Electrophysiological assessment of audiovisual integration in speech perception

Eskelund, Kasper; Andersen, Tobias; MacDonald, Ewen; Dau, Torsten

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 16

Kasper Eskelund

Electrophysiological assessment of audiovisual integration in speech perception

Centre for Applied Hearing Research
Department of Electrical Engineering

Electrophysiological assessment of audiovisual integration in speech perception

PhD thesis by

Kasper Eskelund

PHD-2014-341

ISSN 0909-3192

Cover illustration by Frederik Kirk Eskelund: Man and robot in conversation

Foreword

This dissertation consists of an introductory report, reviewing relevant theory and key prior, experimental findings, in order to introduce the work represented in three articles (Publications A-C) listed below, which present findings of novel experimental paradigms.

Together these parts comprise my PhD dissertation, submitted to the Department of Applied Mathematics and Computer Science at the Technical University of Denmark.

Publication A:

> Eskelund, K., Frølich, L. and Andersen, T.S.: Facial configuration and audiovisual integration of speech: a mismatch negativity study. Proceedings of ISAAR 2013.

Publication B:

> Eskelund, K., MacDonald, Ewen N. and Andersen, T.S.: Face configuration affects speech perception: Evidence from a McGurk Mismatch Negativity study. Neuropsychologia **66**, 2015, 48-54.

Publication C:

> Eskelund, K., Andersen, T.S. and Stekelenburg, J.J.: Electrophysiological correlates of the temporal window of audiovisual integration in speech perception. Unsubmitted manuscript

In addition, a fourth publication produced during the PhD study but based on data produced before admission to the PhD study is included for reference:

Publication D:

> Eskelund, K., Tuomainen, J. and Andersen, T.S.: Audiovisual integration in speech perception: a multi-stage process. Proceedings of ISAAR 2011.

The PhD study and dissertation was supervised by Associate Professor, PhD Tobias S Andersen, DTU Compute (main supervisor), Associate Professor, PhD Ewen MacDonald, DTU Electrical Engineering (co-supervisor) and Professor, PhD Torsten Dau, DTU Electrical Engineering (co-supervisor).

The PhD study was generously funded entirely by the Oticon Foundation.

Kasper Eskelund

Vanløse, June 17, 2014

Abstract

Speech perception integrates signal from ear and eye. This is witnessed by the hearing benefit provided by seeing the speaker while listening, but also by a wide range of audiovisual integration effects, such as ventriloquism and the McGurk illusion. Some behavioral evidence suggest that audiovisual integration of specific aspects is special for speech perception.

However, our knowledge of such bimodal integration would be strengthened if the phenomena could be investigated by objective, neurally based methods. One key question of the present work is if perceptual processing of audiovisual speech can be gauged with a specific signature of neurophysiological activity, the mismatch negativity response (MMN).

MMN has the property of being evoked when an acoustic stimulus deviates from a learned pattern of stimuli. In three experimental studies, this effect is utilized to track when a coinciding visual signal alters auditory speech perception.

Visual speech emanates from the face of the talker. Perception of faces and of speech shares the trait, that they are learned from infancy and seemingly specialized behaviorally and neurally. Due to this, speech and face encoding functions quasi-automatically and with high efficiency. However, perhaps owing to our long experience with human faces, which all are variations on a relatively constrained space of features, face perception is sensitive to manipulations of the structure of the face, the relation between its segments, and the properties of the segments. Does this sensitivity alter the influence of visual speech on the auditory speech percept? In two experiments, which both combine behavioral and neurophysiological measures, an uncovering of the relation between perception of faces and of audiovisual integration is attempted. Behavioral findings suggest a strong effect of face perception upon audiovisual integration, whereas the MMN results are less clear.

Another interesting property of speech perception is that it is relatively tolerant towards temporal shifts between acoustic and visual speech signals. Here, behavioral studies report that perception of audiovisual speech exhibits far greater temporal tolerance than perception of audiovisual non-speech stimuli.

Current findings on neural correlates of this tolerance, however, are few and limited. Here, a novel experimental MMN paradigm is used in effort to shed light on integration asynchronous audiovisual speech. Based on individual behavioral estimates of temporal windows of tolerance, we ask if the MMN signal can be evoked at different points within and outside this window. Behavioral findings match earlier behavioral studies, whereas the MMN findings are ambiguous.

