 3, and 8. *Journal of Educational Psychology, 107*(1), 126–140.

Fletcher, J. M., & Wennekers, T. (2018). From structure to activity: Using centrality measures to predict neuronal activity. *International Journal of Neural Systems, 28*(02), 1750013.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology, 2*.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica, 66*(1–2), 113–126.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience, 6*.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron, 56*(6), 1127–1134.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience, 7*.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology, 11*(12), e1001752.

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences, 98*(2), 694-699.

Groß, J., Tass, P. A., Salenius, S., Hari, R., Freund, H. J., & Schnitzler, A. (2000). Cortico-muscular synchronization during isometric muscle contraction in humans as revealed by magnetoencephalography. *The Journal of Physiology, 527*(3), 623-631.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*(5), 408-422.

Gwin, J. T., & Ferris, D. P. (2012). Beta-and gamma-range human lower limb corticomuscular coherence. *Frontiers in Human Neuroscience, 6*, 258.

Halliday, D. M., Conway, B. A., & Farmer, S. F. (1995). A frameword for the analysis of mixed time series/point process data—Theory and application to the study of physiological tremor, single motor unit discharges and electromyograms. *Progress in Biophysics and Molecular Biology, 64*(2/3), 237–278.

Hämäläinen, J. A., Ortiz-Mantilla, S., & Benasich, A. A. (2011). Source localization of event-related potentials to pitch change mapped onto age-appropriate MRIs at 6 months of age. *NeuroImage, 54*(3), 1910-1918.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing, 32*(1), 35-42.

Häyrinen, T., Serenius-Sirve, S., & Korkman, M. (1999). Lukilasse. Lukemisen, kirjoittamisen ja laskemisen seulontatestistö peruskoulun ala-asteen luokille, 1-6.

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience, 27*(30), 7881–7887.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135-145.

Hillman, R. E., Heaton, J. T., Masaki, A., Zeitels, S. M., & Cheyne, H. A. (2006). Ambulatory monitoring of disordered voices. *Annals of Otology, Rhinology & Laryngology, 115*(11), 795–801.

Houde, J. F., & Chang, E. F. (2015). The cortical computations underlying feedback control in vocal production. *Current Opinion in Neurobiology, 33*, 174-181.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks, 13*(4–5), 411–430.

Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M.-C., Mazziotta, J.C. & Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the National Academy of Sciences, 98*(24), 13995–13999.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Frontiers in Psychology, 2*, 255.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47–B57.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport, 14*(8), 1129–1133.

Kalashnikova, M., Peter, V., Di Liberto, G. M., Lalor, E. C., & Burnham, D. (2018). Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific Reports, 8*(1).

Kilner, J. M., Baker, S. N., Salenius, S., Hari, R., & Lemon, R. N. (2000). Human cortical muscle coherence is directly related to specific motor parameters. *The Journal of Neuroscience, 20*(23), 8838–8845.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research, 18*(1), 65-75.

Korkman, M., Kirk, U., & Kemp, S. L. (1998). A developmental neuropsychological assessment (NEPSY). Psychological Corporation.

Korkman, M., Kirk, U., & Kemp, S. L. (2008). NEPSY-II: Lasten neuropsykologinen tutkimus [NEPSY-II: A developmental neuropsychological assessment]. Psykologien Kustannus Oy.

Kraus, N., McGee, T., Carrell, T., Sharma, A., Micco, A., & Nicol, T. (1993). Speech-evoked cortical potentials in children. *Journal of the American Academy of Audiology, 4*(4), 238-248.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97*(22), 11850–11857.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831–843.

Larsen, L. H., Zibrandtsen, I. C., Wienecke, T., Kjaer, T. W., Christensen, M. S., Nielsen, J. B., & Langberg, H. (2017). Corticomuscular coherence in the acute and subacute phase after stroke. *Clinical Neurophysiology, 128*(11), 2217-2226.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*(5), 187-196.

Lindstrom, F., Ohlsson, A.-C., Sjöholm, J., & Waye, K. P. (2010). Mean F0 values obtained through standard phrase pronunciation compared with values obtained from the normal work environment: A study on teacher and child voices performed in a preschool environment. *Journal of Voice, 24*(3), 319–323.

Lindstrom, F., Ren, K., Li, H., & Waye, K. P. (2009). Comparison of two methods of voice activity detection in field studies. *Journal of Speech, Language, and Hearing Research, 52*(6), 1658–1663.

Liu, J., Sheng, Y., Zeng, J., & Liu, H. (2019). Corticomuscular coherence for upper arm flexor and extensor muscles during isometric exercise and cyclically isokinetic movement. *Frontiers in Neuroscience, 13*, 522.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001–1010.

Maezawa, H., Mima, T., Yazawa, S., Matsuhashi, M., Shiraishi, H., & Funahashi, M. (2016). Cortico-muscular synchronization by proprioceptive afferents from the tongue muscles during isometric tongue protrusion. *NeuroImage, 128*, 284–292.

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

Marty, B., Bourguignon, M., Jousmäki, V., Wens, V., Op de Beeck, M., Van Bogaert, P., Goldman, S., Hari, R., & De Tiège, X. (2015a). Cortical

kinematic processing of executed and observed goal-directed hand actions. *NeuroImage, 119*, 221–228.
Marty, B., Bourguignon, M., Op de Beeck, M., Wens, V., Goldman, S., Van Bogaert, P., Jousmäki, V., & De Tiège, X. (2015b). Effect of movement rate on corticokinematic coherence. *Neurophysiologie Clinique/Clinical Neurophysiology, 45*(6), 469–474.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.
Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25*(25), 5884–5893.
Molinaro, N., & Lizarazu, M. (2018). Delta (but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience, 48*(7), 2642–2650.
Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., & Carreiras, M. (2016). Out-of-synchrony speech entrainment in developmental dyslexia: Altered cortical speech tracking in dyslexia. *Human Brain Mapping, 37*(8), 2767–2783.
Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neuroscience Letters, 363*(2), 112–115.
Munding, D., Dubarry, A. S., & Alario, F. X. (2016). On the cortical dynamics of word production: A review of the MEG evidence. *Language, Cognition and Neuroscience, 31*(4), 441-462.
Näätänen, R., & Picton, T. W. (1986). N2 and automatic versus controlled processes. *Electroencephalography & Clinical Neurophysiology Supplement, 38*, 169-186.
Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vaionio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*(6615), 432–434.
Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590.
Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31*(5), 1704-1714.
Nishitani, N., & Hari, R. (2002). Viewing Lip Forms: Cortical Dynamics. *Neuron, 36*(6), 1211–1220.
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience,* 1–9.
Orlikoff, R. F. (1995). Vocal stability and vocal tract configuration: An acoustic and electroglottographic investigation. *Journal of Voice, 9*(2), 173–181.

Parviainen, T., Helenius, P., & Salmelin, R. (2019). Children show hemispheric differences in the basic auditory response properties. *Human Brain Mapping, 40*(9), 2699–2710.
Parviainen, T., Helenius, P., Poskiparta, E., Niemi, P., & Salmelin, R. (2011). Speech perception in the child brain: Cortical timing and its relevance to literacy acquisition. *Human Brain Mapping, 32*(12), 2193–2206.
Pattamadilok, C., Morais, J., Colin, C., & Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language, 137*, 103-111.
Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex, 23*(6), 1378–1387.
Piitulainen, H., Bourguignon, M., De Tiège, X., Hari, R., & Jousmäki, V. (2013). Corticokinematic coherence during active and passive finger movements. *Neuroscience, 238*, 361–370.
Piitulainen, H., Bourguignon, M., Hari, R., & Jousmäki, V. (2015). MEG-compatible pneumatic stimulator to elicit passive finger and toe movements. *NeuroImage, 112*, 310–317.
Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245–255.
Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Current Opinion in Neurobiology, 28*, 142–149.
Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience, 21*(6), 322–334.
Ponton, C. W., Eggermont, J. J., Kwong, B., & Don, M. (2000). Maturation of human central auditory system activity: evidence from multi-channel evoked potentials. *Clinical Neurophysiology, 111*(2), 220–236.
Ponton, C., Eggermont, J. J., Khosla, D., Kwong, B., & Don, M. (2002). Maturation of human central auditory system activity: separating auditory evoked potentials by dipole source modeling. *Clinical Neurophysiology, 113*(3), 407–420.
Popescu, A., & Noiray, A. (2019). Reading proficiency and phonemic awareness as predictors for coarticulatory gradients in children. *Proceeding of BUCLD, 44*.
Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience, 18*(6), 2188–2199.
Ríos-López, P., Molinaro, N., Bourguignon, M., & Lallier, M. (2020). Development of neural oscillatory activity in response to speech in children from 4 to 6 years old. *Developmental Science 23*(6), e12947
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Molholm, S., Javitt, D. C., & Foxe, J. J. (2007). Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophrenia Research, 97*(1-3), 173-183.

Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., & Salmelin, R. (2012). Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2-3 Hz. *Journal of Neuroscience, 32*(11), 3786–3790.

Salmelin, R. (2007). Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology, 118*(2), 237–254.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*(1), 141–145.

Scott, S. K. (2005). Auditory processing—speech, space and auditory objects. *Current Opinion in Neurobiology, 15*(2), 197-201.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 546.

Simonyan, K. (2014). The laryngeal motor cortex: Its organization and connectivity. *Current Opinion in Neurobiology, 28*, 15–21.

Simonyan, K., & Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist, 17*(2), 197–208.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964-1973.

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

Takeshita, K., Nagamine, T., Thuy, D. H. D., Satow, T., Matsuhashi, M Yamamoto, J., Takayama, M., Fujiwara, N. and Shibasaki, H. (2002). Maturational change of parallel auditory processing in school-aged children revealed by simultaneous recording of magnetic and electric cortical responses. *Clinical Neurophysiology, 113*(9), 1470-1484.

Taulu, S., & Kajola, M. (2005). Presentation of electromagnetic multichannel data: The signal space separation method. *Journal of Applied Physics, 97*(12), 124905.

Taulu, S., Simola, J., & Kajola, M. (2005). Applications of the signal space separation method. *IEEE Transactions on Signal Processing, 53*(9), 3359–3372.

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *Journal of Neuroscience, 29*(47), 14726–14733.

Telkemeyer, S., Rossi, S., Nierhaus, T., Steinbrink, J., Obrig, H., & Wartenburger, I. (2011). Acoustic processing of temporally modulated sounds in infants: Evidence from a combined near-infrared spectroscopy and EEG Study. *Frontiers in Psychology, 1*.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young

children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*(4), 579.
Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981.
Uhlhaas, P. J., Roux, F., Rodriguez, E., Rotarska-Jagiela, A., & Singer, W. (2010). Neural synchrony and the development of cortical networks. *Trends in Cognitive Sciences, 14*(2), 72-80.
van Atteveldt, N. M., Formisano, E., Goebel, R., & Blomert, L. (2007). Top–down task effects overrule automatic multisensory responses to letter–sound pairs in auditory association cortex. *NeuroImage, 36*(4), 1345–1360.
van Vliet, M., Liljeström, M., Aro, S., Salmelin, R., & Kujala, J. (2018). Analysis of functional connectivity and oscillatory power using DICS: from raw MEG data to group-level statistics in Python. *Frontiers in Neuroscience, 12*, 586.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, 102*(4), 1181–1186.
Vihla, M., Lounasmaa, O. V., & Salmelin, R. (2000). Cortical processing of change detection: Dissociation between natural vowels and two-frequency complex tones. *Proceedings of the National Academy of Sciences, 97*(19), 10590–10594.
Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology, 108*(1), 1–27.
Wechsler, D. (2003a). Weschler preschool and primary scale of intelligence – Third Edition (WPPSI-III). NCS Pearson, Inc., USA
Wechsler, D. (2003b). WISC-IV: Administration and scoring manual. Psychological Corporation.
Wechsler, D. (2008). Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). NCS Pearson.
Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R.J. & Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology, 36*(5), 638–642.
Wunderlich, J. L., Cone-Wesson, B. K., & Shepherd, R. (2006). Maturation of the cortical auditory evoked potential in infants and young children. *Hearing Research, 212*(1-2), 185-202.
Xu, W., Kolozsvari, O. B., Oostenveld, R., Leppänen, P. H. T., & Hämäläinen, J. A. (2019). Audiovisual processing of chinese characters elicits suppression and congruency effects in MEG. *Frontiers in Human Neuroscience, 13*, 18.
Yang, Y., Solis-Escalante, T., van de Ruit, M., van der Helm, F. C., & Schouten, A. C. (2016). Nonlinear coupling between cortical oscillations and muscle activity during isotonic wrist flexion. *Frontiers in Computational Neuroscience, 10*, 126.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review, 5*(4), 683-689.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3.

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and Language, 122*(3), 151–161.

ORIGINAL PAPERS

I

TOP-DOWN PREDICTIONS OF FAMILIARITY AND CONGRUENCY IN AUDIO-VISUAL SPEECH PERCEPTION AT NEURAL LEVEL

by

Orsolya Beatrix Kolozsvári, Weiyong Xu, Paavo Herman Tapio Leppänen & Jarmo Arvid Hämäläinen, 2019

Frontiers in Human Neuroscience, *13*, 243

Available online: https://doi.org/10.3389/fnhum.2019.00243

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 12 July 2019
doi: 10.3389/fnhum.2019.00243

Top-Down Predictions of Familiarity and Congruency in Audio-Visual Speech Perception at Neural Level

Orsolya B. Kolozsvári[1,2], Weiyong Xu[1,2], Paavo H. T. Leppänen[1,2] and Jarmo A. Hämäläinen[1,2]*

[1] Department of Psychology, University of Jyväskylä, Jyväskylä, Finland, [2] Jyväskylä Centre for Interdisciplinary Brain Research (CIBR), University of Jyväskylä, Jyväskylä, Finland

During speech perception, listeners rely on multimodal input and make use of both auditory and visual information. When presented with speech, for example syllables, the differences in brain responses to distinct stimuli are not, however, caused merely by the acoustic or visual features of the stimuli. The congruency of the auditory and visual information and the familiarity of a syllable, that is, whether it appears in the listener's native language or not, also modulates brain responses. We investigated how the congruency and familiarity of the presented stimuli affect brain responses to audio-visual (AV) speech in 12 adult Finnish native speakers and 12 adult Chinese native speakers. They watched videos of a Chinese speaker pronouncing syllables (/pa/, /pha/, /ta/, /tha/, /fa/) during a magnetoencephalography (MEG) measurement where only /pa/ and /ta/ were part of Finnish phonology while all the stimuli were part of Chinese phonology. The stimuli were presented in audio-visual (congruent or incongruent), audio only, or visual only conditions. The brain responses were examined in five time-windows: 75–125, 150–200, 200–300, 300–400, and 400–600 ms. We found significant differences for the congruency comparison in the fourth time-window (300–400 ms) in both sensor and source level analysis. Larger responses were observed for the incongruent stimuli than for the congruent stimuli. For the familiarity comparisons no significant differences were found. The results are in line with earlier studies reporting on the modulation of brain responses for audio-visual congruency around 250–500 ms. This suggests a much stronger process for the general detection of a mismatch between predictions based on lip movements and the auditory signal than for the top-down modulation of brain responses based on phonological information.

Keywords: speech perception, magnetoencephalography, audio-visual stimuli, audio-visual integration, familiarity

OPEN ACCESS

Edited by:
Xiaolin Zhou,
Peking University, China

Reviewed by:
Lihan Chen,
Peking University, China
Nai Ding,
Zhejiang University, China

****Correspondence:***
Orsolya B. Kolozsvári
orsolya.b.kolozsvari@jyu.fi

Received: *16 January 2019*
Accepted: *28 June 2019*
Published: *12 July 2019*

Citation:
Kolozsvári OB, Xu W, Leppänen PHT and Hämäläinen JA (2019) Top-Down Predictions of Familiarity and Congruency in Audio-Visual Speech Perception at Neural Level. Front. Hum. Neurosci. 13:243. doi: 10.3389/fnhum.2019.00243

INTRODUCTION

In most cases speech perception relies on the seamless interaction and integration of auditory and visual information. Listeners need to efficiently process a rapid and complex stream of multisensory information, making use of both visual and auditory cues. We wanted to examine how lifelong exposure to audio-visual speech affects the brain mechanisms of cross-modal integration

and mismatch. Auditory and visual cues can be presented either congruently or incongruently and this match or mismatch of features could be used to study the audio-visual processing of speech. Using magnetoencephalography (MEG), we studied how the effects of congruency and familiarity (i.e., whether the speech stimuli are part of the listener's phonology or not) of the auditory and visual features are reflected in brain activity.

Audio-visual speech has been shown to activate (in sequence) the sensory areas around 100 ms from stimulation onset in the auditory and visual cortices (Sams et al., 1991; Möttönen et al., 2004; Salmelin, 2007), then the superior temporal sulcus around 150 ms (Nishitani and Hari, 2002), which has been shown to play an important role in the perception and interpretation of movements (both facial and body) of the speaker (Puce et al., 1998; Iacoboni et al., 2001). The inferior parietal cortex has been shown to be activated at around 200 ms, which is suggested to be related to the connection of the STS to the inferior frontal lobe (Broca's area) (Nishitani and Hari, 2002) with stronger activations in the left hemisphere than in the right (Capek et al., 2004; Campbell, 2008). This is followed by activation in the frontal areas close to Broca's area around 250 ms (Nishitani and Hari, 2002).

It has been suggested (Campbell, 2008) that seeing speech can affect what is perceived in either a complementary or correlated way. In the complementary mode, vision offers further information about some aspects of speech, which are harder to detect only auditorily and which may depend on the clear visibility of the speaker's lower face. In the correlated mode, on the other hand, successful speech processing depends on the speech stream's temporal-spectral signature showing similar dynamic patterning across both the audible and visible channels.

Audio-visual mismatch is often examined from the point of view of congruency (Jones and Callan, 2003; Hein et al., 2007), where congruent and incongruent audio-visual pairs are contrasted. The assumption is that congruency should only have an effect on perception when the inputs of unimodal sources have been integrated (van Atteveldt et al., 2007). In terms of brain responses, the STS has been shown to be a critical brain area for multisensory integration and congruency of auditory and visual information in the case of both speech and non-speech stimuli. For example, Beauchamp et al. (2010) used TMS to disrupt brain activity in STS, while participants viewed audio-visual stimuli that have been shown to cause the McGurk effect (where incongruent auditory and visual speech cues presented together produce an illusory percept; McGurk and Macdonald, 1976). When TMS was applied to the left STS during the perception of McGurk pairs, the frequency of the McGurk percept was greatly reduced. This reduction, in the likelihood of the McGurk effect, demonstrates that the STS is an important cortical locus for the McGurk effect and plays an important part in auditory-visual integration in speech.

Furthermore, a broad network of brain regions in addition to the STS have been found in fMRI studies to show differences between brain responses to incongruent and congruent audio-visual speech, including the precentral gyrus (Jones and Callan, 2003), the inferior parietal lobule (Jones and Callan, 2003), the supramarginal gyrus (Jones and Callan, 2003), the superior frontal gyrus (Miller and D'Esposito, 2005), Heschl's gyrus (Miller and D'Esposito, 2005) and the middle temporal gyrus (Callan et al., 2004).

Previous studies examining audio-visual speech have found relatively early event-related brain potential (ERP) effects around N1 and P2 responses (Stekelenburg and Vroomen, 2007; Baart et al., 2014). In this case the visual information leads the auditory information, that is, lip movements can precede actual phonation for up to several hundreds of milliseconds (Stekelenburg and Vroomen, 2007). This visual information allows the observer to make predictions about several aspects of the auditory signal (e.g., content, timing). Studies have shown that the auditory-evoked N1 and P2 components of ERPs, at latencies of 100–150 and 200–250 ms, respectively, are attenuated and speeded up when the auditory signal is accompanied by visual speech (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007). This suggests early predictive effects of the visual information on the auditory stimulation. Furthermore, no attenuation in N1 was found when no visual anticipatory information about sound onset is present, indicating that the temporal information present in the visual stimulus, rather than the content of the sound, is key in audio-visual interaction (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010).