In conclusion, the work presented here sheds light onto two aspects of speech perception. It also presents methodological conclusions on the use of MMN as a neural marker of audiovisual integration.

Populært resumé

Elektrofysiologiske målinger af audiovisual integration i taleperception

Når vi lytter til tale, integreres syn og hørelse. Når vi kan se den talendes ansigt og dermed kan supplere det akustiske talesignal med et visuelt, forbedres afkodningen af information væsentligt. Effekten af denne integration af syn og hørelse kan også opleves i en række audiovisuelle fænomener, såsom bugtalerens illusionstrick og den såkaldte McGurk-effekt, hvor en visuel konsonant kan forandre høreoplevelsen dramatisk. Visse aspekter af denne integration synes at være specifik for taleperception, hvorimod den enten er fraværende eller er anderledes udtrykt for perception af andre stimuli.

Vor viden om multimodal perception baserer sig i vid udstrækning på adfærdsstudier. Et centralt spørgsmål for nærværende afhandling er derfor, hvorvidt audiovisual taleperception kan studeres i neurofysiologiske mål. I gengivelsen af tre eksperimenter belyses her brugen af den såkaldte mismatch negativity-respons (MMN) til dette formål.

MMN har den særlige egenskab, at den udløses af en uventet afvigelse i en række af identiske akustiske stimuli. I eksperimenterne undersøges, om MMN kan anvendes til at måle en forskel i det auditive percept alene udløst af en forandring i en samtidig visuel talestimulus.

Visuel tale udgår fra ansigtet. Ansigts- og taleperception deler den egenskab, at de er tillærte fra den tidligste barndom og frem. Derfor er perception af begge quasi-automatisk og meget effektiv. Men på baggrund af vor erfaring med at aflæse ansigter, er denne proces meget følsom overfor forandringer i ansigtets opbygning, relationen mellem dets dele, de enkelte segmenters form og størrelse, etc. Har denne følsomhed indflydelse på integrationen af det visuelle talesignal med det akustiske? Skal det talende ansigts struktur kunne perciperes som et normalt ansigts før det visuelle talesignal kan påvirke det hørte? I to eksperimenter, der begge kombinerer adfærds- og neurofysiologiske mål, forsøges dette afdækket. Adfærdsresultaterne herfra er i overensstemmelse med tidligere fund, mens de neurofysiologiske resultater udtrykker et andet mønster.

En anden vigtig egenskab ved taleperception er dens relativt store tolerance overfor tidsmæssige forskydninger mellem visuel og akustisk tale. Adfærdsstudier har her vist, at perception af tale er væsentlig mere tolerant overfor disse forskydninger end perception af andre audiovisuelle stimuli. De eksisterende neurofysiologiske undersøgelser af denne tolerance er dog få og begrænsede. Her undersøges fænomenet med MMN-responsen. I et todelt forsøg estimeres først individuelle tolerancetærskler i et adfærdseksperiment. Baseret på disse undersøges det om MMN-responsen som neural indikator af audiovisuel integration kan udløses af stimuli med forskellige grader af audiovisuel asynkroni indenfor og udenfor de estimerede tolerancetærskler. Adfærdsresultaterne er her i overensstemmelse med tidligere fund, mens de neurofysiologiske data er tvetydige.

I sin helhed præsenterer afhandlingen ny viden om to aspekter af taleperception. Derudover rapporterer den metodologiske fund om brugen af MMN som neurofysiologisk indikator af audiovisuel integration.

Acknowledgements

A great many people were influential in the coming to be of this work. I here wish to extend my thanks to all the colleagues, friends and loved ones, who supported my efforts:

First of all, I will express my deep gratitude towards the Oticon Foundation who found my project worthy of funding.

Tobias Andersen – for supporting my endeavors during both my Master and PhD studies, for setting and teaching a high scientific standard, which will be my critical yardstick in future scientific efforts, for inspiring me to think experimentally.

Ewen MacDonald – for truly useful and empathic guidance in the writing process.

Torsten Dau – for including me in the admirable CAHR and CHeSS groups and boosting my motivation.

Jeroen Stekelenburg – for letting me look over his shoulder and learn how audiovisual speech perception actually is studied.

Lars Kai Hansen – for listening.

Rasmus Kjelsmark Olsson – for teaching me important time management lessons.

Laura Frølich – for dragging me into the depths of independent components.

Simon Nielsen – for a great collaboration around getting our EEG equipment up and running.

Michael Smed Kristensen and Johan Wich – my brothers in arms in auditory signal processing.

The staff at CAHR – for generously supplying loads of fun, discussions and a truly great experimental work environment.

Marian Solrun Probst – for making communication with all sorts of systems and administrations transparent and simply having time to talk.

Søren Bo Andersen and my new colleagues in the Military Psychology Unit and the Veterans Centre of Research and Expertise - for offering me an exit strategy.

Jens Hjortkjær – for getting me into all this in the first place – keeping me interested - and getting me out again.

Ulrik Schmidt – for connecting me with the common ground.

Frede Madsen and Anita Kirk – for making life easier.

M, AJ, F – for their love.

Abbreviations

AO	Auditory-only
AV	Audiovisual
AV MMN	Audiovisual mismatch negativity
AV-VO	Audiovisual with subtraction of visual-only
EEG	Electroencephalography
ERP	Event-related potential
fMRI	Functional magnetic resonance imaging
McGurk-MMN	Mismatch negativity produced by means of the McGurk illusion
MEG	Magnetoencephalography
MMN	Mismatch negativity
PET	Positron-emission tomography
SJ	Simultaneity judgment
SOA	Stimulus onset asynchrony
TOJ	Temporal order judgment
VO	Visual-only

Notation conventions

For ease of reading, phonetic stimuli are notated in slashes (e.g. /tabi/).

In graphical representation of ERPs and EEG, negative is plotted upwards.

Audiovisual asynchrony is notated in SOA in ms. Negative SOAs denote the time interval that the acoustic signal precedes the visual signal. Positive SOAs denote the time interval that the visual signal precedes the acoustic signal.

TABLE OF CONTENTS

1 Introduction

Imagine yourself standing at a busy subway station, conversing with a friend. Trains are passing by and noise levels are high. Would seeing the face of your friend be helpful in getting her message? It most probably would. Although acoustic speech is perceived effortlessly under good listening conditions, the benefit of seeing the face of the talker becomes evident under noisy listening conditions. The advantage of the visual face for speech intelligibility and detection becomes progressively larger with increasing acoustic noise level until the acoustic signal is saturated by noise (Grant and Seitz, 2000; Sumby and Pollack, 1954). Such phenomena – and many other - witness how visual speech may facilitate auditory speech perception, and how the modalities of vision and hearing integrate.

Due to this dependency on binding of acoustic and visual signals, speech may be considered an audiovisual signal. The binding of acoustic and visual partial signals relies on three basic properties: 1) visual articulatory movements and the acoustic signal must correspond to some degree (articulatory matching) (Fogassi and Gallese, 2004). 2) acoustic and visual signals must coincide in time within some tolerance interval (temporal matching) (Vroomen and Keetels, 2010). 3) acoustic and visual speech sources must coincide in space, also within some tolerance distance (spatial matching) (Stein and Meredith, 1993). Cues in these three aspects may be fine-grained and ambiguous or masked by noise. Thus, successful binding of acoustic and visual signals relies on a complex processing of the cues available. This processing must reflect a trade-off between exact matching to learned speech pattern on one hand and flexibility to deal with variability in the signals and their correlation on the other. If the perceptual system matches visual and auditory signals too tightly, adaptation to environments with e.g. unexpected acoustic properties becomes difficult, and binding would only happen under specific conditions. If flexibility is too high, on the other hand, signals from disjunct sources (i.e. separate in time) would be integrated, leading to artifactual binding, or, bimodal illusions (Stein and Meredith, 1993). Although the subject of much experimental work, relatively little is known of the mechanisms involved in audiovisual binding in speech perception (e.g. which processes are specific for speech stimuli, which are

effective for other stimulus categories). Even though studies of the neurophysiological properties of speech perception are numerous, there still is a need for objective measures in gauging audiovisual integration. This dissertation is an attempt at investigating mechanisms involved in articulatory matching (chapter 3) and temporal matching (chapter 4) by means of neurophysiology. While doing so, a separate methodological aim is pursued in developing and assessing objective methods for measuring audiovisual integration.