However, the N1 and P2 responses seem to be sensitive to the stimulus material. This was shown by Baart et al. (2014), who investigated speech-specific audio-visual integration, where they used speech stimuli and sinewave speech, and found that N1 suppression occurs regardless of the type of stimuli, but P2 amplitude was only suppressed in relation to speech stimuli. They found congruency effects for responses to speech stimuli from around 200 ms after audio-visual incongruency became apparent, with ERPs being more negative for congruent stimuli than for incongruent stimuli. These early suppression effects were found when comparing the brain responses between the unimodal and the multimodal stimuli.

In addition, audio-visual speech congruency effects have also been reported in later time-windows. Arnal et al. (2009) investigated how the visual signal of an audio-visual stimulus affects auditory speech processing. In their experiment they recorded early visual and auditory responses to matching (congruent) and non-matching (incongruent) audio-visual syllables using MEG and found no effect of audio-visual incongruence in the early time-window (M100). They detected the earliest mismatch effect 120 ms after voice onset, followed by three more maxima at 250, 370, and 460 ms. Their findings indicated a multistep comparison between the top-down visual prediction and the bottom-up auditory signal.

Another aspect affecting audio-visual speech is the long-term memory representations of speech, that is, the familiarity of the speech itself. It has been documented that speech perception is altered by an individual's language experience. Iverson et al. (2003) found that listeners of different languages respond to distinct acoustic aspects of the same speech stimulus. They compared Japanese, German, and English speakers' responses to contrasts of /ra/ and /la/, where they had to rate whether the stimulus presented was a good exemplar of their own

native-language phoneme. They found that American listeners attend to the third formant, which reliably distinguishes /r/ from /l/, while Japanese listeners attend more strongly to the second formant, which is critical for distinguishing Japanese phonemes, but is not at all helpful in distinguishing /r/ from /l/.

This and other studies suggest that the effects of language experience on speech perception are due to neural coding of the acoustic components that are critical to native-language processing (e.g., Kuhl, 2000, 2004). Such effects of language exposure are reflected in brain responses around 150–200 ms, for example in the modulation of the strength of the mismatch negativity (MMN), which is thought to tap into language-specific perceptual sensitivity (Näätänen et al., 1997, 2007; Winkler et al., 1999; Zhang et al., 2005, 2009). Language-specific phonetic-phonological analysis has been shown to start 100–200 ms following stimulus onset (Vihla et al., 2000; Näätänen et al., 2007). MMN or mismatch field (MMF) in EEG and MEG studies, respectively, have indicated access to phonological categories (Vihla et al., 2000; Näätänen et al., 2007) and the distinct processing of native and non-native phonetic contrasts (Näätänen et al., 1997, 2007) in this time-window.

By comparing two groups with different native languages (Finnish and Chinese), we aimed to see how long-term audio-visual representations affect speech perception by examining the congruency effects. Additionally, we aimed to distinguish the effects of familiarity, which is a learned aspect of speech, from congruency, which should be an inherent aspect of the audio-visual stimuli related to the general correspondence between mouth movements and speech signal.

To this end, we compared brain responses measured with MEG to unfamiliar and familiar (called aspirated and unaspirated, respectively, see section "Materials and Methods" below) and also congruent and incongruent audio-visual speech stimuli. We expected to find significant differences in responses to congruent and incongruent stimuli for both Chinese and Finnish participants with larger responses to incongruent stimuli starting from 150 ms or later based on the previous literature (e.g., Arnal et al., 2009). However, in the case of the Finnish participants, we expected differences between the familiar and unfamiliar stimuli specifically starting in the same time-window as the congruency effect (150 ms onward), with the unfamiliar stimuli producing a larger response than the familiar stimuli if long-term phonological representations facilitate the processing of audio-visual speech.

MATERIALS AND METHODS

Participants

Participants were adult Finnish native speakers and adult Chinese native speakers studying in Jyväskylä, Finland. None of the participants had neurological or learning problems, hearing difficulties, using medication affecting the central nervous system, head injuries, ADHD or language-specific disorders. They all had normal or corrected-to-normal sight. The Finnish participants had no exposure to the Chinese language. In total, 19 Finnish native speakers and 18 Chinese native speakers participated in the study. Of these, 13 were excluded from the analysis due to excessive head movement (two participants), poor vision after correction (two participants), technical problems during recording (three participants), strong noise interference (two participants), or otherwise bad signal quality (four participants). Data included in the analysis were from 12 Finnish participants and 12 Chinese participants (see **Table 1** for characteristics of participants included).

Ethical approval for the study was provided by the Ethical Committee of the University of Jyväskylä. Participants gave their written informed consent to participate in the study. All participants received movie tickets as compensation for participating in the study.

Stimuli

The stimuli were video recordings of the syllables /pa/, /pha/, /ta/, /tha/ and /fa/. Of these five syllables, /fa/ was used for a cover task to maintain participants' attention on the stimuli [see **Figure 1** for oscillograms, spectrograms and acoustic features of the stimuli. Figures were created using Praat (Boersma and Weenink, 2018), see **Table 2** for description of the stimuli]. The videos were recorded using a Canon Legria HF200 HD video camera and were edited in Adobe Premier Pro CS5.5 to be 1800 ms long. The videos were recordings of a male native Mandarin Chinese speaker.

For the Finnish participants, /pa/ and /ta/ were considered familiar stimuli because they are part of their native phonology. For the Chinese participants all four syllables were familiar. The recordings could be audio only, in which the participant was presented with the audio track and the still image of the speaker; visual only, in which the video was presented without any sound; and audio-visual, where both audio track and video were presented at the same time. The audio-visual condition could be congruent, where what they saw was what they heard, or incongruent, where the audio did not match the video.

Procedure

Participants sat in a magnetically shielded, sound-attenuated room. They sat under the MEG helmet in a 68° sitting position.

Stimuli were presented using Presentation software (version 18.1; Neurobehavioral Systems, Inc., Albany, CA, United States) running on a Microsoft Windows computer using a Sound Blaster Audigy RX sound card and NVIDIA Quadro K5200 video card.

The stimuli were presented on a projector screen. Stimuli were projected from outside of the measurement room onto a mirror then reflected onto the projector screen using a Barco

TABLE 1 | Participant characteristics.

Native language	Finnish	Chinese
Mean age (*SD*)	23.92 (1.98)	24.75 (3.39)
Gender ratio (male:female)	6:6	3:9
Handedness ratio (right:left)	12:0	12:0

Mean age, gender ratio and handedness are for those included in the analysis.

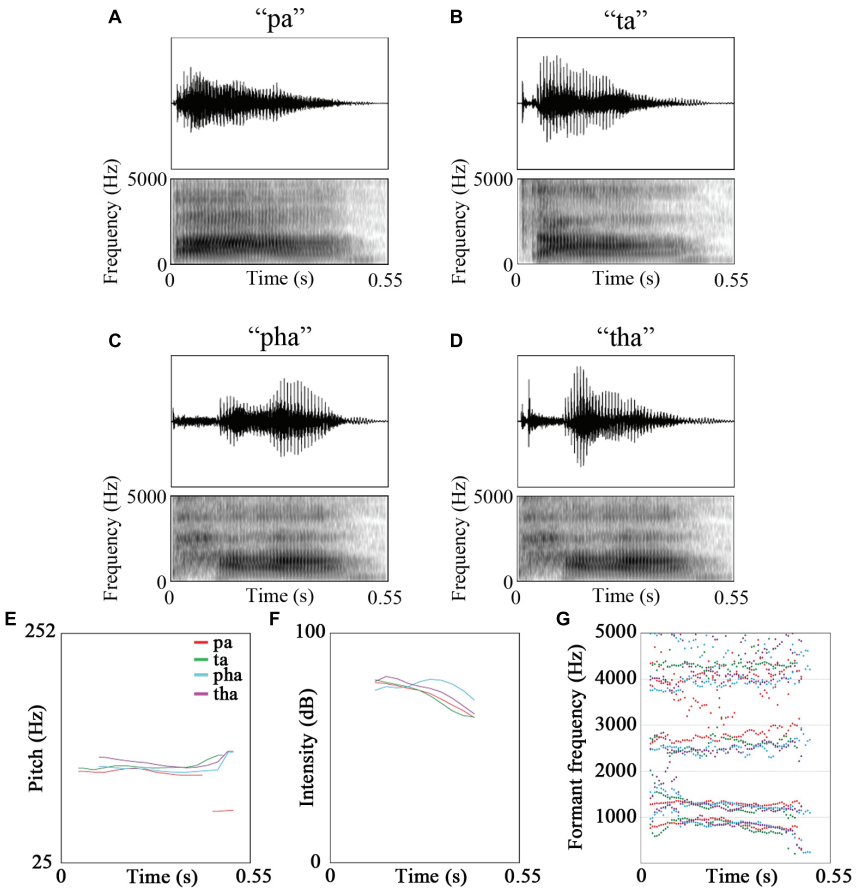


FIGURE 1 | Oscillograms, spectrograms and the acoustic features of the stimuli **(A)** pa, **(B)** ta, **(C)** pha, **(D)** tha, **(E)** pitch, **(F)** intensity, **(G)** formant frequencies (red - /pa/, green - /ta/, cyan - /pha/, purple - /tha/).

TABLE 2 | Stimuli description.

Modality	Target	Familiar / Unaspirated		Unfamiliar / Aspirated	
Audio	fa A	pa A	ta A	pha A	tha A
Visual	fa V	pa V	ta V	pha V	tha V
AV congruent	fa V / fa A	pa V / pa A	ta V/ ta A	pha V /pha A	tha V / tha A
AV incongruent	–	pa V / tha A	ta V / pha A	pha V / ta A	tha V / pa A

FL35 projector. The participants were sitting 1 m from the projection screen.

The participants were asked to watch short videos of a speaker uttering syllables and to attend to all stimuli presented. The videos were cropped to the mouth area of the speaker (from just above the nose to the clavicles). The fixation cross before the onset of the video clip was centered on where the lips of the speaker were in the videos. Videos were presented on a black background, in the center of the screen. The lights were dimmed. Sounds were presented through insert earphones (Lo-Fi auditory stimulation system, Elekta MEGIN Triux) at ∼70 dB sound pressure level.

The participants were presented with a blank screen for 500 ms, then a fixation cross for 550 ms, followed by a still image of the speaker for 500 ms and finally the stimuli, which was 1800 ms long.

The participants received eight practice trials. In the actual experiment 220 stimuli (20 targets for the cover task, and 50 audio-visual congruent, 50 audio-visual incongruent, 50 audio and 50 visual stimuli; /pa/ and /ta/ repeated 12 times each, /pha/ and /tha/ repeated 13 times each) were presented in pseudo-random order with no immediate repetitions of the same stimuli. Stimuli were presented in two blocks, with a short break (duration determined by the participant) in between the blocks (see **Figure 2** for a schematic representation of the video sequence and timings).

As a cover task the participants were asked to press a button to indicate if they saw and/or heard the target syllable /fa/.

Magnetoencephalography Recording and Preprocessing

The MEG data were recorded by a whole-head 306 channel Elekta Neuromag TRIUX MEG device in Jyväskylä, Finland, including

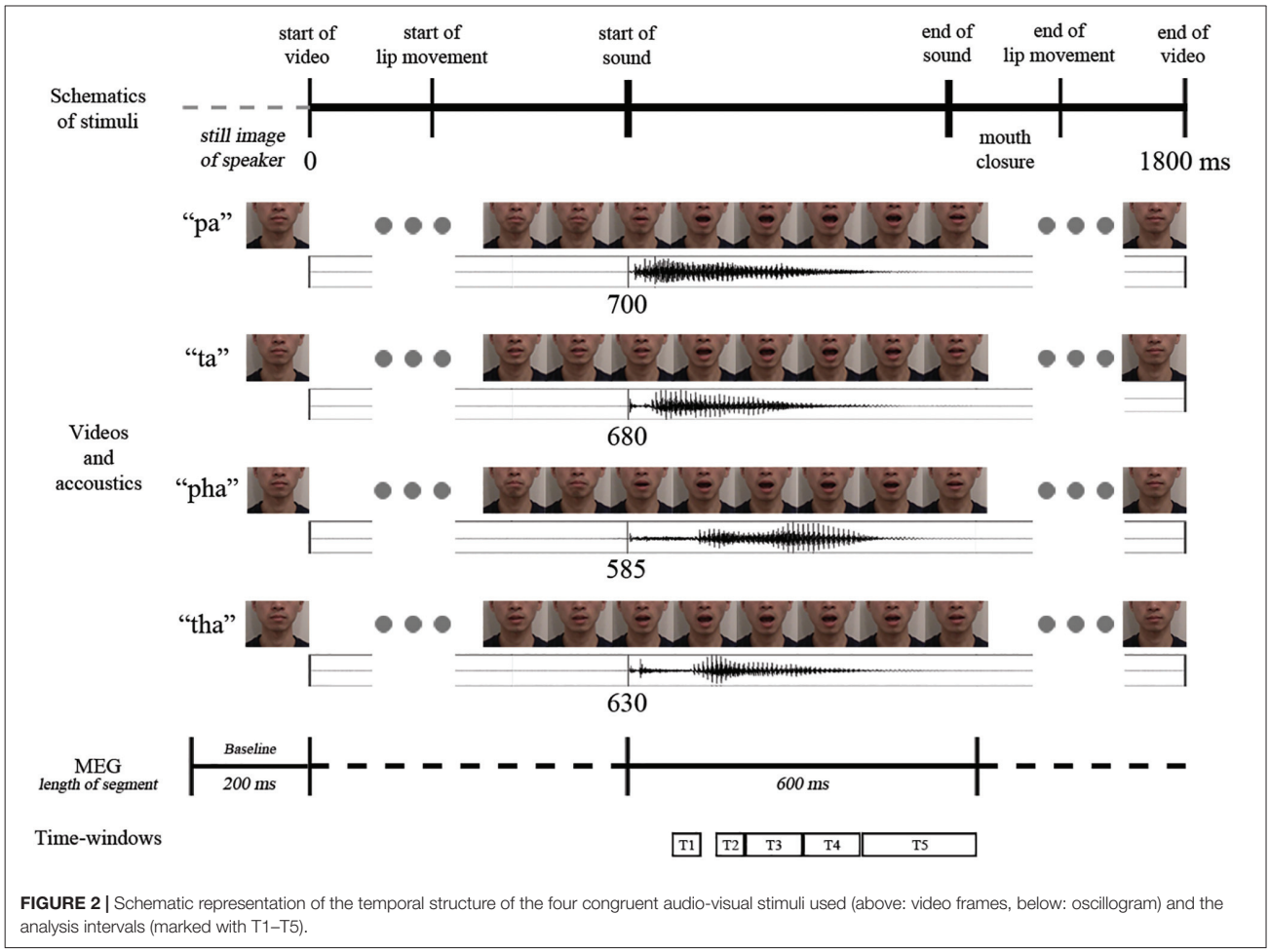


FIGURE 2 | Schematic representation of the temporal structure of the four congruent audio-visual stimuli used (above: video frames, below: oscillogram) and the analysis intervals (marked with T1–T5).

102 magnetometers and 204 orthogonal planar gradiometers. EOG was measured from two diagonally placed electrodes, slightly above the right eye and slightly below the left eye, with the ground electrode on the right clavicle. Five head position indicator (HPI) coils were attached to the scalp, three on the forehead and one behind each ear, and were used to monitor the location of the head in relation to the sensors during the recording by sending 293, 307, 314, 321, and 328 Hz sinusoidal currents into the five coils, respectively. The Polhemus Isotrak digital tracker system (Polhemus, Colchester, VT, United States) was used to determine the position of the HPI coils in relation to three anatomical landmarks (the nasion, left and right preauricular points). For co-registration purposes an additional set of scalp points (>100) were also digitized, distributed randomly over the skull.

Magnetoencephalography data were collected with a sampling rate of 1000 Hz and an online filter of 0.1–330 Hz. All data were preprocessed using the temporal extension of the signal space separation (tSSS) method with buffers of 30 s (Taulu and Kajola, 2005; Taulu et al., 2005) in Maxfilter 3.0^{TM} (Elekta AB) to remove external interference and correct for head movements. Bad channels were identified by visual inspection and marked for exclusion and reconstructed by the MaxFilter program. Head position was estimated in 200 ms time-windows and 10 ms steps for movement compensation.

Data were preprocessed using MNE Python (0.16.2) (Gramfort et al., 2013). Independent component analysis (ICA) using the fastICA algorithm (Hyvärinen and Oja, 2000) was applied to remove eye blinks, horizontal eye movements and cardiac artifacts. Data were low-pass filtered at 35 Hz using a zero-phase FIR filter with a bandwidth of 8.8 Hz. Then the continuous MEG recording was epoched into 200 ms before to 1800 ms after the onset of the video stimuli in the audio-visual condition. The epoched data were baselined using the 200 ms preceding the onset of stimuli. The epochs were shortened and realigned to 200 ms before and 1000 ms after the start of sound in the audio-visual condition. Data were then manually checked to remove any head movement–related artifacts and electronic jump artifacts. MEG epochs exceeding 2 pT/cm for gradiometer or 4 pT for magnetometer peak-to-peak amplitudes were excluded from further analysis. After artifact rejection, an average of 96.50% of trials were used for analysis. Event-related fields were obtained by averaging trials for different conditions separately. The data were then

resampled to 250 Hz to shorten the computation time in the statistical analysis.

Statistical analysis of sensor-level data was done in FieldTrip toolbox (downloaded 20 October 2016; Oostenveld et al., 2011) for MATLAB R2016b (The MathWorks Inc., Natick, MA, 2000) while source-level analyses were run in MNE Python.

Time-Windows

Based on previous literature, five time-windows were investigated: 75–125, 150–200, 200–300, 300–400, and 400–600 ms (where 0 ms is the start of the sound in the section "Stimuli" as described above). The first time-window encompasses the basic auditory N1 m response (Poeppel et al., 1996; Parviainen et al., 2005; Salmelin, 2007), where the brain extracts speech sounds and their sequences from the incoming auditory signal and the responses are expected to be in the auditory cortices. The second time-window has been shown to be involved in further phonemic processing of the stimulus (Näätänen et al., 1997, 2007; Salmelin, 2007) with responses localized to the temporal cortex. The third time-window has been shown to be responsive to lexical-semantic manipulations (Helenius et al., 2002; Kujala et al., 2004) as well as to audio-visual manipulations (e.g., Raij et al., 2000; Arnal et al., 2009, around 250 ms), as have the fourth (Arnal et al., 2009, around 370 ms; Baart et al., 2014, 300–500 ms after onset of AV congruency) and the fifth time-windows (Arnal et al., 2009, around 460 ms).

Sensor-Level Analysis

Averaged planar gradiometer data were transformed into combined planar gradients using the vector sum of the two orthogonal sensors at each position implemented in the Fieldtrip toolbox (Oostenveld et al., 2011), which were then used for sensor-level analysis. Gradiometers were chosen because they are less sensitive to noise sources originating far from the sensors than magnetometers are.

Permutation tests with spatial and temporal clustering based on *t*-test statistics were carried out for planar gradients of individual averaged ERFs (Maris and Oostenveld, 2007). The five time-windows defined (see above) were investigated separately, with a cluster α level of 0.05 and the number of permutations 3000.