What are the factors that facilitate or influence integration of acoustic and visual speech? This question has been asked in many different ways, addressing different aspects of the stimulus and the perceptual mechanisms involved.

Speech perception is a highly trained capacity, learned in infancy and childhood (Pisoni and Remez, 2005). From the viewpoint of the listener, phonetic encoding cannot be disengaged once speech cues are perceived (Remez et al., 1981; Tuomainen et al., 2005). Natural visual speech emanates from the articulator's face. Perception of faces is also a highly trained capacity, traces of which can be observed already in neonates (Meltzoff and Moore, 1983). Like the encoding of speech, faces are immediately and involuntarily perceived as faces. And we are not in doubt when a stimulus violates our expectations to the structure of a human face (Bruce and Young, 2012). But there are also constraints on the structure of faces, as to when face perception processes accepts them as normal, human faces (Maurer et al., 2002). Face perception thus appears to rely on learned processing of face-specific visual features and their spatial configuration. And it seems that we have a special sensitivity towards stimuli transgressing these limits. Furthermore, faces appear to be processed by dedicated perceptual systems, in turn associated with dedicated cortical processing (Kanwisher et al., 1997).

Perception of both speech and of faces is thus very basic to our social interaction, and they share a key stimulus. But do processes involved in these distinct domains interact? Does face perception modulate speech perception? Previous behavioral studies report highly different results depending on stimulus material and methods applied. Here, we attempt at verifying existing behavioral findings while examining neurophysiological correlates of the interaction.

While natural AV speech involves interaction between face movements and speech sounds, acoustic and visual signals also coincide in space and time. As this is the case with the stimulus material, which natural speech perception is trained on (e.g. when infants and parents engage in audiovisual communication), we might assume that speech perception relies on strict temporal (and spatial) coincidence. Surprisingly, this is not the case: Behavioral experimental studies have revealed that speech perception is tolerant to considerable delays between the acoustic and visual speech signals (e.g. Conrey and Pisoni, 2006; Massaro et al., 1996; Munhall et al., 1996; Navarra et al., 2005; van Wassenhove et al., 2007). The temporal window within which binding of acoustic and visual signals is effective seems to be remarkably wide.

But is synchronous speech processed in the same way as acoustic and visual speech separated by perhaps a fifth of a second (Conrey and Pisoni, 2006)? Estimates on the exact width of this window have varied considerably depending on experimental method. Here, we attempt at verifying behavioral estimates of the window of audiovisual integration with neurophysiological methods.

1.1 Aim of the dissertation

There is a growing amount of evidence for the importance of cross-sensory influences in speech perception. The present work is driven by an interest in understanding the multimodal nature of speech perception. But the importance and possible benefit of integrating acoustic and visual speech signals are underlined by recent findings in hearing impaired populations. Cochlear implant (CI) users exhibit a cortical plasticity after implantation, which results in a higher sensitivity to visual stimulation in the auditory cortex (Sandmann et al., 2012). With a permanent degraded auditory signal quality, CI-assisted perception adapts to this and increases the influence of other cues, such as visual signals (Sandmann et al., 2012). Here, research in audiovisual processing in the normal-hearing and hearing-impaired can be useful in alleviating hearing impairment further. Recently, a technical development produced a prototype of a visually guided hearing aid, letting gaze information control an acoustic beam-forming microphone array (Kidd et al., 2013). Knowledge of stimulus features and

perceptual mechanisms that support binding of acoustic and visual speech is important in developing such technologies.

The present work serves two parallel goals. First, by combined behavioral and neurophysiological methods to investigate two important factors on binding of audiovisual speech. These include (a) the role of face processing in speech perception and (b) the tolerance towards asynchrony between acoustic and visual speech. Second, by applying a specific neurophysiological method in this investigation, the work presented is also directed at an important methodological question: how can audiovisual integration in speech be investigated by objective, neurophysiological means?

The influence of face perception on speech perception has only been subject to few studies. In a behavioral study, Rosenblum and colleagues (2000) found a strong effect, but primarily for a specific speech token. In a similar behavioral study, Hietanen and coworkers reported a weak modulation of speech perception using manipulated faces (2001). However, the stimuli used by the two groups differed considerably. This leaves a twofold need for further research: First, the ambivalent findings reported so far call for further investigation, supplementing the limited findings available. Second, presenting unusual face stimuli may influence perception in many ways. This may be a problem for purely behavioral experimentation, where an unfamiliar visual speech signal may induce a response bias towards a more familiar acoustic speech signal. Thus, there is a need for methods that circumvent such biases.

Here, our aim is replicate the behavioral effect found by Rosenblum and fellow researchers (2000) and further to investigate the effect by neurophysiological measures. By adding this, we target a response less influenced by behavioral response bias.

Tolerance towards audiovisual asynchrony in speech has been targeted by a large number of behavioral studies (e.g. Conrey and Pisoni, 2006; Grant et al., 2004; Munhall et al., 1996; Navarra et al., 2005; van Wassenhove et al., 2007). Most of these agree on a quite wide temporal window, but there still is considerable variability in estimates. This may be due to experimental paradigm differences and stimulus specifics.

Here, we first record behavioral responses to asynchronous speech, as to produce behaviorally based temporal window estimates. We then attempt at verifying these estimates by measuring neural responses to stimuli with select asynchronies, reflecting specific integration levels in the behavioral domain.

We are searching for a neural correlate of audiovisual speech perception. To observe this, the chosen methods must be able to represent a visual influence upon auditory speech perception. Here, we look for a neural representation of a change in the speech percept, induced by integration of visual speech into the auditory percept. Preferably, this method should be pre-attentive, so that attentional bias is preempted. Mismatch Negativity is a component in the auditory event-related potential (ERP), which has exactly the latter characteristic (Näätänen et al., 1978). It has also been shown to be evoked by audiovisual speech stimuli presenting phonetically incongruent audiovisual speech (Colin, 2002; Ponton et al., 2009; Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012). Thus, the audiovisual MMN (AV MMN) is used as a neural marker of audiovisual integration (for a detail description, see Chapter 2.3.2).

Using AV MMN as a marker of integration of acoustic and visual speech, the studies comprising the main part of this dissertation look for neural correlates to the two behavioral phenomena targeted.

2 Factors and indicators of audiovisual speech perception

Multiple factors influence the binding of acoustic and visual speech signals. As one aim of the present dissertation is to investigate the role of two such factors, this chapter presents a brief review of select influences on binding in multisensory speech perception.

To gain an understanding of how audiovisual binding can be studied experimentally, the chapter continues with a review of select behavioral and neurophysiological indicators of cross-modal influences on speech perception. In the sections covering neurophysiological measures, the method applied in the empirical work (Experiments 1-3, Publications A-C in the Appendix), the mismatch negativity component will be given most weight.

2.1 Factors influencing audiovisual integration in speech

2.1.1 The temporal factor

Coincidence in the temporal domain is a supporting factor in AV integration (Stein and Meredith, 1993). Auditory speech is intelligible at lower intensities when presented with synchronous visual speech than with asynchronous (Grant and Greenberg, 2001). Perception of non-speech stimuli exhibits a relatively low tolerance to asynchrony, where intersensory delays of approx. 50 ms are detected (cf. e.g. Zampini et al., 2003). In comparison, speech perception is considerably more tolerant to asynchrony. Tolerable asynchronies of up to 200 ms are commonly reported (e.g. Munhall et al., 1996). Behavioral estimates of the temporal window of tolerable asynchronies, however, are highly dependent on methodology: Different aspects of speech perception may exhibit different sensitivities to asynchrony (see e.g. Soto-Faraco and Alsius, 2009).

2.1.2 The spatial factor

Spatial coincidence of the talking face and the sound source supports binding of acoustic and visual signals (Stein and Meredith, 1993). However, speech perception also shows tolerance towards spatially disjoint acoustic and visual signals, such that separate sources may be integrated into one percept (see e.g. Colin et al., 2001). In this way, a visual signal may influence localization of an

auditory signal. This phenomenon is colloquially known as ventriloquism, a name that also denotes its speech-related origin, in which auditory speech source localization is affected by a co-occuring visual speech signal, to either fully shift the perceived acoustic source to the location of the visual source, or partially doing so (Witkin et al., 1952). Ventriloquism with speech stimuli tolerates audiovisual asynchrony up to 200 ms (Jack and Thurlow, 1973) and greater spatial shifts of the acoustic source can be obtained in the vertical plane than in the horizontal (Thurlow and Jack, 1973).