Source-Level Analysis

Source analysis was carried out with a minimum-norm estimate on the event-related fields of the magnetometers and gradiometers (Hämäläinen and Ilmoniemi, 1994). The noise covariance matrix was calculated from the baseline period of 200 ms preceding the start of the video (i.e., the participants were viewing the still image of the speaker).

Individual magnetic resonance images (MRI) were not available from the participants and therefore Freesurfer (RRID:SCR_001847) average brain (FSAverage) was used as a template for the source analysis (see below). Three-parameter scaling was used to co-register FSAverage with individual digitized head points. The average co-registration error was 3.54 mm (*SD*=0.27). A single layer BEM (Boundary Element Method) solution was used for the forward modeling.

Depth-weighted L2-minimum-norm estimate (wMNE) (Hämäläinen and Ilmoniemi, 1994; Lin et al., 2006) was calculated for 4098 current dipoles with free orientation distributed on the cortical surface in each hemisphere. Dynamic statistical parametric mapping (dSPM) (Dale et al., 2000) was used to noise-normalize the inverse solution for further statistical analysis. Cluster-based permutation statistics in MNE Python were run on the dSPM source waveforms.

Statistical Analyses

Accuracy and reaction times in the cover task were examined using Target type (Audio only, Visual only, Audio-Visual) by Native language (Finnish, Chinese) ANOVAs.

Congruency and familiarity effects were examined using the interaction of Stimulus by Native language by comparing difference waves between the groups. If no significant results were obtained, Stimulus main effects were investigated between the stimuli. For comparisons investigating congruency, we compared responses to the congruent and incongruent audio-visual stimuli. For comparisons investigating familiarity, we compared responses to the congruent unaspirated audio-visual (/pa/ and /ta/ syllables) and the congruent aspirated audio-visual (/pha/ and /tha/ syllables) stimuli.

RESULTS

Behavioral Performance

Participants' accuracy scores were close to 100% (Finnish: 97.88%; Chinese: 98.35%) (**Table 3**), indicating that they were indeed paying attention to the stimuli. Accuracy (percentage of correct responses) were averaged for each participant, and a 3 (Target type: Audio only, Visual only, Audio-Visual) × 2 (Native language: Finnish, Chinese) repeated measures ANOVA resulted in no significant interaction or main effects.

Reaction times were on average 1189.72 ms (*SD*: 125.86) (**Table 4**). Reaction times were averaged for each participant, and a 3 (Target type: Audio only, Visual only, Audio-Visual) × 2 (Native language: Finnish, Chinese) repeated measures mixed ANOVA resulted in a significant Target type main effect [$F(1.954,42.985) = 6.338$, $p = 0.004$, partial $\eta 2 = 0.224$]. *Post hoc* *t* tests revealed that there was a significant difference between response time to visual only and audio only targets [$t(23) = 2.943$, $p = 0.007$], and audio-visual and audio only targets [$t(23) = 3.518$, $p = 0.002$] with audio only targets having longer reaction times than the other targets.

TABLE 3 | Accuracy scores for the Finnish and Chinese participants in detecting the target syllable /fa/.

	Accuracy (% of correct response to the target stimulus)			
	AV stimuli (%)	A stimuli (%)	V stimuli (%)	All stimuli (%)
Finnish (*n* = 12)	100	97.22	96.43	97.88
Chinese (*n* = 12)	98.81	98.61	97.62	98.35
Total (*n* = 24)	99.40	97.92	97.02	98.12

TABLE 4 | Reaction times for the Finnish and Chinese participants in detecting the syllable /fa/.

	Reaction times in ms (*SD*)			
	AV stimuli	**A stimuli**	**V stimuli**	**All stimuli**
Finnish (*n* = 12)	1170.56 (94.06)	1230.43 (94.51)	1187.56 (141.20)	1193.69 (103.84)
Chinese (*n* = 12)	1152.81 (151.20)	1201.29 (83.87)	1142.48 (160.23)	1163.16 (127.68)
Total (*n* = 24)	1161.69 (123.48)	1215.86 (88.64)	1165.02 (149.48)	1178.42 (114.88)

MEG

Our focus was on the native language interactions and we first examine, and report results with significant native language effects. In the absence of interactions, we report the main effects of congruency and familiarity.

Grand average plots of responses at sensor and source level for the congruency comparison and the familiarity comparison can be seen in **Supplementary Figures S1**, **S2**, respectively.

Sensor-Level Analysis

Congruency Effects

No significant effects were found in the first, second, third or fifth time-windows.

In the fourth time-window, two clusters were found to be significant for the Congruency main effect (responses to the incongruent stimuli compared to responses to the congruent stimuli) after the cluster permutation tests. One cluster (p = 0.036654) was found in the left frontal areas and another cluster (p = 0.046651) was found in the right temporal areas. See **Figure 3** for the topographic maps of brain responses in this time-window. See **Figure 4** for the topographic maps of the clusters and the average evoked responses from the channels forming the clusters.

Familiarity Comparison (Audio-Visual)

No significant statistical effects were found in the five time-windows examined using the cluster permutation tests.

Source-Level Analysis

Congruency Effects

No significant differences were found in the first, second, third and fifth time-windows.

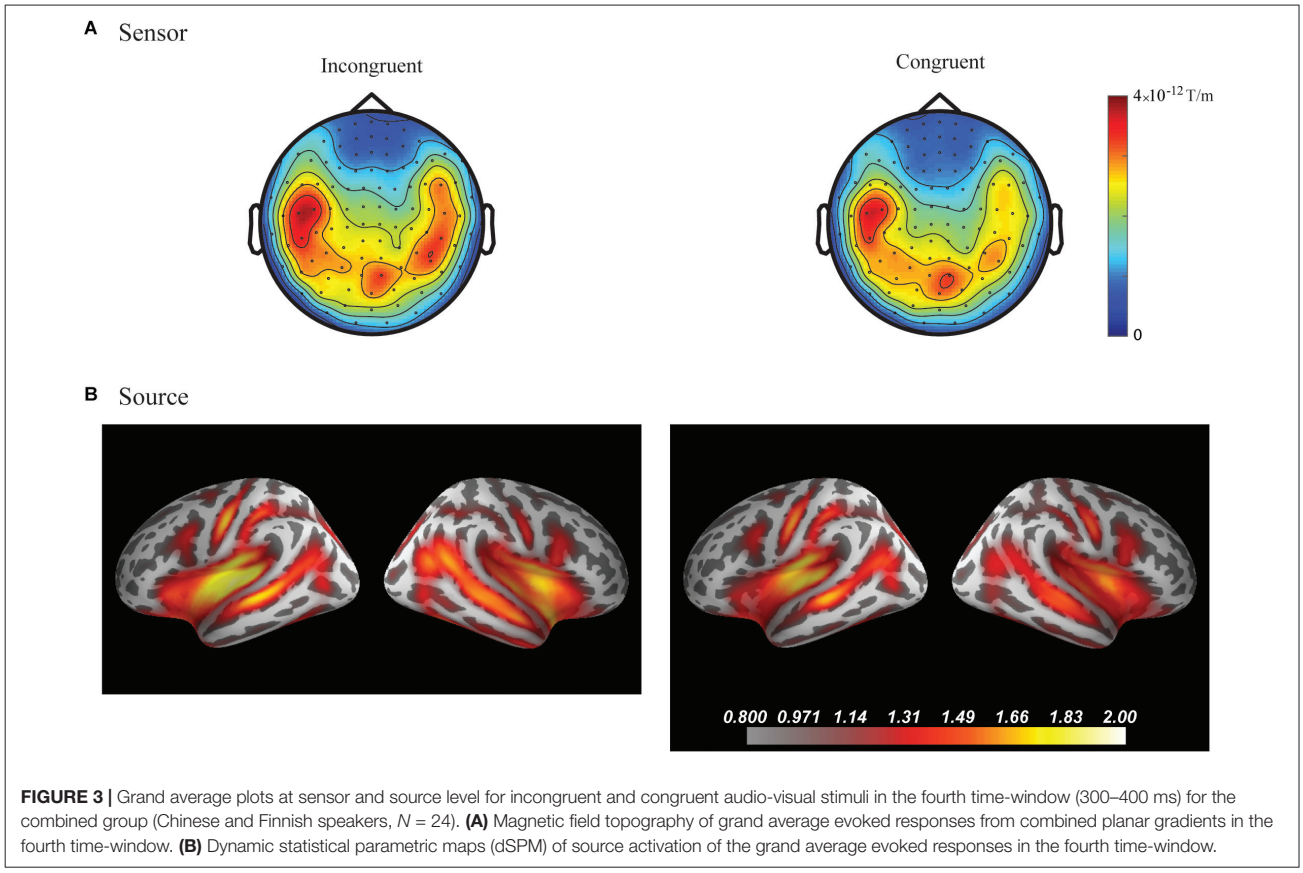


FIGURE 3 | Grand average plots at sensor and source level for incongruent and congruent audio-visual stimuli in the fourth time-window (300–400 ms) for the combined group (Chinese and Finnish speakers, *N* = 24). **(A)** Magnetic field topography of grand average evoked responses from combined planar gradients in the fourth time-window. **(B)** Dynamic statistical parametric maps (dSPM) of source activation of the grand average evoked responses in the fourth time-window.

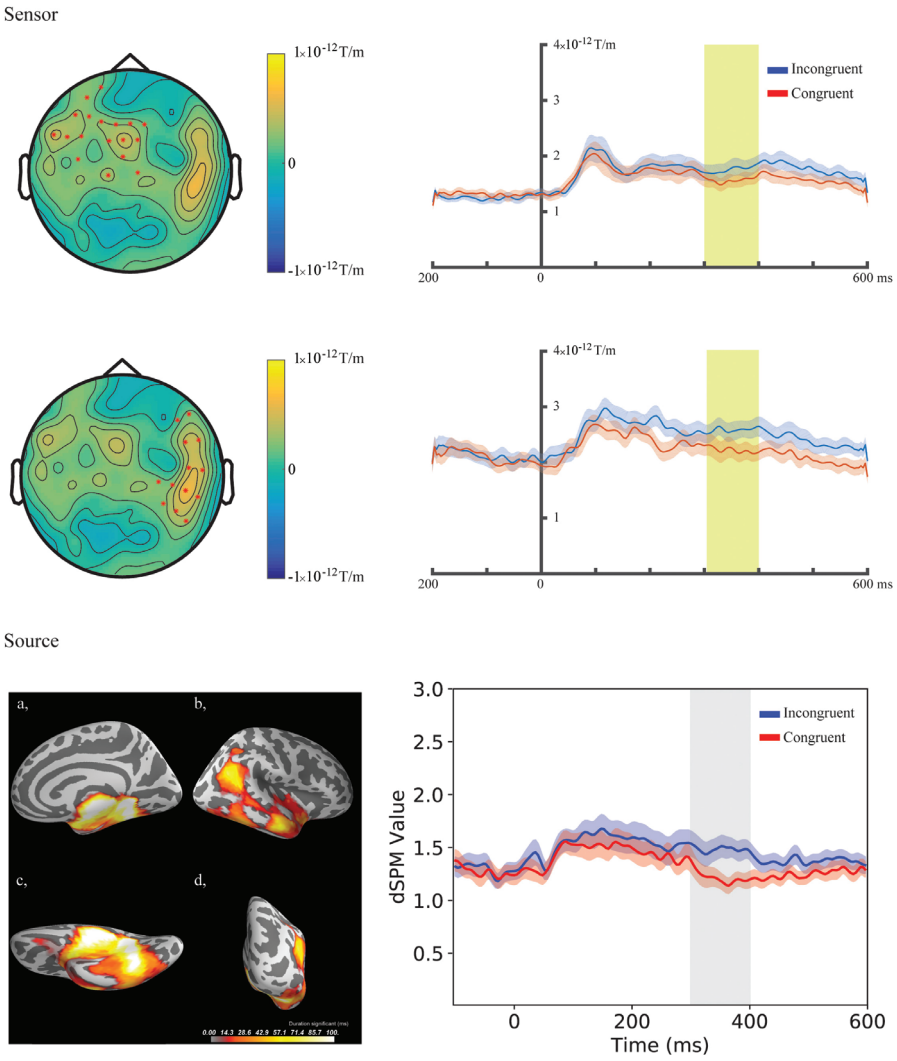


FIGURE 4 | Sensor- and source-level clusters based on the permutation statistics of the congruency effects for the combined group (Chinese and Finnish speakers, $N = 24$). **Left:** Clusters are represented by red dots in the sensor space and yellow and red colouring on the cortical surfaces for the source space. The brightness of the cluster was scaled by the temporal duration of the cluster in the source space. Lower left corner shows the cluster-based permutation statistics results for the incongruent vs congruent comparison in the source space: (a) Medial view, (b) Lateral view, (c) Ventral view, (d) Caudal view. **Right:** average evoked responses from the channels forming the cluster for the sensor space results and the source waveform (dSPM value) extracted from the clusters for the source space results. The red and blue shaded area represents the standard error of the mean and the yellow and gray shaded area indicates the time-window of the cluster.

In the fourth time-window, one cluster was found to be significant ($p = 0.039$) after the cluster permutation tests for the Congruency main effect (responses to the incongruent stimuli compared to responses to the congruent stimuli). The cluster encompassed the right temporal-parietal and medial areas. See **Figure 3** for dynamic statistical parametric maps (dSPM) source activation in this time-window. See **Figure 4** for the source waveform (dSPM value) extracted from the significant cluster.

Familiarity Comparison (Audio-Visual)

No significant statistical effects were found in the five time-windows examined using the cluster permutation tests.

All non-significant results of the permutation tests in the five time-windows, with lowest p-values, are reported in the **Supplementary Material 3**.

DISCUSSION

We investigated how the congruency and familiarity of a stimulus could affect audio-visual speech perception in two groups of adults, native speakers of Chinese and those of Finnish. The Chinese participants had long-term exposure to all of the stimuli because they belonged to their native language, but

some of the speech sounds were not part of Finnish phonology, thus making them unfamiliar for the Finnish participants. We found significant differences in the congruency comparisons across these groups. A significant congruency main effect was found in the frontal and temporal regions at the sensor level and in the right temporal-parietal regions at the source level 300–400 ms following the onset of sound, but no significant effects were found for familiarity comparisons. Matching and mismatching audio-visual speech thus produces robust and replicable processing differences in the brain, which is consistent with findings in earlier studies. Direct comparison of responses to stimuli familiar (unaspirated) and unfamiliar (aspirated) to the Finnish participants do not show evidence for strong cross-modal top-down predictions that would modulate obligatory sensory brain responses.

We found a significant difference between the responses to the congruent and incongruent stimuli for Chinese and Finnish participants in the 300–400 ms time-window bilaterally at the sensor level at the left frontal and right temporal areas as well as in the right hemisphere at the source level in the temporal-parietal areas, indicating that both groups detected the incongruency. The time-window is in line with similar earlier studies using native language stimuli where the incongruence effects were found around 300–500 ms (Arnal et al., 2009; Baart et al., 2014). The localization of the congruency effect seems to depend on the task and contrast used. For example, left hemisphere emphasis was found using more complex stimulation with six different syllables (Arnal et al., 2009) and left frontotemporal regions for symbol–speech sound comparisons (Xu et al., 2019).

The direction of the congruency effect was also in line with earlier studies using audio-visual stimuli showing more brain activity for the incongruent compared to the congruent stimuli (e.g., Arnal et al., 2009; Xu et al., 2019). The direction of the effect likely indicates the benefit of using two modalities to decode the speech signal reflected in less allocation of neuronal resources to the process when the two modalities match (e.g., Bernstein and Liebenthal, 2014). For the incongruent stimuli, the brain response likely includes an error detection signal for the mismatching auditory and visual input. Similar to Arnal et al. (2009), we compared responses to congruent and incongruent stimuli. In their study, they found significant differences in relatively late time-windows, which showed multiple steps for audio-visual processing (with differences at ∼250, ∼370, and ∼460 ms, with responses being larger for the congruent stimuli at the first time-point, and larger for the incongruent stimuli at the later time-points) localized to the auditory cortex and the STS.

The lack of congruency effects in the time-windows after 400 ms in this study could be due to the differences in the complexity of the experimental design used, the features of the stimulus material and the timing parameters between the auditory and visual features of the present study and earlier studies. For example, in Arnal et al. (2009) audio-visual combinations of five different syllables were used, which made the identification of congruency more difficult and possibly required further processing steps compared to the current study.

Furthermore, we found no early effects of congruency at N1 m response (75–25 ms following sound onset), which is in line with previous observations (Stekelenburg and Vroomen, 2007). Our results corroborate the assumption that early responses are predominantly sensitive to the stimulus material used for the comparisons. Differences found in the N1 and P2 time-windows were related to suppression effects of audio-visual stimuli compared to audio only stimuli, and not to the direct comparison of congruent and incongruent audio-visual stimuli (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007).

The source localization result of the current study was in line with the sensor-level results in terms of the time-window. However, the clusters at the source level were observed only in the right hemisphere and in a widely spread area encompassing the superior temporal areas as well as the medial and ventral surfaces of the temporal lobe. The superior temporal cortex is roughly in line with that found in Arnal et al. (2009). The widely spread clustering at the source level could be due to methodological limitations. It is important to note that we used a template MRI, and this could have increased the localization error of the brain responses in the source-level analysis. Furthermore, the difference was found in a relatively late time-window and appears quite widespread in time, and the localization of ongoing activation can be more challenging than those of clear time-locked evoked responses. These might explain the differences in the locations of the clusters between the sensor and source level, although we assume they reflect the same effect.

We found no significant effects of familiarity when directly comparing the responses to stimuli that were part of the participants' native language and to stimuli that were not part of their native language. The earlier studies have mostly examined this in auditory oddball experiments investigating deviance detection based on categorical perception of phonemes (e.g., Näätänen et al., 1997; Winkler et al., 1999). First, having equal probabilities of presentation for each stimulus type allows examination of the obligatory sensory responses without overlap from other processes. However, our null results comparing the responses to these stimuli in a passive cross-modal task suggest that the use of either an active comparison involving phonological representations or an identification task which would actively engage these representations is needed to lead to differences in brain activity for familiar and unfamiliar speech stimuli. Second, we examined evoked responses to audio-visual stimuli instead of induced brain activity. It is possible that the familiarity effects could produce brain activity that is not phase-locked to the stimuli. In this case the effect would not be observable in evoked responses. However, we did not have a hypothesis on the specific frequency band or time-window, where the difference in induced activity could be observed. Future studies could examine this in more detail.

The familiarity of speech in our study referred to whether participants perceiving the stimuli had prior knowledge of them, i.e., whether the syllables were present in their native phonology or not. Our stimuli (syllables) were produced by a native Chinese, non-finish speaker. This was required as native Finnish speakers would not be able to naturally produce all stimuli used in

the experiment. Future studies could examine the effect of the speaker identity by using recordings of both native Chinese speaker and native Finnish speaker and how it might interact with the phonological familiarity of speech sounds.

CONCLUSION

Our results show that in the case of audio-visual speech stimuli, congruency has an effect around 300 to 400 ms after the start of voicing. This effect was found in the temporal-parietal brain areas, partly replicating earlier findings. We found no significant differences between Chinese and Finnish speakers in their brain responses depending on the familiarity of the speech stimuli, that is, whether the syllables belonged to the native language or not. This suggests that the congruency effect is a result of a general detection of a mismatch between prediction based on lip movements and the auditory signal.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Ethics Committee of the University of Jyväskylä with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of the University of Jyväskylä.

AUTHOR CONTRIBUTIONS

OK, JH, and WX designed the study, performed the MEG experiments, and analyzed the data. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by the European Union Projects ChildBrain (Marie Curie Innovative Training Networks, # 641652), Predictable (Marie Curie Innovative Training Networks, # 641858), and the Academy of Finland (MultiLeTe #292 466).