2.1.3 The role of attention

Whereas the perceptual set or mode of the listener influences crossmodal interaction (see Publication D below, cf. Tuomainen et al., 2005), attention also impacts binding of acoustic and visual speech. The strength of integration depends on visual attention as observed in the McGurk illusion (for a detailed description of the McGurk illusion, see Section 2.2) (Alsius et al., 2005; Tiippana et al., 2004). Visual attention, however, does not influence integration of spatially separate sources, as in ventriloquism (Bertelson et al., 2000).

2.1.4 Articulatory matching: The role of visual speech and face perception

The relation between the structure and dynamics of the visual speech signal and the acoustic speech signal underlies articulatory matching (Fogassi and Gallese, 2004). The source of the natural visual speech signal is the articulator's face. The structure and direction of the face has been shown to impact speech perception as well. Visual speech with modifications violating the configuration of human faces impedes audiovisual integration. Hietanen and colleagues (2001) found that a scrambled face, where mouth, eye and nose segment were shifted both in their horizontal and vertical positions evoked less McGurk responses, reducing McGurk responses from above 80% to 75% (for detailed description of the McGurk illusion see Section 2.2 below). Rosenblum and colleagues (2000) varied both the orientation of the face (facial context) and of the mouth segment. A stimulus presenting an upright face with inverted mouth segment produced less McGurk responses. Here, an unaltered face produced 95% McGurk responses, whereas the manipulated face resulted in only 45%. Stimuli with inverted facial

context also had this effect though to a lesser degree (McGurk response rate at and above 84%).

Another way of indicating the contribution of the articulatory movements of the face is to contrast it with a non-speech visual stimulus, which exhibits the same dynamic behavior. If measuring the contribution to speech detection by the two classes of visual signals, the natural speech signal is more beneficial than other visual signals with the same dynamic information (Bernstein et al., 2004).

2.2 Behavioral indicators of audiovisual integration in speech

Integration of acoustic and visual speech occurs seamlessly and unnoticed whenever presented with coinciding natural acoustic and visual speech emanating from the same source. To study perception of audiovisual speech in behavioral experimentation, it is necessary to evoke a response, which may indicate bimodal interaction.

2.2.1 The McGurk illusion

One way to target audiovisual integration is to produce an illusory integration effect, such as the McGurk illusion (McGurk and MacDonald, 1976). To produce this speech illusion, video of an incongruent visual phoneme (e.g. /ga/) is dubbed unto an acoustic phoneme (e.g. /ba/). For most listeners, McGurk stimuli give rise to either hearing a third phoneme, present in neither modality, which could either be a *fusion* (i.e. /da/) or a *combination* of the visual and acoustic phonemes (i.e. /bga/), or hearing the visual phoneme (i.e. /ga/) in a *visual dominance* response. The propensity to generate McGurk responses varies between specific incongruent syllable combinations. Here, a /ba/+/ga/ stimulus results in 91% McGurk responses, whereas a /pa/+/ta/ combination results in 50% McGurk responses (John MacDonald and McGurk, 1978).

The McGurk illusion is special for speech in two ways: On the side of the stimulus, experimentation with incongruent audiovisual stimuli in other domains (e.g. music) has as of yet not revealed a comparable illusion (see e.g. Saldaña and Rosenblum, 1993). On the perceptual side, influence of the visual signal upon the auditory percept requires that the acoustic stimulus is perceived as speech. In case of sine wave speech stimuli, which are not heard as speech

unless instructed to, there is no influence of the visual signal on the auditory phonetic percept (see Publication D below, cf. Tuomainen et al., 2005). While the McGurk illusion presents the perhaps most convincing demonstration of sensory merging in speech, it also plays a central role in speech and multisensory research as an indicator that audiovisual integration takes place.