ACKNOWLEDGMENTS

The authors would like to thank Chunhan Chiang for his help with data collection and constructive discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00243/full#supplementary-material

FIGURE S1 | Grand average plots at sensor and source level for incongruent and congruent audio-visual stimuli for the two groups. **(a)** Grand averaged waveform for the combined planar gradient (vector sum of the paired orthogonal gradiometer channels) channels grouped (channels included indicated by circles) over the left and right temporal channels in the Chinese (above, *N*=12) and Finnish (below, *N*=12) groups. **(b)** Magnetic field topography and dynamic statistical parametric maps (dSPM) source activation of the grand average evoked responses in the five time-windows investigated in the study (75–125, 150–200, 200–300, 300–400, and 400–600 ms) for the two conditions.

FIGURE S2 | Grand average plots at sensor and source level for unfamiliar and familiar congruent audio-visual stimuli for the two groups. **(a)** Grand averaged waveform for the combined planar gradient (vector sum of the paired orthogonal gradiometer channels) channels grouped (channels included indicated by circles) over the left and right temporal channels in the Chinese (above, *N*=12) and Finnish (below, *N*=12) groups. **(b)** Magnetic field topography and dynamic statistical parametric maps (dSPM) source activation of the grand average evoked responses in the five time-windows investigated in the study (75–125, 150–200, 200–300, 300–400, and 400–600 ms) for the two conditions.

REFERENCES

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 53, 115–121. doi: 10.1016/j.neuropsychologia.2013.11.011

Beauchamp, M. S., Nath, A. R., and Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the mcgurk effect. *J. Neurosci.* 30, 2414–2417. doi: 10.1523/JNEUROSCI.4865-09.2010

Bernstein, L. E., and Liebenthal, E. (2014). Neural pathways for visual speech perception. *Front. Neurosci.* 8:386. doi: 10.3389/fnins.2014.00386

Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Boersma, P., and Weenink, D. (2018). *Praat: Doing Phonetics by Computer [Computer Program] (Version 6.0.37)*. Available at: http://www.praat.org/ (accssed July 07, 2017).

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., and Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cogn. Neurosci.* 16, 805–816. doi: 10.1162/089892904970771

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb.2007.2155

Capek, C. M., Bavelier, D., Corina, D., Newman, A. J., Jezzard, P., and Neville, H. J. (2004). The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 tesla. *Cogn. Brain Res.* 20, 111–119. doi: 10.1016/j.cogbrainres.2003.10.014

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R., Belliveau, J. W., and Lewine, J. D. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* 26, 55–67. doi: 10.1016/S0896-6273(00)81138-1

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Hämäläinen, M. S., and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.* 32, 35–42. doi: 10.1007/BF02512476

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in

cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/JNEUROSCI.1740-07.2007

Helenius, P., Salmelin, R., Connolly, J. F., Leinonen, S., and Lyytinen, H. (2002). Cortical activation during spoken-word segmentation in nonreading-impaired and dyslexic adults. *J. Neurosci.* 22, 2936–2944. doi: 10.1523/JNEUROSCI.22-07-02936.2002

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi: 10.1016/S0893-6080(00)00026-5

Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M.-C., et al. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13995–13999. doi: 10.1073/pnas.241474598

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47–B57. doi: 10.1016/S0010-0277(02)00198-1

Jones, J. A., and Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport* 14, 1129–1133. doi: 10.1097/00001756-200306110-00006

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Kuhl, P. K. (2000). A new view of language acquisition. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11850–11857. doi: 10.1073/pnas.97.22.11850

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Kujala, A., Alho, K., Service, E., Ilmoniemi, R. J., and Connolly, J. F. (2004). Activation in the anterior left auditory cortex associated with phonological analysis of speech input: localization of the phonological mismatch negativity response with MEG. *Cogn. Brain Res.* 21, 106–113. doi: 10.1016/j.cogbrainres.2004.05.011

Lin, F. H., Witzel, T., Ahlfors, S. P., Stufflebeam, S. M., Belliveau, J. W., and Hämäläinen, M. S. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage* 31, 160–171. doi: 10.1016/j.neuroimage.2005.11.054

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

McGurk, H., and Macdonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893. doi: 10.1523/JNEUROSCI.0896-05.2005

Möttönen, R., Schürmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neurosci. Lett.* 363, 112–115. doi: 10.1016/j.neulet.2004.03.076

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385, 432–434. doi: 10.1038/385432a0

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Nishitani, N., and Hari, R. (2002). Viewing lip forms: cortical dynamics. *Neuron* 36, 1211–1220. doi: 10.1016/S0896-6273(02)01089-9

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Parviainen, T., Helenius, P., and Salmelin, R. (2005). Cortical differentiation of speech and nonspeech sounds at 100 ms: implications for dyslexia. *Cereb. Cortex* 15, 1054–1063. doi: 10.1093/cercor/bhh206

Poeppel, D., Yellin, E., Phillips, C., Roberts, T. P. L., Rowley, H. A., Wexler, K., et al. (1996). Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. *Cogn. Brain Res.* 4, 231–242. doi: 10.1016/S0926-6410(96)00643-X

Puce, A., Allison, T., Bentin, S., Gore, J. C., and McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *J. Neurosci.* 18, 2188–2199. doi: 10.1523/JNEUROSCI.18-06-02188.1998

Raij, T., Uutela, K., and Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron* 28, 617–625. doi: 10.1016/S0896-6273(00)00138-0

Salmelin, R. (2007). Clinical neurophysiology of language: the MEG approach. *Clin. Neurophysiol.* 118, 237–254. doi: 10.1016/j.clinph.2006.07.316

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Taulu, S., and Kajola, M. (2005). Presentation of electromagnetic multichannel data: the signal space separation method. *J. Appl. Phys.* 97:124905. doi: 10.1063/1.1935742

Taulu, S., Simola, J., and Kajola, M. (2005). Applications of the signal space separation method. *IEEE Trans. Signal Process.* 53, 3359–3372. doi: 10.1109/TSP.2005.853302

van Atteveldt, N. M., Formisano, E., Goebel, R., and Blomert, L. (2007). Top–down task effects overrule automatic multisensory responses to letter–sound pairs in auditory association cortex. *NeuroImage* 36, 1345–1360. doi: 10.1016/j.neuroimage.2007.03.065

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Vihla, M., Lounasmaa, O. V., and Salmelin, R. (2000). Cortical processing of change detection: dissociation between natural vowels and two-frequency complex tones. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10590–10594. doi: 10.1073/pnas.180317297

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., et al. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology* 36, 638–642. doi: 10.1017/S0048577299981908

Xu, W., Kolozsvari, O. B., Oostenveld, R., Leppänen, P. H. T., and Hamalainen, J. A. (2019). Audiovisual processing of chinese characters elicits suppression and congruency effects in MEG. *Front. Hum. Neurosci.* 13:18. doi: 10.3389/fnhum.2019.00018

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., et al. (2009). Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. *NeuroImage* 46, 226–240. doi: 10.1016/j.neuroimage.2009.01.028

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., and Tohkura, Y. (2005). Effects of language experience: neural commitment to language-specific auditory patterns. *NeuroImage* 26, 703–720. doi: 10.1016/j.neuroimage.2005.02.040

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

II

COHERENCE BETWEEN BRAIN ACTIVATION AND SPEECH ENVELOPE AT WORD AND SENTENCE LEVELS SHOWED AGE-RELATED DIFFERENCES IN LOW FREQUENCY BANDS

by

Orsolya Beatrix Kolozsvári, Weiyong Xu, Georgia Gerike, Lea Nieminen, Tiina Parviainen, Aude Noiray & Jarmo Arvid Hämäläinen, 2021

Neurobiology of Language, *2*(2), 226-253

Available online: https://doi.org/10.1162/nol_a_00033

Neurobiology of Language

an open access journal

Citation: Kolozsvári, O. B., Xu, W., Gerike, G., Parviainen, T., Nieminen, L., Noiray, A., & Hämäläinen, J. A. (2021). Coherence between brain activation and speech envelope at word and sentence levels showed age-related differences in low frequency bands. *Neurobiology of Language*, *2*(2), 226–253. https://doi.org/10.1162/nol_a_00033

DOI:
https://doi.org/10.1162/nol_a_00033

Supporting Information:
https://doi.org/10.1162/nol_a_00033

Received: 17 July 2020
Accepted: 17 February 2021

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Orsolya Beatrix Kolozsvári
orsolya.b.kolozsvari@jyu.fi

Handling Editor:
David Poeppel

The MIT Press

RESEARCH ARTICLE

Coherence Between Brain Activation and Speech Envelope at Word and Sentence Levels Showed Age-Related Differences in Low Frequency Bands

Orsolya B. Kolozsvári[1,2], Weiyong Xu[1,2], Georgia Gerike[1,2,3], Tiina Parviainen[1,2], Lea Nieminen[4], Aude Noiray[5], and Jarmo A. Hämäläinen[1,2]

[1]Department of Psychology, University of Jyväskylä, Finland
[2]Centre for Interdisciplinary Brain Research (CIBR), University of Jyväskylä, Finland
[3]Niilo Mäki Institute, Jyväskylä, Finland
[4]Centre for Applied Language Studies, University of Jyväskylä, Finland
[5]Laboratory for Oral Language Acquisition (LOLA), University of Potsdam, Germany

Keywords: speech perception, development, magnetoencephalography, speech tracking, coherence, auditory responses

ABSTRACT

Speech perception is dynamic and shows changes across development. In parallel, functional differences in brain development over time have been well documented and these differences may interact with changes in speech perception during infancy and childhood. Further, there is evidence that the two hemispheres contribute unequally to speech segmentation at the sentence and phonemic levels. To disentangle those contributions, we studied the cortical tracking of various sized units of speech that are crucial for spoken language processing in children (4.7–9.3 years old, $N = 34$) and adults ($N = 19$). We measured participants' magnetoencephalogram (MEG) responses to syllables, words, and sentences, calculated the coherence between the speech signal and MEG responses at the level of words and sentences, and further examined auditory evoked responses to syllables. Age-related differences were found for coherence values at the delta and theta frequency bands. Both frequency bands showed an effect of stimulus type, although this was attributed to the length of the stimulus and not the linguistic unit size. There was no difference between hemispheres at the source level either in coherence values for word or sentence processing or in evoked response to syllables. Results highlight the importance of the lower frequencies for speech tracking in the brain across different lexical units. Further, stimulus length affects the speech–brain associations suggesting methodological approaches should be selected carefully when studying speech envelope processing at the neural level. Speech tracking in the brain seems decoupled from more general maturation of the auditory cortex.

INTRODUCTION

Brain structure and function continue to develop into early adulthood, with some evidence for different trajectories for the left and right hemispheres (Gogtay et al., 2004; Pang & Taylor, 2000; Parviainen et al., 2019). In adults, important functional differences between the left

and right hemispheres have been demonstrated when processing syllable and phonemic information (e.g., Poeppel, 2014). However, little is known about the development of this functional specialization in children. Functional near-infrared spectroscopy (fNIRS) and magnetic resonance imaging (MRI) have provided evidence for a significant leftward asymmetry for speech processing that is already present from birth (Dehaene-Lambertz et al., 2002; Pena et al., 2003). Drawing on these findings, we used magnetoencephalography (MEG) to examine how hemispheric specialization is reflected in brain responses to various speech units (sentence, words, syllables) and to uncover whether this specialization differs between children and adults. To achieve those goals, we combined two experimental approaches: examining general indices of auditory maturation as reflected in the age-related changes of onset-responses (event-related fields [ERF]) to simple speech sounds alongside examination of word and sentence tracking in different frequency bands, as measured by coherence.

Functional near-infrared spectroscopy (fNIRS):
Functional neuroimaging technique using near-infrared spectroscopy where cerebral hemodynamic responses are measured.

Coherence:
A value that reflects how similar the oscillatory activity present in two signals is.

Event-related potential (ERP):
The brain response to a presented stimulus, measured using electroencephalography (EEG).

N1m:
Auditory evoked response related to N1 response, measured using magnetoencephalography

Previously, long lasting maturational effects have often been studied using the event-related potentials (ERPs) and their magnetic equivalent ERFs to short sounds with EEG and MEG. The auditory ERPs in infancy and in the preschool age show prominent P1 and N2 responses, which as children enter childhood start to become earlier in latency and decrease in amplitude. Additionally, P1 and N2 responses are separated by emerging N1 and P2 responses around the age of 8 to 9 years (Albrecht et al., 2000; Ponton et al., 2000).

Differences in hemispheric maturation rates have also been observed using ERFs. The N1m patterns measured with MEG were more adult-like in 7- to 8-year-olds in the right hemisphere than in the left (Parviainen et al., 2019). This suggests fine-grained developmental trajectories of the different auditory regions with clearly immature patterns of activation in the auditory cortex around early school age (8 to 9 years old).

While studying the event-related potentials and fields in response to individual phonemes and syllables is a useful method to investigate the well-known maturational effects of auditory processing, auditory information in speech spans across multiple timescales encompassing phonemes, syllables, words, and phrases. Multi-time resolution models of speech processing (Ghitza, 2011; Ghitza & Greenberg, 2009; Poeppel, 2003; Poeppel & Assaneo, 2020) propose that speech information is processed and integrated in a hierarchical and interdependent manner by phase alignment or neural entrainment of the involved oscillatory networks in the auditory cortices with different specialization for the left and right auditory areas.

Coherence analysis can be used to study speech perception in these longer speech segments. Coherence is the computation of synchrony between two signals in the frequency domain. The coherence value reflects the consistency of phase difference between two signals (here between the speech envelope and brain activity) at any given frequency. This technique can be used to investigate tracking of the speech signal in the brain, which has been argued to reflect relevant linguistic operations such as parsing and chunking of hierarchical linguistic structures of speech (Bourguignon et al., 2013; Ding et al., 2016; Gross et al., 2013; Molinaro & Lizarazu, 2018; Peelle & Davis, 2012).

Neuronal oscillations in frequency bands present in speech (delta, 1–3 Hz, theta, 4–8 Hz, beta, 15–30 Hz, and low gamma, 30–50 Hz; Poeppel, 2014) have been theorised to provide a basis for parsing the continuous speech signal into different linguistic units (e.g., delta: syllable stress patterns; theta: syllables; beta: onset-rime units; low gamma: phonetic information; Ghitza et al., 2013; Leong & Goswami, 2014; Poeppel, 2014; Poeppel & Assaneo, 2020). In this framework, the linguistic information associated with the different timescales would be then integrated to give the final speech percept. Low frequency cortical activity appears to synchronise to the rhythms of multiple linguistic units (Ding et al., 2016, 2017), while higher

frequencies (such as beta and gamma) may be more sensitive to syntactic and semantic information (Ding et al., 2016). Together, these results suggest that during listening to connected speech, the brain synchronizes cortical rhythms to track the rhythm of the different linguistic units (Ding et al., 2017).

Speech processing involves both left and right auditory cortices (Poeppel, 2003; Poeppel & Assaneo, 2020). In its early stage the representation of the input speech signal has a bilateral symmetry, which then branches out in subsequent processing steps. Left auditory areas have been suggested to sample information from short (20–40 ms) integration windows (Giraud et al., 2007; Poeppel, 2003; Poeppel & Assaneo, 2020), and right areas to sample information from longer (150–200 ms) integration windows (Giraud et al., 2007; Luo & Poeppel, 2007; Poeppel, 2003; Poeppel & Assaneo, 2020). These differences are reflected in oscillatory neuronal activity in different bands (mostly in gamma and theta bands, respectively).

However, changes in brain activity have been reported during childhood with respect to general auditory sound processing as well as more specific speech processing (e.g., Ríos-López et al., 2020; Uhlhaas et al., 2010). Developmental changes in neural synchrony have been demonstrated (for a review, see Uhlhaas et al., 2010) using auditory stimulation (Müller et al., 2009), whereby young children showed reduced synchronisation in the delta and theta (Müller et al., 2009) frequencies compared to adolescents and adults.

There is converging evidence that hemispheric specialisation to different windows of integration for auditory information and speech is present from the first year of life; however, results differ as to which hemisphere shows the strongest response to long speech-like chunks (Telkemeyer et al., 2009, 2011). The developmental pattern of hemispheric dominance for processing spoken sentences seems to shift between brain hemispheres with age. Greater entrainment to speech was found in the left hemisphere compared to right in the theta band with 7-month-old infants (Kalashnikova et al., 2018). However, this specialization was not found in young children between the ages of 4 and 7 years (Ríos-López et al., 2020) in the delta band. Finally, a higher correlation in the right as compared to the left hemisphere between the amplitude envelope of sentences and their corresponding brain responses was found in older 9- to 13-year-old children (Abrams et al., 2008, 2009).

Building on those findings, the current study investigated (a) age-related differences and (b) hemispheric balance in word and sentence tracking in low frequency bands to separate the word to phrasal levels of processing. Based on previous studies on adults and older children, we expected hemispheric differences to already be present in 5- to 9-year-olds in the delta (1–4 Hz) and theta (4–8 Hz) bands with the right hemisphere showing higher coherence than the left hemisphere.

To examine if and how the maturation of the synchrony measures is related to the established maturation of the onset response (reflected in the changes in ERFs to syllables), we compared the coherence values for words and sentences with the age-related changes in the N1m response to syllables. Evoked brain activity to sounds has been shown to change from preschool to school age and to adulthood. While the specific N1m response is absent in early childhood, it seems to emerge at around 8 to 9 years of age and only become fully mature in adulthood (Albrecht et al., 2000; Ponton et al., 2000). If the N1m amplitude has a common underlying maturational mechanism with the speech tracking index, our results should show similar developmental effects. On the other hand, synchronization of brain activity to speech could utilize partly separate brain mechanisms that follow a different developmental trajectory and are affected more by environmental input than by developmental changes reflected by N1m.

We also examined correlations between the processing of speech envelopes and phonological skills. Speech envelope processing has been related to segmentation into syllable and phoneme level elements (Poeppel, 2014). As for phonological skills, broadly defined they include the awareness of various speech units (e.g., phonemes, syllables, words), working memory operations for speech sounds, and access to phonological representations (e.g., Fowler, 1991; Goswami & Bryant, 2016; for a review, see Noiray, Popescu et al., 2019). These are thought to be represented, for example, by rapid naming, phoneme deletion, and speech repetition tasks. Based on this, we hypothesized that speech envelope processing could be linked to phonological skill development (Goswami, 2011).

MATERIALS AND METHODS

Participants

Two age groups participated in the study: typically developing children and young adults. The adults were studying at the University of Jyväskylä, Finland. Table 1 shows the number of participants, mean age and age-range, gender, handedness, and average hearing level for each group. All participants were Finnish native speakers.

The children were recruited via the National Registry of Finland and the adults via email lists of the university. Exclusion criteria at the time of recruitment were head injuries, ADHD or learning difficulties, neurological diseases and medication affecting the central nervous system, or any reported hearing deficits. Children recruited for the study were typically developing and did not present any neurological, cognitive, or language-related deficiency. In addition, the hearing level of the participants was tested using audiometry, with most of them performing at or below 25 dBs for 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz sounds in the left and right ears.

After data collection 13 participants were excluded overall, all of them from the child group. Five were excluded based on the measurement because of too much movement and inability to follow instructions during the recording, two because of noisy data, four because of technical problems (instrumentation failure or software issues), one based on incidental findings during the measurements (based on the neurologist's report), and one because of high amplitude fluctuations in the data.

Enrolment in the study was voluntary; all adults and children participants as well as their parent/caregiver provided written informed consent prior to their participation in the study. Subsequent to the MEG study, all participants received either a movie ticket or a gift card

Table 1. Description of participants

	Children	Adults
# of participants included in the analysis (measured in MEG)	34 (47)	19 (19)
Mean age (*SD*)	7.53 (1.34)	24.80 (3.73)
Age range (Minimum–Maximum; y = years, m = months)	4y8m–9y4m	20y3m–35y2m
Gender ratio (M:F)	18:16	2:17
Handedness (left:both:right)	5:1:28	0:1:18
Average hearing level in DBs (left:right ear)	21.25:21.37	Self-report of normal hearing level

as compensation for their participation. Individual structural MR images were acquired from a private company offering MRI services (Synlab Jyväskylä). T1-weighted 3D-SE images were collected on a GE 1.5 T (GoldSeal Signa HDxt) MRI scanner using a standard head coil and with the following parameters: TR/TE = 540/10 ms, flip angle = 90, matrix size = 256 × 256, slice thickness = 1.2 mm, sagittal orientation.

This study was carried out in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Jyväskylä, Finland.

Behavioural Test Battery

First, we conducted a battery of behavioural tests assessing the children's general cognitive abilities, with an emphasis on language-related skills. For a description of the behavioural tests, see Table 2.

Three different age-appropriate tests, WPPSI-III (Wechsler, 2003a), WISC-IV (Wechsler, 2003b), and WAIS-IV (Wechsler, 2008), were used to measure participants' visuo-spatial reasoning and vocabulary, and two tests, WISC-IV and WAIS-IV, were used for working memory. The motor development of the participants was tested using subtests from the Developmental Neuropsychological Assessment (NEPSY; Korkman et al., 1998), the oro-motor task, and the

Table 2. Description of behavioural test scores

Behavioural measure	Subtest	Mean (*SD*) score reported	Children		Adults
WPPSI-III, WISC-IV, WAIS-IV	Block design	sp	10.00 (3.82)	10.96 (3.15)	11.26 (3.21)
	Vocabulary	sp	11.36 (3.01)	10.46 (3.07)	13.79 (2.02)
	Digit span	sp	x	9.96 (2.15)	11.26 (2.96)
NEPSY	Repetition of nonsense words	sp	10.09 (2.43)		x
	Oro-motor task	sp	10.30 (2.81)		x
NEPSY-II	Visuo-motor task (car and motorcycle)	combined sp	9.42 (3.02)		x
	Phonological processing	sp	11.33 (2.33)		x
	Repetition of sentences	# correct	25.62 (3.08)		x
Rapid automatized naming (RAN)	Objects	time (s)	64.25 (13.66)		34.37 (7.78)
	Letters	time (s)	34.70 (10.64)		18.90 (4.62)
Letter knowledge task		total	22.67 (8.27)		x
Lukilasse	Word reading	percentile	46.38 (40.32)		x
Pseudoword list reading		total	34.45 (13.98)		x
Pseudoword text reading		fluency	99.59 (0.22)		x
Lukilasse	Dictation	percentile	58.10 (39.06)		x

Note. sp: standard point; *SD*: standard deviation; WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 2003a); WISC-IV: Wechsler Intelligence Scale for Children (Wechsler, 2003b); WAIS-IV: Wechsler Adult Intelligence Scale (Wechsler, 2008); NEPSY: Neuropsychological Assessment test battery I (Korkman et al., 1998); NEPSY II (Korkman et al., 2008); RAN: Rapid automatized naming (Denckla & Rudel, 1976); Lukilasse (Häyrinen et al., 1999).

NEPSY II visuo-motor task (Korkman et al., 2008). Participants' phonological processing was tested using the NEPSY II subtest. To assess speed of lexical retrieval, the Rapid automatized naming (RAN; Denckla & Rudel, 1976) Objects and Letters subtests were used. To measure memory for sentences, the NEPSY II Sentence Repetition subtest was used.

Reading skills were tested using the word reading task from the Lukilasse test battery (Häyrinen et al., 1999), the pseudoword reading task adapted from TOWRE (Torgesen et al., 1999), and the pseudoword text reading task (Eklund et al., 2015).

For a detailed description of the behavioural tests, see Supplementary Material 1 (supporting information can be found online at https://www.mitpressjournals.org/doi/suppl/10.1162/nol_a_00033).

Stimuli

Three types of stimuli characterizing various temporal and linguistic structures were used for the speech tracking task: syllables, words, and sentences. Syllables varied in consonants' place of articulation (moving from front to back: bilabial stop /p/, dental stop /t/, and palatal stop /k/), while the vowel remained identical (/a/).

Each syllable was presented 18 times (total of 54 syllable presentations), and words starting with the same syllables (18 words for each syllable, total of 54 words), as well as 54 sentences, each starting with one of the word category stimuli. For a description and exemplars of the stimuli, see Table 3.

All words were common, everyday nouns. The words were 2 to 3 syllables long. Sentences were composed of 3 to 4 words and always started with a noun followed by a form of the verb "to be" in the present tense. Stimuli were chosen with the help of an expert developmental linguist. The stimuli were produced by a female native Finnish speaker. All stimuli were separate, unique tokens produced separately. Stimuli were recorded using a 44 kHz sampling frequency, 32-bit quantisation in a professional recording studio. The sound files were cut into individual segments using Praat (Boersma & Weenink, 2018).

The same syllables and words were used for each stimulus type to get comparable onset evoked brain responses. To see the list of stimuli used, see Supplementary Material 2.

Procedure

Experimental design

Each speech tracking trial consisted of a fixation cross in the middle of the screen for 500 ms, then an exclamation mark appeared in the same space for 1,000 ms signalling that a sound is going to come soon, followed by the fixation cross for 750 ms. The auditory stimuli were then

Table 3. Description of stimuli

Stimulus type	Average duration (ms)	*SD*	Range (ms)	Exemplars: Finnish	English translation
Syllable	209.33	25.58	185–236	ka, ta, pa	
Word	574.54	103.22	352–797	kala, paju, talo	*fish, willow, house*
Sentence	1,438.54	240.49	1,039–2,051	Kala on akvaariossa. Paju on taipuisa puu. Talo on aivan uusi.	*The fish is in the aquarium.* *A willow is a flexible tree.* *The house is brand new.*

presented via earphones, with the fixation cross on the screen. The fixation cross remained on the screen for 750 ms after the end of sound. This was followed by a still image of a parrot appearing for 1,250–4,250 ms (presentation duration depended on the type of stimuli heard) which provided the cue for the participants to repeat the previously heard stimuli aloud (see Figure 1).

Participants were instructed to first listen to a speech sequence (i.e., a syllable, a word, or a sentence) and to repeat it after seeing the visual cue on the screen (a parakeet). The visual stimuli were presented on a black background with white standard characters (a cross for fixation and an exclamation mark alerting to the auditory stimuli) in Times New Roman font and a font size 64. The bird stimuli were 9 × 15 cm in size on the projection screen. Here only the time-window of the auditory stimulus presentation was analysed.

Participants were first given instructions and 6 practice trials (2 of each type of stimuli, presented in random order). In the actual experiment 162 stimuli (the 3 syllables repeated 18 times each, 54 words, and 54 sentences) were presented in random order.

Stimuli were presented in 9 blocks, with 2 longer breaks after 3 blocks and shorter breaks (duration determined by the participant) in between the blocks. Three blocks lasted approximately 8 min, and it took approximately 30 min to complete the task, instructions and practice included.

The task was embedded in a child-friendly narrative to stimulate children's attention and motivation to complete the task. Participants were told they are teaching 3 parrots how to "speak." Their task was to wait for the parrot to start listening (when the cue appeared on the screen) and their instructions included keeping eye-contact with the parrot to make sure the bird is paying attention (to minimize movement-related artefacts in the recording). Furthermore, they were asked to repeat what they heard at a normal speaking loudness (i.e., to not mumble the syllables, words, or sentences) since the parrots will "not be able to learn if they don't hear the speech properly" (to be able to record the production as clearly as possible). This also ensured the children were fully engaged in the task. Correct production was on average for children 88.41% and for adults 97.86%. At the end of each third block (i.e., before the longer breaks and the end of the test), the parrots "repeated" some of the heard sounds, which were new sounds created by raising the pitch of the original stimuli, to give the impression that the parrots were the ones repeating them. The first and second time it was the syllables, while at the end of the MEG recording it was one sentence from each syllable type.

Participants sat in a magnetically shielded, sound attenuated room under the MEG helmet, at a 68 degree position. The stimuli were presented through insert earphones (Rotel RA-1570

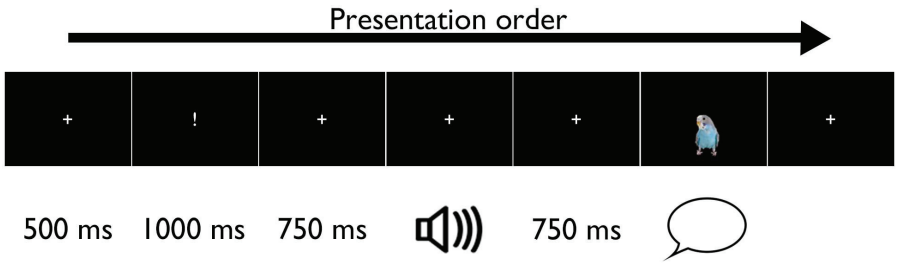


Figure 1. Schematic representation of one trial of the experimental paradigm. Data analysis was focused on the time-window during the stimuli presentation, indicated by the picture of the loudspeaker.

system; eartips were ER3-14B for children and ER3-14A for adults) at a comfortable loudness level. The participants sat 1 m from the projection screen. During measurement, a research assistant was also present in the room when necessary for the children. Presentation software (version 18.1; Neurobehavioral Systems, Inc., Albany, CA, USA) was used to present the stimuli, running on a Microsoft Windows computer (sound card: Sound Blaster Audigy RX; video card: NVIDIA Quadro K5200). Measurements were video monitored to make sure participants were paying attention and doing the task.

MEG recording

306-channel (102 magnetometers and 102 planar gradiometer pairs) MEG data were recorded in a magnetically shielded room using the Elekta Neuromag® TRIUX™ system (Elekta AB, Stockholm, Sweden) at the Centre for Interdisciplinary Brain Research, at the University of Jyväskylä, Finland.

The head position in relation to the sensors in the helmet was monitored continuously with five digitised head position indicator (HPI) coils attached to the scalp. Three HPI coils were placed on the forehead and one behind each ear. The position of the HPI coils was determined by three anatomic landmarks (nasion, left and right preauricular points) using the Polhemus Isotrak digital tracker system (Polhemus, Colchester, VT) at the beginning of the recording. An additional set of points (>100) randomly distributed over the scalp was also digitised. Electrooculogram was recorded with two electrodes attached diagonally close to the left and right eyes and one ground electrode attached to the collar bone.

The sampling rate of the recording was 1000 Hz and a 0.03–330 Hz online band-pass filter was used.

Data Analysis

Pre-processing

All data were pre-processed using the temporal extension of the signal space separation method with buffers of 30 s (Taulu & Kajola, 2005; Taulu et al., 2005) in Maxfilter 2.2™ (Elekta AB) to remove external interference and correct for head movements. Bad channels were identified and reconstructed by the Maxfilter program. Head position was estimated in 200 ms time-windows and 10 ms steps for movement compensation. Data were saved in three separate files containing three recording blocks. Initial head position of the first file was used for transforming the head position to the same position across the files.

Data were pre-processed using independent component analysis (ICA) using fastICA algorithm (Hyvärinen & Oja, 2000) to remove eye blinks, horizontal eye movements, and cardiac artifacts in MNE Python (0.16.2; Gramfort et al., 2013), and the separate MEG recordings were concatenated. The rest of the data analysis was done in the FieldTrip toolbox (Oostenveld et al., 2011) in MATLAB R2016 (https://www.mathworks.com/).

The continuous MEG recording was epoched to 100 ms before and 1,000 ms after the onset of sound in the syllable stimuli (for analysis of the evoked fields), and 100 ms before the onset of sound and 100 ms after the end of the sound in the word and sentence stimuli (for analysis of the frequency contents). Epochs were visually inspected and bad trials were rejected, with an average of 2.18% of epochs rejected for the children and 0.78% of epochs rejected for the adults. Data were low-pass filtered at 45 Hz. The epoched data was baseline corrected using the 100 ms preceding the onset of the stimuli.

We examined the data using two approaches. First, to examine how closely the brain follows the frequency contents of the speech signal, coherence was calculated between the MEG signal and the speech signal. Second, the evoked fields to the syllable stimuli were calculated to examine possible associations between the relatively well-known developmental changes of the evoked fields (particularly responses around 100 ms) and the coherence measures.

Coherence measures

We conducted coherence analysis at different frequency bands to investigate how brain activity changes while tracking the speech envelope of stimuli with different durations at different ages.

The speech stimuli were downsampled to 1000 Hz from 44.1 kHz. The absolute hilbert envelope was calculated for each stimulus separately in MATLAB (abs(hilbert(audiosignal))). The envelope was then appended to the epoched MEG data as a 307th channel.

Earlier studies looking into cross-correlations between the speech envelope and brain activity removed the first 250 ms of brain activity to avoid the onset evoked response (e.g., Abrams et al., 2008). However, the effects of the onset response on the coherence measures have not been reported before. Therefore, we performed the coherence analyses two times: first, for data without the evoked response, second, for the whole epoch length (see Supplementary Material 3). As shown in Supplementary Material 3, this did not have a large effect on the results. The results reported in the main text are based on analysis conducted using data where the evoked response was removed.

Frequency analysis of the data was done to compute the cross and power spectra of the trials using a multitaper frequency transformation method, where the maximum trial length was rounded up to the next power of 2 (cfg.pad = nextpow2) using FieldTrip's ft_freqanalysis function, between 1 and 45 Hz with a 3 Hz smoothing and keeping the trials. This was followed by coherence analysis between the sound envelope and the MEG data using the ft_connectivityanalysis function.

Further, to see if the coherence between the brain and speech signals was significant at the individual level, we calculated 1,000 permutations of coherence, where the sound envelopes were randomly paired with the brain activity of another sound envelope, then compared with the original coherence value. For each participant at least one channel of the original speech–brain pair showed a coherence value larger than 95% of the permuted values (for visualization, see Supplementary Material 4).

To examine the effect of the stimulus length on the coherence values, we first checked the lengths of trials for word stimuli. Second, we cut out the end of the sentence stimuli to be of equal length with the word stimuli (i.e., the initial part of the sentence was used in the new analysis). We then recalculated the coherence between these shortened sentence stimuli and brain activity (see Supplementary Material 5). The results showed that shortened trials also had larger coherence values in both frequencies.

For further analyses, channels were grouped together by hemispheres (see Supplementary Material 6 for grouping of sensors across hemispheres). In the statistical analysis, data from magnetometers were averaged based on hemispheres and separated into two frequency bands: 1–3.5 Hz (delta), 4.5–8 Hz (theta).

For children, source reconstruction was based on their own T1 MRIs, while for adults the fsaverage brain template from Freesurfer (RRID: SCR_001847; Martinos Center for Biomedical Imaging, Charlestown, MA, USA) was used. Coregistration was done between the digitized head points and the brain template with 3-parameter scaling.

Source analysis was done using the ft_sourceanalysis, using the dynamic imaging of coherent sources method (Gross et al., 2001) between 1–8 Hz for every 0.5 Hz. The resulting coherence values were then averaged together according to the frequency band defined—delta band: 1–3.5 Hz, theta band: 4.5–8 Hz. The coherence values were then extracted based on the Desikan-Killiany Atlas (Desikan et al., 2006). Two regions of interest (ROIs) were selected a priori: the temporal area, including the superior temporal, transverse gyrus, and bank of superior temporal sulcus areas; and the inferior frontal area, including the pars opercularis, pars orbitalis, pars triangularis, and precentral areas (see, e.g., Molinaro et al., 2016).

Identification of responses around 100 ms to syllable stimuli and correlation with coherence values for the word and sentence stimuli

Global mean field power (GMFP): A measure used to characterize global MEG activity.

Trials for syllables were averaged together for each participant separately. Global mean field power (GMFP) was calculated for each group separately, and the time-window of auditory response was identified. Based on the GMFP peaks, the time-windows were defined by automatically finding the peak near 100 ms, and using a time-window of +/−25 ms for each hemisphere and group. Thus, the time-windows used in further analyses were 94–144 ms in the left hemisphere and 92–142 ms in the right hemisphere for adults, and 114–164 ms in the left and 113–163 ms in the right hemisphere for children. We averaged together the squared values from the temporal channels from the two hemispheres separately. The values were then correlated with the coherence values in the left and right hemispheres.

Topography of the averages was visually inspected to confirm the correct N1m response pattern or its equivalent in children. Earlier ERP/ERF research has shown that the N1m pattern reflects current direction towards inferior-posterior direction, and the opposite direction was referred to as P1m/P1m-like response. Indeed, averaging or grouping together opposite field patterns would obscure the outcome, and these patterns are likely to reflect distinct processes. Responses were separated based on hemisphere, then squared. The squared amplitude of the response was then correlated with the coherence values from the left and right hemispheres for the delta and theta bands.

A missing response could be due to noisy ERF signal. Therefore, signal-to-noise ratio was calculated by averaging and squaring together the baseline periods of the ERFs (time-window: −100–0 ms), and used as a covariate in separate ANOVAs to ensure that it was not the source of the differences found at sensor level. We found that it did not affect the significant effects.

Source analysis of the ERFs was done using ft_sourceanalysis, using the minimum-norm estimate (MNE) method (Hämäläinen & Ilmoniemi, 1994), and the power of each source component was calculated using ft_sourcedescriptives and used in the statistical analyses.

MNE source estimates were calculated for ERFs, and source power waveforms were extracted based on the Desikan-Killiany Atlas (Desikan et al., 2006). One ROI was selected a priori from the temporal areas around the auditory cortex including the temporal area, including the superior temporal, transverse gyrus, and bank of superior temporal sulcus areas, postcentral and supramarginal areas. The same time-windows were used as in the sensor level analysis. The literature clearly defines the sources of the N1m response near auditory cortex (Parviainen et al., 2019; Ponton et al., 2002). The ROIs for the coherence value analysis and ERFs were therefore expected to be slightly different with the former encompassing more frontal regions (Molinaro et al., 2016).

Statistical analyses

The age, hemisphere, and stimulus type effect on the coherence values for the different frequency bands were analysed in SPSS (IBM SPSS Statistics v. 24) using a 2 (Type: Word,

Sentence) × 2 (Hemisphere: Left, Right) × 2 (Group: Children, Adults) repeated measures mixed ANOVA at both sensor and source levels. Significant interactions were further examined using independent samples *t* tests, and paired samples *t* tests where groups were involved in the interaction.

Pearson correlation was calculated between the coherence values at source level and the children's ages in years rounded to months.

The averaged and squared responses around N1m to syllables were compared in a 2 (Hemisphere: Left, Right) × 2 (Group: Children, Adults) repeated measures mixed ANOVA. Further, Pearson correlation coefficients were calculated to examine the relationship between the peak amplitudes of the auditory responses around 100 ms and coherence values.

Pearson correlation coefficients were calculated to examine the relationship between the scores of three behavioural tests (RAN: objects subtests, NEPSY: Phonological processing and Sentence repetition subtests) and coherence values at source level.

Alpha level was 0.05. False discovery rate (FDR) correction for multiple comparisons was calculated for each analysis.

RESULTS

Coherence Between Brain and Speech Signals for Words and Sentences

Sensor level

The results of the repeated measures ANOVA revealed first, that adults had the largest coherence values (see Tables 4 and 5, Group main effect; and Figure 2). Second, larger coherence values were observed for words as compared to sentences for both delta and theta frequency bands (see Tables 4 and 5, Type main effect; Figure 3). Further, we found that coherence values in the delta band were larger in the left compared to right hemispheres in adults' brain responses and that adults had larger coherence values in the left hemisphere than children (see Table 4, Hemisphere × Group interaction; Figure 4).

Adults showed larger coherence values in the left hemisphere compared to the right hemisphere in the delta band ($t(18) = 5.437$, $p = 0.000$) when compared in a paired samples *t* test.

Table 4. Results of repeated measures mixed ANOVA for the delta frequency band at sensor level

Delta	Main effects and interactions	*df*	*F* value	*p* value	partial η^2
	Type	**1,51**	**227.754**	**0.000**	**0.817**
	Hemisphere	**1,51**	**11.631**	**0.001**	**0.186**
	Group	**1,51**	**12.739**	**0.001**	**0.200**
	Type × Group	1,51	0.295	0.589	0.006
	Hemisphere × Group	**1,51**	**5.822**	**0.019**	**0.102**
	Type × Hemisphere	1,51	3.670	0.061	0.067
	Type × Hemisphere × Group	1,51	0.996	0.323	0.019

Note. Bold values remained significant after false discovery rate correction.

Table 5. Results of repeated measures mixed ANOVA for the theta frequency band at sensor level

Theta	Main effects and interactions	*df*	*F* value	*p* value	partial η^2
	Type	**1,51**	**259.307**	**0.000**	**0.836**
	Hemisphere	1,51	0.171	0.681	0.003
	Group	**1,51**	**14.089**	**0.000**	**0.216**
	Type × Group	1,51	0.836	0.365	0.016
	Hemisphere × Group	1,51	0.132	0.718	0.003
	Type × Hemisphere	1,51	0.017	0.896	0.000
	Type × Hemisphere × Group	1,51	0.051	0.822	0.001

Note. Bold values remained significant after false discovery rate correction.

Children's coherence values did not differ significantly in the two hemispheres ($t(33) = 0.730$, $p = 0.470$). Further, independent samples t tests showed that adults had larger coherence values in the delta band in the left hemisphere compared to children ($t(51) = -4.044$, $p = 0.000$), and the groups did not differ significantly in their coherence values in the right hemisphere ($t(51) = -1.386$, $p = 0.172$).

Source level

Similar to the sensor level, the results of the repeated measures ANOVA revealed that adults had the largest coherence values (see Tables 6 and 7, Group main effect; Figure 5) at source level. Second, larger coherence values were observed for words compared to sentences for both delta and theta frequency bands (see Tables 6 and 7: Type main effect; Figure 6). Third, we found that the adults had larger coherence values compared to children in the delta

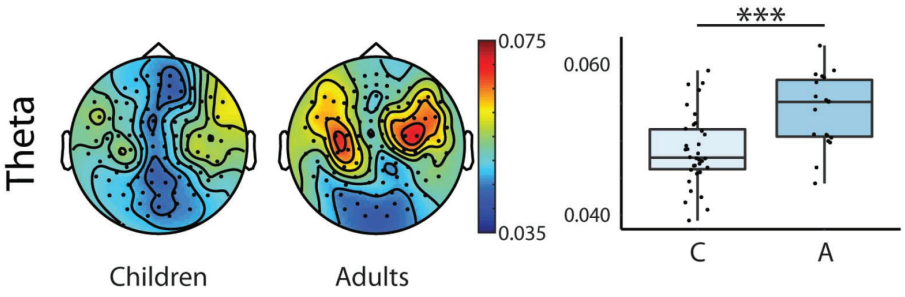


Figure 2. Topographic distribution of the coherence values and box plots of the theta frequency band showing Age main effect in the repeated measures mixed ANOVA for the two groups (Children, $N = 34$; Adults, $N = 19$) collapsed across hemispheres and stimulus types. Topographies: Warmer colours reflect higher coherence between the stimuli envelope and the brain data. Right box plots: Bold lines denote the median of the coherence values; the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Light blue boxes show average coherence for children, dark blue boxes for adults (C = Children, A= Adults). (*** < 0.001)

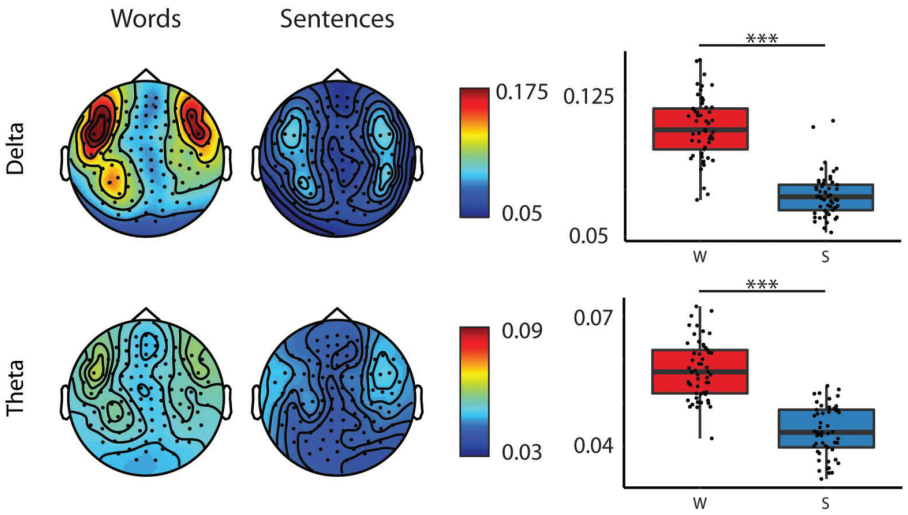


Figure 3. Topographic maps of coherence values of the two different frequency bands to word (W) and sentence (S) stimuli and box plots of averaged coherences for the delta and theta frequency bands collapsed across hemispheres and ages. Topographies: Warmer colours reflect higher coherence between the stimuli envelope and the brain data. Boxplots: Bold lines denote the median of the coherence values; the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Red boxes represent average coherence values for words, and blue boxes for sentences. (*** < 0.001)

band in the temporal region in case of both words and sentences, and that adults had larger values for words than sentences (See Table 6, Type × Group interaction; Figure 5).

Post hoc independent samples *t* tests revealed that adults had significantly larger coherence values for words ($t(51) = -4.467$, $p = 0.000$) and for sentences ($t(51) = -1.598$, $p = 0.002$)

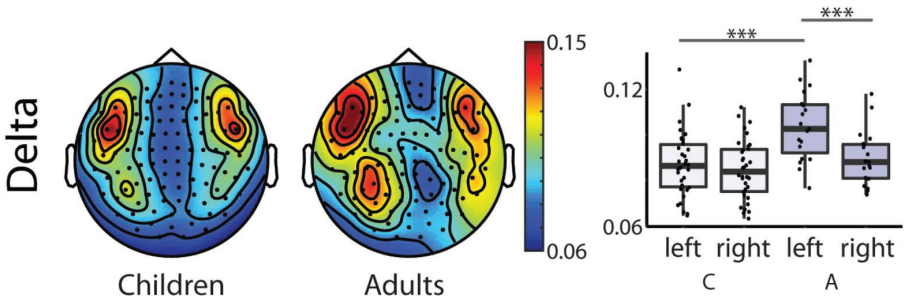


Figure 4. Topographic distribution of the coherence values and box plots of the delta frequency band showing a Hemisphere × Group interaction in the repeated measures mixed ANOVA (Children, $N = 34$; Adults, $N = 19$) collapsed across stimulus types. Topographies: Warmer colours reflect higher coherence between the stimuli envelope and the brain data. Right box plots: Bold lines denote the median of the coherence values; the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Light purple boxes show average coherence for children, dark purple boxes for adults (C = Children, A = Adults). (*** < 0.001)

Table 6. Results of repeated measures mixed ANOVA for the delta frequency band at source level in the two regions of interests

Delta – Temporal region	Main effects and interactions	*df*	*F* value	*p* value	partial η^2
	Type	**1,51**	**12.939**	**0.001**	**0.202**
	Hemisphere	1,51	5.266	0.026	0.094
	Group	**1,51**	**13.897**	**0.000**	**0.214**
	Type* × *Group	**1,51**	**6.519**	**0.014**	**0.113**
	Hemisphere × *Group*	1,51	1.727	0.195	0.033
	Type × *Hemisphere*	1,51	0.182	0.672	0.004
	Type × *Hemisphere* × *Group*	1,51	0.996	0.323	0.019
Delta – Inferior-frontal region	**Main effects and interactions**	***df***	***F* value**	***p* value**	**partial η^2**
	Type	**1,51**	**9.143**	**0.004**	**0.152**
	Hemisphere	1,51	0.014	0.907	0.000
	Group	**1,51**	**13.476**	**0.001**	**0.209**
	Type × *Group*	1,51	1.291	0.261	0.025
	Hemisphere × *Group*	1,51	1.960	0.168	0.037
	Type × *Hemisphere*	1,51	0.737	0.395	0.014
	Type × *Hemisphere* × *Group*	1,51	0.037	0.849	0.001

Note. Bold values remained significant after false discovery rate correction.

compared to children. Paired samples *t* tests revealed that adults also had significantly larger coherence values for words compared to sentences ($t(18) = 3.200$, $p = 0.005$), and children's coherence values did not differ significantly between words and sentences ($t(33) = 1.000$, $p = 0.325$).

Because the child group spanned a relatively large age range (4.7–9.3 years), we examined whether age was linearly related to changes in coherence values. We did not find any significant correlation between the observed coherence values and age (see Table 8).

Evoked Responses to Syllables

Sensor level

The averaged evoked responses' topographies were typical of the N1m response in adults. In children the topography reminiscent of the N1m was slightly later in time in the right hemisphere. The left hemisphere showed a less clear pattern for children (see Figure 7). The topographies were also examined individually.

The averaged squared responses were compared in a 2 (Hemisphere: left, right) × 2 (Group: Children, Adults) repeated measures mixed ANOVA (see Table 9). No significant differences were found.

Table 7. Results of repeated measures mixed ANOVA for the theta frequency band at source level in the two regions of interests

Theta – Temporal region	Main effects and interactions	*df*	*F* value	*p* value	partial η^2
	Type	**1,51**	**44.799**	**0.000**	**0.468**
	Hemisphere	1,51	5.850	0.019	0.103
	Group	**1,51**	**6.849**	**0.012**	**0.118**
	Type × Group	1,51	2.131	0.151	0.040
	Hemisphere × Group	1,51	0.743	0.393	0.014
	Type × Hemisphere	1,51	0.253	0.617	0.005
	Type × Hemisphere × Group	1,51	0.190	0.665	0.004
Theta – Inferior-frontal region	**Main effects and interactions**	***df***	***F* value**	***p* value**	**partial η^2**
	Type	**1,51**	**50.638**	**0.000**	**0.498**
	Hemisphere	1,51	0.540	0.465	0.011
	Group	**1,51**	**14.688**	**0.000**	**0.224**
	Type × Group	1,51	0.001	0.977	0.000
	Hemisphere × Group	1,51	0.865	0.357	0.017
	Type × Hemisphere	1,51	0.398	0.531	0.008
	Type × Hemisphere × Group	1,51	0.003	0.960	0.000

Note. Bold values remained significant after false discovery rate correction.

The averaged squared values were then correlated with the corresponding hemisphere's coherence values in the frequency bands. No significant correlations were found. (For the table of correlation coefficients and *p* values, see Supplementary Material 7, Table 7.1.)

Topography of the averages was visually inspected then to confirm the correct N1m response pattern in each participant, for left and right hemispheres separately.

The N1m response in the left hemisphere was observed in 4 (11.76%) children, with an average latency of 130 ms, and 17 (89.47%) adults, with an average latency of 104 ms. Six (17.65%) children with an average latency of 151 ms showed an activation pattern with an opposite current direction to the adult-like N1m.

The N1m response in the right hemisphere was observed in 8 (23.53%) of the children's evoked responses, with an average latency of 141 ms, and in 17 (89.47%) of the adults' evoked responses, with an average latency of 105 ms. Five (14.71%) of the children with an average latency of 143 ms showed an activation pattern with an opposite current direction to the adult-like N1m.

To examine whether the N1m amplitude and coherence values in the delta and theta bands would follow a similar developmental pattern, correlations were calculated to quantify the possible developmental relationship between the measures. Coherence values were plotted

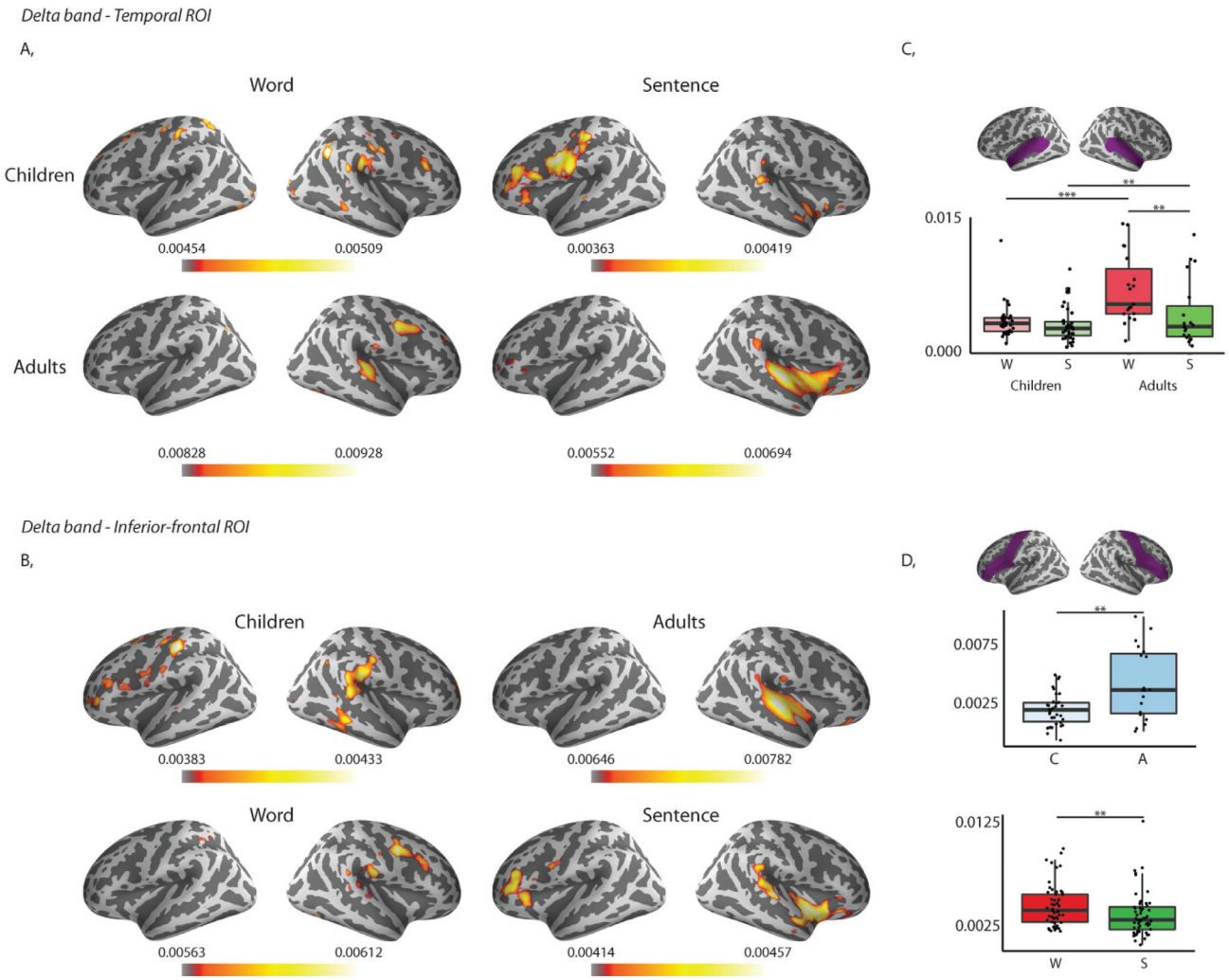


Figure 5. Left panels: Grand average of source level coherence values of children and adults to words and sentences. (A) In the delta frequency band in the temporal region of interest. (B) In the delta frequency band in the inferior-frontal region of interest; top row: grand averages of children and adults; bottom row: grand averages to words and sentences. Warmer colours reflect higher coherence between the stimuli envelope and the brain data. Right panels: Region of interest highlighted in purple (as defined in the Desikan-Killiany Atlas; Desikan et al., 2006). (C) Box plots of averaged coherence values in the delta frequency band in the temporal region collapsed across hemispheres. (D) Box plots of averaged coherence values in the delta frequency band in the inferior-frontal region collapsed across hemispheres and ages (top) or stimulus types (bottom). Bold lines denote the median of the coherence values; the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. W = words, S = sentences, C = children, A = adults. (*** < 0.001, ** < 0.01)

against the N1m responses in the child and adult groups (see Figure 8). No significant correlations were found between N1m amplitude to syllables and delta and theta coherence values to words and sentences in either the left or right hemispheres after correction for multiple comparisons. (For the table of correlation coefficients and *p* values, see Supplementary Material 7, Table 7.2.)

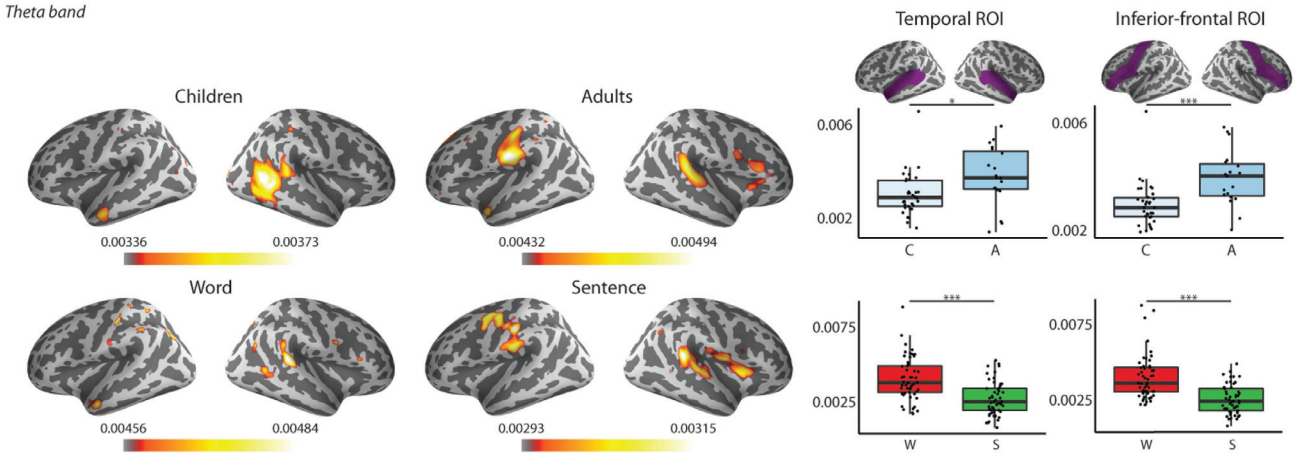


Figure 6. Left: Grand average of source level coherence values in the theta frequency band. Warmer colours reflect higher coherence between the stimuli envelope and the brain data. Top row: Grand averages of children and adults. Bottom row: Grand averages to words and sentences. Right: Region of interest highlighted in purple (as defined in the Desikan-Killiany Atlas; Desikan et al., 2006) and box plots of averaged coherences of the ROIs in the theta frequency band collapsed across hemispheres and ages (top) or stimulus types (bottom). Bold lines denote the median of the coherence values; the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Top plot: Light blue boxes show average coherence for children (C), and dark blue boxes for the adults (A). Bottom plot: Red boxes represent average coherence values for words (W), and green boxes for sentences (S). (*** < 0.001, * < 0.05)

Table 8. Results of correlations between the coherence values at source level and age in the children group

		Correlation coefficient	Sig	*N*
Delta	Temp	0.049	0.785	34
	Inf-front	−0.036	0.840	34
Theta	Temp	0.165	0.352	34
	Inf-front	0.056	0.752	34

Note. Sig = significance. Temp = Temporal. Inf-front = Inferior-frontal.

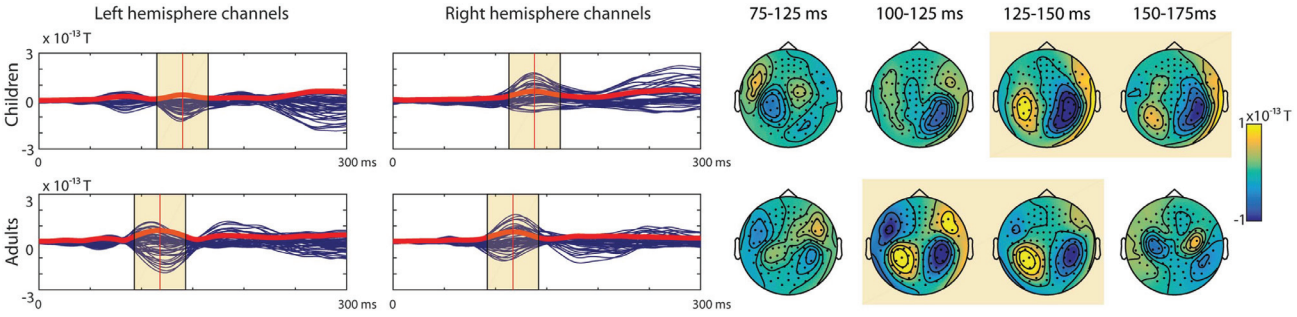


Figure 7. Blue butterfly plots of the group-averaged magnetometers with the global mean field power (GMFP) (red line) and topographic maps for the evoked responses to the syllable stimuli for the two age groups (Children, *N* = 34; Adults, *N* = 19). The yellow boxes highlight where the auditory response was expected in the groups based on the GMFP.

Table 9. Results of repeated measures mixed ANOVA for the averaged squared responses based on the GMFP peaks at sensor level

Main effects and interactions	*df*	*F* value	*p* value	partial η^2
Hemisphere	1,51	0.013	0.531	0.000
Group	1,51	3.761	0.058	0.069
Hemisphere × Group	1,51	0.414	0.523	0.008

Source level

The responses were compared in a 2 (Hemisphere: left, right) × 2 (Group: Children, Adults) repeated measures mixed ANOVA (see Table 10 and Figure 9). No significant differences were found.

The averaged power was then correlated with the corresponding hemisphere's coherence values in the frequency bands. No significant correlations were found. (For the table of correlation coefficients and *p* values, see Supplementary Material 7, Table 7.3.)

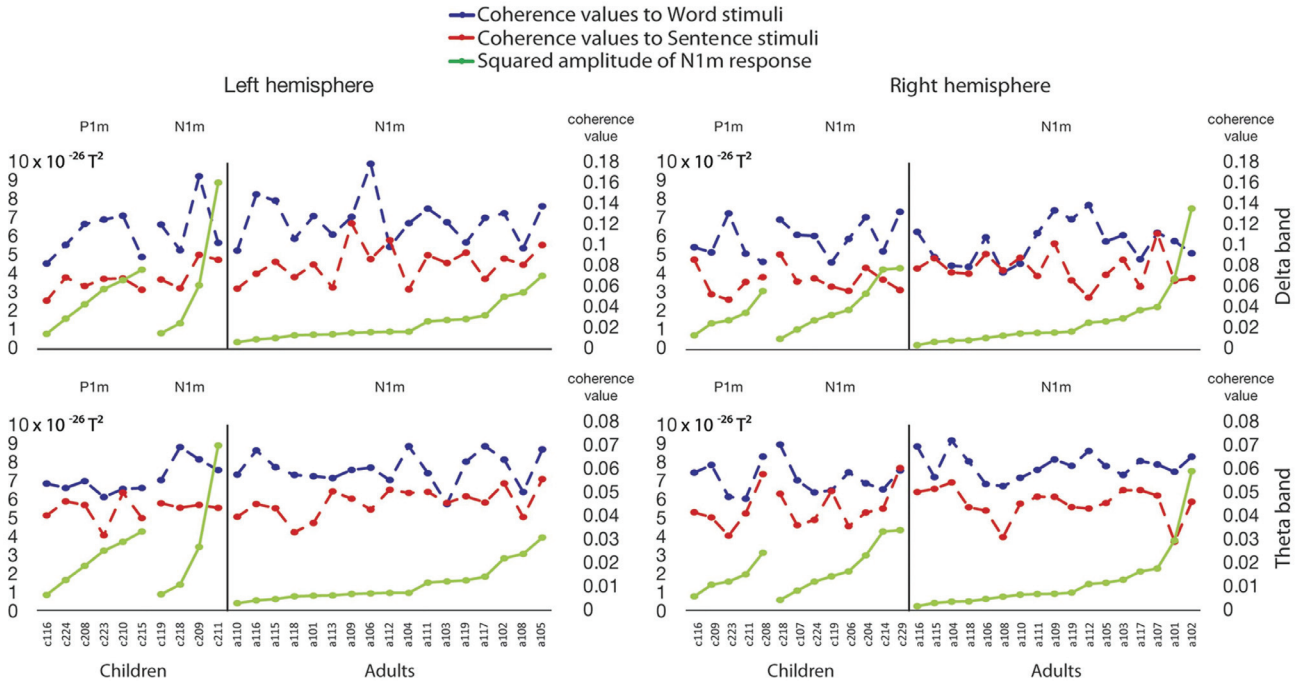


Figure 8. Line plots for coherence values in the left and right hemispheres to word and sentence stimuli in comparison to the squared amplitude of the N1m (T^2) in the participants who showed a P1m or N1m pattern in their ERF responses. In each plot, the scale on the left side shows squared amplitude of N1m response, and the scale on the right side shows coherence values. Blue dashed line: coherence values to word stimuli; red dashed line: coherence values to sentence stimuli; green solid line: the squared N1m amplitude. Top row plots show coherence values for the delta band, left and right hemispheres respectively; bottom row plots show coherence values for the theta band, left and right hemispheres respectively. Values are organized in order of the squared amplitudes.

Table 10. Results of repeated measures mixed ANOVA for the averaged squared responses based on the GMFP peaks at source level

Main effects and interactions	*df*	*F* value	*p* value	partial η^2
Hemisphere	1,51	0.3976	0.531	0.008
Group	1,51	0.105	0.747	0.002
Hemisphere × *Group*	1,51	3.469	0.068	0.064

Correlations of Source Level Coherence with Behavioural Scores

Behavioural scores in the Phonological processing and Sentence repetition tasks did not correlate with coherence values from the delta and theta bands. RAN objects did correlate inversely with both frequency bands and both ROIs (see Table 11 and Figure 10), but when age was controlled for, the correlation was no longer significant (see Table 12).

Coherence values were correlated with the child groups scores on NEPSY's phonological processing task and sentence repetition task (see Tables 13 and 14). No significant correlations were found.

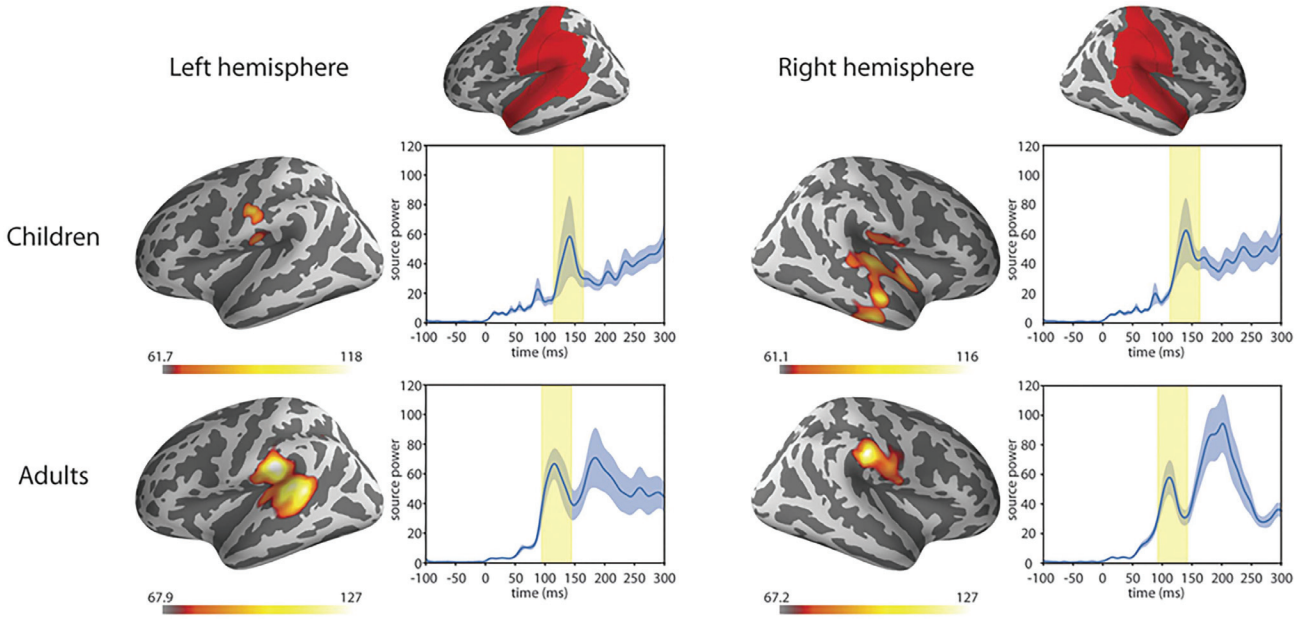


Figure 9. Grand averaged source level ERFs to syllables. Warmer colours reflect higher source power of the event-related field. The red areas highlighted were included in the region of interest (as defined in the Desikan-Killiany Atlas; Desikan et al., 2006). Right panels show the average source waveform (MNE estimate) extracted from the brain regions. The blue shading represents the standard error of the mean, and the yellow shading shows the time-windows used for the N1m response.

Table 11. Correlations between performance on RAN of objects (time in seconds) and the coherence values from the two regions of interest in the delta and theta frequency bands at source level

		Correlation coefficient	Sig	*N*
Delta	Temp	**−0.377**	**0.005**	**53**
	Inf-front	**−0.350**	**0.010**	**53**
Theta	Temp	**−0.292**	**0.034**	**53**
	Inf-front	**−0.330**	**0.016**	**53**

Note. Sig = significance. Temp = Temporal. Inf-front = Inferior-frontal.

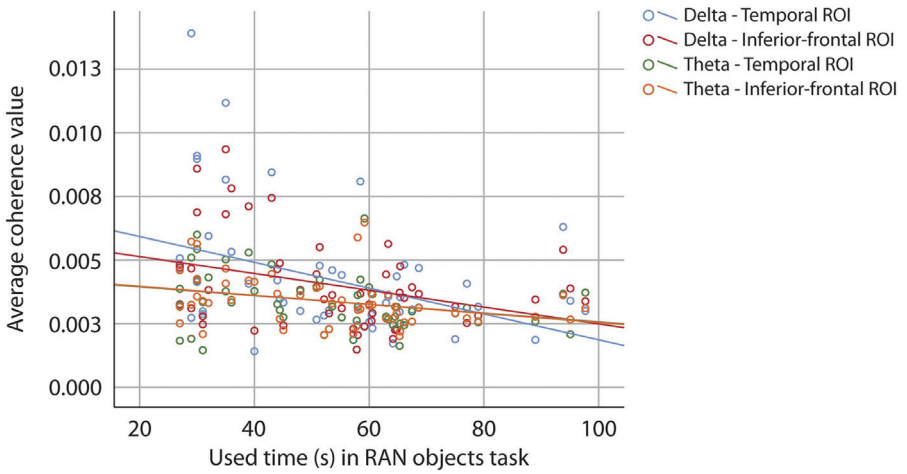


Figure 10. Scatter plot of correlation between performance on RAN objects and coherence values at source level. Blue dots and line represent coherence values from the temporal region of interest, and red dots and line represent values from the inferior-frontal region of interest in the delta frequency band. Green dots and line represent values from the temporal region of interest, and orange dots and line represent values from the inferior-frontal region of interest in the theta frequency band.

Table 12. Correlations between performance on RAN of objects and the coherence values at source level controlled for age

		Correlation coefficient	Sig	*N*
Delta	Temp	−0.070	0.624	53
	Inf-front	0.020	0.888	53
Theta	Temp	−0.005	0.971	53
	Inf-front	0.055	0.699	53

Note. Sig = significance. Temp = Temporal. Inf-front = Inferior-frontal.

Table 13. Correlations between performance on the phonological processing task and the coherence values at source level

		Correlation coefficient	Sig	*N*
Delta	Temp	0.037	0.836	34
	Inf-front	−0.058	0.743	34
Theta	Temp	0.234	0.182	34
	Inf-front	0.131	0.462	34

Note. Sig = significance. Temp = Temporal. Inf-front = Inferior-frontal.

Table 14. Correlations between performance on the repetition of sentences task and the coherence values at source level

		Correlation coefficient	Sig	*N*
Delta	Temp	0.097	0.585	34
	Inf-front	0.200	0.256	34
Theta	Temp	0.076	0.670	34
	Inf-front	−0.055	0.757	34

Note. Sig = significance. Temp = Temporal. Inf-front = Inferior-frontal.

DISCUSSION

This study investigated whether children and adults differ in overall brain activity as well as in left and right auditory cortex activity while listening to various speech units that are essential for spoken language processing. More specifically, we examined how auditory processing of words and sentences is reflected in the level of coherence and hemispheric lateralization across development, and how these are related to the processing of syllables, at both the sensor and source levels. To this end, two age groups were tested for comparison: children between the ages of 4.7 and 9.3 years and adults. Coherence is an interesting and useful measure of the brain's ability to track the speech envelope across different frequency bands by quantifying the similarity in frequency content between brain activity and the speech envelope. The higher the coherence between brain activity and the speech envelope, the better the speech tracking.

First, we found an improvement with age in the brain's ability to track speech evidenced as increased coherence values in the delta and theta frequency bands between brain signal and the speech envelope. Second, at the sensor level, where the whole hemispheres were examined, we found an interaction between hemispheres and age groups in the delta band with adults showing larger values at the left than the right hemisphere. However, at the source level, hemispheric differences in coherence values did not interact with age, which suggests no differences in maturation rates for the left and right auditory and frontal cortices in the degree to which the brain can synchronize to the speech envelope. Third, we also found differences in the coherence values observed for the word and sentence stimuli independent of age, although this was attributed to physical stimulus length rather than linguistic unit size, suggesting that the methodological approach should be taken into account when interpreting findings

about speech perception. Last, we found no relationship between the general maturation of auditory processing and speech tracking as indicated by early ERFs to syllables.

Developmental differences were found as an overall increase in the coherence values in both the delta and theta frequency bands between adults and children, with adults showing largest values compared to children. Further, the topography of the coherence for these frequencies exhibited a clear pattern of auditory cortex activation at the sensor level. This was mostly confirmed by the source level analysis. The coherence values reflect how similar the frequency contents between the brain signal and the speech envelope are; therefore, our findings could be interpreted as increased precision of the auditory system to track the speech from childhood to adulthood. However, when examining the coherence values as a continuous variable within the child groups, we did not find any correlation between age and coherence values. There may be several reasons for the observed differences between adults and children.

First, basic auditory processing matures slowly with major changes in, for example, ERP responses noted at around ages 8–9 years with further changes until late adolescence (Ponton et al., 2000). This slow maturation of basic processing could affect the precision of speech processing in a bottom-up manner.

Second, the bottom-up process could be affected by genetically driven maturation or continued exposure to speech that refines the bottom-up pathway of the auditory system (Kuhl, 2000; Ponton et al., 2000). At the same time, continuous exposure to speech refines and changes the brain's ability to perceive speech in a top-down manner as well (Kuhl, 2000). This environmental input would shape long-term memory representations, therefore affecting speech processing.

Last, the development of speech tracking may interact with other co-developing cognitive and language-related abilities (e.g., receptive and expressive vocabulary, speech motor and phonological developments) in addition to maturational factors such as age. There is for instance evidence that children initially process large units that are lexically based (e.g., words) before developing representations for smaller units (syllables, individual phonemes; for a review, see Vihman, 2017). This process may also be affected by reading acquisition that places emphasis on phonemes (e.g., Brennan et al., 2013; Popescu & Noiray, 2019; Ziegler & Goswami, 2005). Further, in speech production research investigating the size of coarticulatory units across age, Noiray and colleagues (Noiray et al., 2018; Noiray, Wieling et al., 2019) noted that children do not mature their coarticulatory patterns in a linear fashion. Instead, they found that preschoolers at the age of 3, 4, and 5 organised their speech in larger chunks compared to primary school children at the age of 7 and adults (Noiray et al., 2018; Noiray, Wieling et al., 2019).

In a subsequent study, Noiray and colleagues further demonstrated that the development of children's phonological awareness, that is, the awareness that the native language, is composed of various size compounds (e.g., syllables, rhyme, and individual phonemes) and the ability to manipulate those units interacts with children's speech motor organisation (Noiray, Popescu et al., 2019). Greater awareness of individual phonemes was associated with greater phonemic differentiation of articulatory gestures. To summarise, relationships between several cognitive and language-related abilities occur in the course of language acquisition, and they seem to evolve dynamically over time (e.g., Noiray, Popescu et al., 2019; Noiray, Wieling et al., 2019; Vihman, 2017). In future research, it will be important to investigate larger samples of children spanning kindergarten to primary school to better understand the dynamics of these relationships and how they contribute to the development of speech tracking specifically.

Our research provides supplementary information about the processing of various speech-sized units. More specifically, we confirmed the role of the lower frequency bands for sentence processing (Molinaro & Lizarazu, 2018; Ríos-López et al., 2020) and extended this finding to the processing of words. Indeed, there is evidence that theta and delta bands play a main role for parsing the continuous speech signal into linguistic and prosodic units (Poeppel, 2014; Poeppel & Assaneo, 2020). Thus, a developmental increase in brain coherence in these frequency bands could be associated with a development in the processing and awareness of those distinct speech units. While we did not find any significant correlation between children's coherence values and their performance on phonological processing or sentence repetition, future studies should further examine the relationship among phonological awareness, reading, and speech tracking in the brain with larger samples of children and longitudinally.

We also found that coherence was higher for words than sentences for all frequency bands. This was somewhat surprising given longer stimuli should provide opportunities for brain activity to lock to the ongoing auditory signal. It is important, however, to note that after checking the coherence at the beginning of sentences trials (with the same length as used for words), we noted that coherence increased compared to the original values for sentences. This suggests that the higher coherence for words than sentences does not reflect differences that would be directly relevant for neural computation of linguistic units, but more the characteristics of calculating the coherence measure for short versus long stimuli. For example, longer stimuli provide greater chances for brain activity unrelated to stimulation to occur, with higher likelihood of this noise in the brain activity interacting with the coherence measure. Therefore, comparison of coherence measures across different length stimuli should be done with care, as pure physical length of the stimulus might have an effect on the results. In general, our findings confirm that speech tracking can happen at a shorter length, such as words, as well as at sentence level.

While we expected to find a significant interaction between the coherence values in the left and right hemispheres and the two age groups in the delta band, this difference was in the opposite direction from our predictions, where we expected larger values in the right hemisphere compared to the left, particularly in the adults (Luo & Poeppel, 2007). We observed significantly larger coherence values in the left hemisphere than right for adults in the delta band only at the sensor level—this was not observed at the source level. One possible reason for this difference between sensor and source level findings could be the selection of channels or regions used in the analysis. The sensor level comparison used an overall average of the hemispheres, while the source level focused on the temporal regions. At the source level, when focusing on the temporal regions, a hemispheric difference was found in the expected direction of larger right side activation compared to the left; however, the difference was no longer significant after FDR correction.

Finally, we investigated the overlap of different maturational processes across linguistic units by examining the age-related changes in brain activity around 125 ms (the time-window of the N1m response in adults to the syllables) and compared those to the coherence values for words and sentences. The amplitudes of evoked responses and the coherence values likely represent different neuronal mechanisms. The first presumably represents a more general maturation of the auditory and speech perception system, and the latter is likely linked to top-down processes such as comprehension of speech (Luo & Poeppel, 2007; Peelle et al., 2013; Ponton et al., 2000, 2002). We compared the development of the evoked responses and coherence using two approaches. In our first approach using GMFP, which included both the P1m and N1m responses, we found no significant difference between the groups. However, the P1m and N1m likely represent different computational processes of the sounds. In early childhood the auditory ERPs show prominent P1 and N2 peaks. During development,

the P1 response shifts to earlier latencies accompanied by a decrease in amplitudes, and the prominent N1 response emerges to the waveform at around early school years (Albrecht et al., 2000; Ponton et al., 2000). Importantly, P1m of young children and N1m of older children and adults show very similar timing, obscuring the interpretation of purely GMFP-based interpretations (Parviainen et al., 2019).

Therefore, as a next step, we checked whether the spatial patterns and timings of responses in the left and right hemispheres for each individual matched with the expected N1m pattern based on the current direction in the magnetometer topography. We found no systematic correlation between the responses in the time-window of the first prominent evoked field (the N1m or P1m) and coherence between the speech envelope in the delta and theta bands. Although the correlations were not significant it should be noted that several factors might affect the result, such as sample size and methodology used. The ERFs and coherence were examined using different approaches (ERFs in the time domain and coherence in the frequency domain). It is possible that the use of these different approaches makes the measures difficult to compare directly. Taking this into consideration, our results suggest that the evoked response to syllables and the speech tracking might develop independently of each other and not share robust maturational mechanisms. If this is the case, the ERF amplitudes could reflect more bottom-up processes while the coherence values more top-down processes. Previous literature shows that ERFs are clearly modulated by the physical features of the sounds (Näätänen & Picton, 1986; Näätänen et al., 1997), and the speech envelope following seems to be linked to speech intelligibility and attention (Peelle et al., 2013).

Furthermore, especially in the case of younger children, while the GMFP did show a response around 100 ms, individual inspection of the responses showed that the response at the time was not actually an N1m response, but rather P1m. Taking this into account, it could perhaps explain why we found no differences between the groups when comparing the responses based on time-windows defined by only the peak in the GMFP.

Likewise, no hemispheric differences were observed in any of the age groups for the GMFP-based values of the evoked response, in contrast to a previous MEG study (Parviainen et al., 2019). However, the difference between these studies most likely reflects the chosen analysis approach. While the GMFP reflects the overall response strength at the sensor level, at the source level, measures of equivalent current dipoles depict the spatially specific amplitude values at different time points. Indeed, our data demonstrated a similarly delayed pattern of N1m topography in children, with more clear response in the right than left hemisphere, as was implied by Parviainen and colleagues (2011, 2019). Inspection of the responses themselves revealed that only about one third of the children actually had the N1m evoked response. Due to our sample size, comparison using brain responses of children who indeed produced the N1m response would be unbalanced, and future research should look into this comparison with a larger sample size for both groups.

One of the limitations of our study is related to the type of stimulus we used, as words uttered in isolation are more pronounced than those in a sentence. However, our post hoc comparison of the coherence values to sentences at word-length trials vs. sentence-length trials revealed that higher coherence was found at the beginning of the sentence regardless of stimulus type. It is possible the word level stimuli could be affected by their short length in the estimation of low frequencies, and therefore the results for the word stimuli should be regarded with caution. Another potential limitation is the number of participants limiting the power of the study to detect more subtle differences related to development or hemispheric processing of the different stimuli.

In summary, we investigated developmental differences in speech processing of various speech sized units that are linguistically relevant for spoken language processing, the syllable, word, and sentence, using MEG. We also examined how the hemispheric specialization is represented in brain responses and whether this specialization varies as a function of age. Overall, we found that both delta and theta frequencies show coherence with speech and seem to be important for speech processing. We also found developmental changes in the coherence values, which could reflect both bottom-up maturation and top-down refinement caused, for example, by continuous refinement of speech sound representations. Our data also suggest that the general functional maturation of the auditory cortices follows a different trajectory to that of the brain activity tracking the speech envelope.

ACKNOWLEDGMENTS

The authors would like to thank Katja Koskialho, Sonja Tiri, Ainomaija Laitinen, Annamaria Vesterinen, Aino Sorsa, Maija Koskio, and Cherie Jenkins for their help with data collection. This work has been supported by the European Union projects Predictable (Marie Curie Innovative Training Networks, # 641858), and ChildBrain (Marie Curie Innovative Training Networks, #641652).

FUNDING INFORMATION

Barbara Höhle, Horizon 2020 Framework Programme (https://dx.doi.org/10.13039/100010661), Award ID: 641858. Paavo Leppänen, Horizon 2020 Framework Programme (https://dx.doi.org/10.13039/100010661), Award ID: 641652.

AUTHOR CONTRIBUTIONS

Orsolya Beatrix Kolozsvári: Conceptualization: Equal; Data curation: Equal; Investigation: Lead; Formal analysis: Equal; Methodology: Equal; Project administration: Equal; Software: Equal; Visualisation: Lead; Writing–Original Draft: Equal; Writing–Review & Editing: Equal. **Weiyong Xu**: Conceptualization: Equal; Data curation: Equal; Formal analysis: Equal; Methodology: Equal; Software: Equal; Visualisation: Supporting; Writing–Original Draft: Equal; Writing–Review & Editing: Equal. **Georgia Gerike**: Formal analysis: Equal; Methodology: Equal; Visualisation: Supporting; Writing–Review & Editing: Equal. **Tiina Parviainen**: Conceptualization: Equal; Methodology: Supporting; Supervision: Supporting; Writing–Review & Editing: Equal. **Lea Nieminen**: Conceptualization: Equal; Methodology: Supporting; Writing–Review & Editing: Equal. **Aude Noiray**: Supervision: Supporting; Writing–Review & Editing: Equal. **Jarmo Hämäläinen:** Conceptualization: Equal; Formal analysis: Equal; Funding acquisition: Lead; Methodology: Equal; Project administration: Equal; Supervision: Lead; Writing–Original Draft: Equal; Writing–Review & Editing: Equal.

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. DOI: https://doi.org/10.1523/JNEUROSCI.0187-08.2008, PMID: 18400895, PMCID: PMC2713056

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2009). Abnormal cortical processing of the syllable rate of speech in poor readers. *Journal of Neuroscience*, *29*(24), 7686–7693. DOI: https://doi.org/10.1523/JNEUROSCI.5242-08.2009, PMID: 19535580, PMCID: PMC2763585

Albrecht, R., Suchodoletz, W., & Uwer, R. (2000). The development of auditory evoked dipole source activity from childhood to adulthood. *Clinical Neurophysiology*, *111*(12), 2268–2276. DOI: https://doi.org/10.1016/S1388-2457(00)00464-8, PMID: 11090781

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program] (Version 6.0.37). Retrieved from https://www.praat.org/

Bourguignon, M., De Tiege, X., de Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., Goldman, S., Hari, R. & Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the

reader's voice. *Human Brain Mapping*, *34*(2), 314–326. DOI: https://doi.org/10.1002/hbm.21442, PMID: 22392861, PMCID: PMC6869855

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems: Learning to read reorganizes language network. *Human Brain Mapping*, *34*(12), 3354–3368. DOI: https://doi.org/10.1002/hbm.22147, PMID: 22815229, PMCID: PMC3537870

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*(5600), 2013–2015. DOI: https://doi.org/10.1126/science.1077066, PMID: 12471265

Denckla, M. B., & Rudel, R. G. (1976). Naming of object-drawings by dyslexic and other learning disabled children. *Brain and Language*, *3*(1), 1–15. DOI: https://doi.org/10.1016/0093-934X(76)90001-8, PMID: 773516

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., & Albert, M. S. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980. DOI: https://doi.org/10.1016/j.neuroimage.2006.01.021, PMID: 16530430

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience*, *11*, 481. DOI: https://doi.org/10.3389/fnhum.2017.00481, PMID: 29033809, PMCID: PMC5624994

Din, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience. *Nature Neuroscience*, *19*(1), 158–164. DOI: https://doi.org/10.1038/nn.4186, PMID: 26642090, PMCID: PMC4809195

Eklund, K., Torppa, M., Aro, M., Leppänen, P. H. T., & Lyytinen, H. (2015). Literacy skill development of children with familial risk for dyslexia through grades 2, 3, and 8. *Journal of Educational Psychology*, *107*(1), 126–140. DOI: https://doi.org/10.1037/a0037121

Fowler, A. E. (1991). How early phonological development might set the stage for phoneme awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 97–117). Lawrence Erlbaum.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2. DOI: https://doi.org/10.3389/fpsyg.2011.00130, PMID: 21743809, PMCID: PMC3127251

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*. DOI: https://doi.org/10.3389/fnhum.2012.00340, PMID: 23316150, PMCID: PMC3539830

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126. DOI: https://doi.org/10.1159/000208934, PMID: 19390234

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127–1134. DOI: https://doi.org/10.1016/j.neuron.2007.09.038, PMID: 18093532

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., Herman, D. H., Clasen, L. S., & Toga, A. W. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*, *101*(21), 8174–8179. DOI: https://doi.org/10.1073/pnas.0402680101, PMID: 15148381, PMCID: PMC419576

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences*, *15*(1), 3–10. DOI: https://doi.org/10.1016/j.tics.2010.10.001, PMID: 21093350

Goswami, U., & Bryant, P. (2016). *Phonological skills and learning to read*. Psychology Press. DOI: https://doi.org/10.4324/9781315695068

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. DOI: https://doi.org/10.3389/fnins.2013.00267, PMID: 24431986, PMCID: PMC3872725

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752. DOI: https://doi.org/10.1371/journal.pbio.1001752, PMID: 24391472, PMCID: PMC3876971

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: Studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences*, *98*(2), 694–699. DOI: https://doi.org/10.1073/pnas.98.2.694, PMID: 11209067, PMCID: PMC14650

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. DOI: https://doi.org/10.1007/BF02512476, PMID: 8182960

Häyrinen, T., Serenius-Sirve, S., & Korkman, M. (1999). *Lukilasse. Lukemisen, Kirjoittamisen Ja Laskemisen Seulontatestistö Peruskoulun Ala-Asteen Luokille*. Helsinki: Psykologien Kustannus Oy.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430. DOI: https://doi.org/10.1016/S0893-6080(00)00026-5, PMID: 10946390

Kalashnikova, M., Peter, V., Di Liberto, G. M., Lalor, E. C., & Burnham, D. (2018). Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific Reports*, *8*(1). DOI: https://doi.org/10.1038/s41598-018-32150-6, PMID: 30214000, PMCID: PMC6137049

Korkman, M., Kirk, U., & Kemp, S. L. (1998). *NEPSY: A developmental neuropsychological assessment*. Psychological Corporation.

Korkman, M., Kirk, U., & Kemp, S. L. (2008). *NEPSY-II: Lasten neuropsykologinen tutkimus* [NEPSY-II: A developmental neuropsychological assessment]. Psykologien Kustannus Oy.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. DOI: https://doi.org/10.1073/pnas.97.22.11850, PMID: 11050219, PMCID: PMC34178

Leong, V., & Goswami, U. (2014). Assessment of rhythmic entrainment at multiple timescales in dyslexia: Evidence for disruption to syllable timing. *Hearing Research*, *308*, 141–161. DOI: https://doi.org/10.1016/j.heares.2013.07.015, PMID: 23916752, PMCID: PMC3969307

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex.

Neuron, *54*(6), 1001–1010. DOI: https://doi.org/10.1016/j.neuron.2007.06.004, PMID: 17582338, PMCID: PMC2703451

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650. DOI: https://doi.org/10.1111/ejn.13811, PMID: 29283465

Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., & Carreiras, M. (2016). Out-of-synchrony speech entrainment in developmental dyslexia: Altered cortical speech tracking in dyslexia. *Human Brain Mapping*, *37*(8), 2767–2783. DOI: https://doi.org/10.1002/hbm.23206, PMID: 27061643, PMCID: PMC6867425

Müller, V., Gruber, W., Klimesch, W., & Lindenberger, U. (2009). Lifespan differences in cortical dynamics of auditory perception. *Developmental Science*, *12*(6), 839–853. DOI: https://doi.org/10.1111/j.1467-7687.2009.00834.x, PMID: 19840040

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., & Allik, J., (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. DOI: https://doi.org/10.1038/385432a0, PMID: 9009189

Näätänen, R., & Picton, T. W. (1986). N2 and automatic versus controlled processes. *Electroencephalography and Clinical Neurophysiology Supplement*, *38*, 169–186. PMID: 3466775

Noiray, A., Abakarova, D., Rubertus, E., Krüger, S., & Tiede, M. (2018). How do children organize their speech in the first years of life? Insight from ultrasound imaging. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1355–1368. DOI: https://doi.org/10.1044/2018_JSLHR-S-17-0148, PMID: 29799996

Noiray, A., Popescu, A., Killmer, H., Rubertus, E., Krüger, S., & Hintermeier, L. (2019). Spoken language development and the challenge of skill integration. *Frontiers in Psychology*, *10*. DOI: https://doi.org/10.3389/fpsyg.2019.02777, PMID: 31920826, PMCID: PMC6938249

Noiray, A., Wieling, M., Abakarova, D., Rubertus, E., & Tiede, M. (2019). Back from the future: Nonlinear anticipation in adults' and children's speech. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 3033–3054. DOI: https://doi.org/10.1044/2019_JSLHR-S-CSMC7-18-0208, PMID: 31465705

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 1–9. DOI: https://doi.org/10.1155/2011/156869, PMID: 21253357, PMCID: PMC3021840

Pang, E., & Taylor, M. (2000). Tracking the development of the N1 from age 3 to adulthood: An examination of speech and non-speech stimuli. *Clinical Neurophysiology*, *111*(3), 388–397. DOI: https://doi.org/10.1016/S1388-2457(99)00259-X

Parviainen, T., Helenius, P., Poskiparta, E., Niemi, P., & Salmelin, R. (2011). Speech perception in the child brain: Cortical timing and its relevance to literacy acquisition. *Human Brain Mapping*, *32*(12), 2193–2206. DOI: https://doi.org/10.1002/hbm.21181, PMID: 21391257, PMCID: PMC6870499

Parviainen, T., Helenius, P., & Salmelin, R. (2019). Children show hemispheric differences in the basic auditory response properties. *Human Brain Mapping*, *40*(9), 2699–2710. DOI: https://doi.org/10.1002/hbm.24553, PMID: 30779260, PMCID: PMC6865417

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 320. DOI: https://doi.org/10.3389/fpsyg.2012.00320, PMID: 22973251, PMCID: PMC3434440

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. DOI: https://doi.org/10.1093/cercor/bhs118, PMID: 22610394, PMCID: PMC3643716

Pena, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences*, *100*(20), 11702–11705. DOI: https://doi.org/10.1073/pnas.1934290100, PMID: 14500906, PMCID: PMC208821

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*(1), 245–255. DOI: https://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Current Opinion in Neurobiology*, *28*, 142–149. DOI: https://doi.org/10.1016/j.conb.2014.07.005, PMID: 25064048, PMCID: PMC4177440

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334. DOI: https://doi.org/10.1038/s41583-020-0304-4, PMID: 32376899

Ponton, C. W., Eggermont, J. J., Khosla, D., Kwong, B., & Don, M. (2002). Maturation of human central auditory system activity: Separating auditory evoked potentials by dipole source modeling. *Clinical Neurophysiology*, *113*(3), 407–420. DOI: https://doi.org/10.1016/S1388-2457(01)00733-7

Ponton, C. W., Eggermont, J. J., Kwong, B., & Don, M. (2000). Maturation of human central auditory system activity: Evidence from multi-channel evoked potentials. *Clinical Neurophysiology*, *111*(2), 220–236. DOI: https://doi.org/10.1016/S1388-2457(99)00236-9, PMID: 10680557

Popescu, A., & Noiray, A. (2019, November 7–10). Reading proficiency and phonemic awareness as predictors for coarticulatory gradients in children. In *Proceedings of the 44th Boston University Conference on Language Development*, Boston, MA.

Ríos-López, P., Molinaro, N., Bourguignon, M., & Lallier, M. (2020). Development of neural oscillatory activity in response to speech in children from 4 to 6 years old. *Developmental Science 23*(6), e12947. DOI: https://doi.org/10.1111/desc.12947, PMID: 32043677, PMCID: PMC7685108

Taulu, S., & Kajola, M. (2005). Presentation of electromagnetic multichannel data: The signal space separation method. *Journal of Applied Physics*, *97*(12), 124905. DOI: https://doi.org/10.1063/1.1935742

Taulu, S., Simola, J., & Kajola, M. (2005). Applications of the signal space separation method. *IEEE Transactions on Signal Processing*, *53*(9), 3359–3372. DOI: https://doi.org/10.1109/TSP.2005.853302

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *Journal of Neuroscience*, *29*(47), 14726–14733. DOI: https://doi.org/10.1523/JNEUROSCI.1246-09.2009, PMID: 19940167, PMCID: PMC6666009

Telkemeyer, S., Rossi, S., Nierhaus, T., Steinbrink, J., Obrig, H., & Wartenburger, I. (2011). Acoustic processing of temporally modulated sounds in infants: Evidence from a combined near-infrared spectroscopy and EEG study. *Frontiers in Psychology*, *1*. DOI: https://doi.org/10.3389/fpsyg.2011.00062, PMID: 21716574, PMCID: PMC3110620

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction.

Journal of Educational Psychology, *91*(4), 579. DOI: https://doi.org/10.1037/0022-0663.91.4.579
Uhlhaas, P. J., Roux, F., Rodriguez, E., Rotarska-Jagiela, A., & Singer, W. (2010). Neural synchrony and the development of cortical networks. *Trends in Cognitive Sciences*, *14*(2), 72–80. DOI: https://doi.org/10.1016/j.tics.2009.12.002, PMID: 20080054
Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology*, *108*(1), 1–27. DOI: https://doi.org/10.1111/bjop.12207, PMID: 27449816
Wechsler, D. (2003a). *Wechsler preschool and primary scale of intelligence – Third Edition (WPPSI-III)*. NCS Pearson, Inc., USA. Psykologien Kustannus Oy, Helsinki. DOI: https://doi.org/10.1037/t15177-000
Wechsler, D. (2003b). *WISC-IV: Administration and scoring manual*. Psychological Corporation.
Wechsler, D. (2008). *Wechsler adult intelligence scale – Fourth Edition (WAIS–IV)*. NCS Pearson. DOI: https://doi.org/10.1037/t15169-000
Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. DOI: https://doi.org/10.1037/0033-2909.131.1.3, PMID: 15631549

III

CORTICO-KINEMATIC AND CORTICO-ACOUSTIC COHERENCE IN ADULTS DURING ACTIVE SPEECH PRODUCTION

by

Orsolya Beatrix Kolozsvári, Weiyong Xu, Georgia Gerike, Jan Kujala, Tiina Parviainen, Paavo Herman Tapio Leppänen, Aude Noiray & Jarmo Arvid Hämäläinen, 2021

Submitted manuscript

Request a copy from author.