2.2.2 Direct measures of temporal perception

Another indicator of integration is when acoustic and visual speech are perceived as occuring simultaneously. This is most often targeted using one of two methods. Simultaneity judgment (SJ) is a simple task, requiring subjects to judge whenever an audiovisual speech stimulus was perceived as simultaneous. The related temporal order judgment (TOJ) task asks subjects to report the onset sequence of the modalities. This is a slightly more difficult task. It recruits the same capability to detect asynchrony as in the SJ task, but also engages the capacity to detect the stimulus sequence. Due to this added difficulty, TOJ tasks usually require training of subjects (Vroomen and Keetels, 2010).

2.2.3 Other behavioral measures

The auditory detection advantage associated with an accompanying visual speech signal (Bernstein et al., 2004; Grant and Seitz, 2000) may also be used as an indicator of audiovisual integration. Using this method, the acoustic signal-to-noise ratio (SNR) is varied (either by varying stimulus or noise masker intensity) in e.g. two-alternative forced-choice paradigms.

2.3 Neurophysiological indicators of audiovisual integration in speech

The neural correlates to some of the AV integration phenomena mentioned above have been investigated with physiological methods, including EEG, MEG, fMRI and PET. Auditory and visual event-related potentials (ERP) are well-studied, and these methods have been applied to research in audiovisual speech perception as well. In the following, key neurophysiological methods are reviewed regarding their importance for the experimental aims of the work reported in this dissertation.

2.3.1 Event-related potentials

In research using event-related potentials (ERPs), modulations of the auditory ERP by visual stimuli have been used as a measure of sensory integration. Studies have focused on alterations of early electrophysiological responses encompassing the negative-going potential peaking at approx. 100 ms (the N1 component) and the positive-going potential peaking at approx. 200-250 ms (the P2 component). One method is to compare ERPs evoked by acoustic and visual in isolation to audiovisual speech. Klucharev and colleagues (2003) investigated this comparison under the hypothesis that if no differences were observed between summed unimodal ERPs and the AV ERP, this would indicate that auditory and visual speech was processed individually also in audiovisual stimuli. However, audiovisual speech produced an ERP of lower amplitude than combined unimodal ERPs. A suppression of the early N1 component due to audiovisual speech was found by Besle and colleagues (Besle et al., 2004) in a similar experiment. In a related study, van Wassenhove and coworkers (2005) found that the N1 peak in the ERP due to audiovisual speech has a shorter latency and lower amplitude than the auditory-only ERP. However, Stekelenburg and Vroomen (2007) studied these effects in both speech and non-speech stimuli and observed N1 and P2 suppression in both classes of stimuli. Moreover, the N1 suppression effect was not sensitive to congruence of acoustic and visual stimuli. The P2 suppression effect, however, showed this sensitivity, but for incongruent speech stimuli only. In a similar study, Vroomen and Stekelenburg (2010) demonstrated that the N1 suppression effect is dependent on the predictability of the timing between acoustic and visual signals. For stimuli with variable asynchrony, N1 was not suppressed.

2.3.2 Mismatch negativity as an indicator of audiovisual integration in speech

Another method based on the ERP is mismatch negativity (MMN). MMN is a component in the auditory ERP, which is evoked by infrequent, deviant stimuli in a sequence of identical standard stimuli, presented with constant SOA (see Figure 1, cf. Näätänen et al., 1978). Stimulus deviance may occur in any basic stimulus property, e.g. pitch, intensity, duration, onset asynchrony, modulation (Näätänen et al., 2004), or phoneme (Pulvermüller et al., 2001). The MMN is

observed as a negative deflection of the ERP due to deviant stimuli. This is usually represented in a differential potential (difference wave), computed by subtracting the average standard ERP from the average deviant ERP (see Figure 2, Garrido et al., 2009).

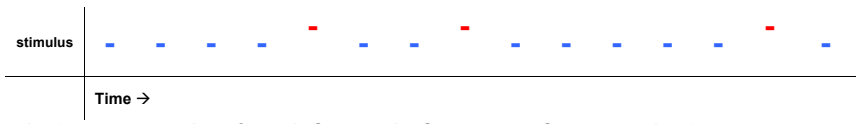


Fig. 1: A representation of a typical MMN stimulus sequence, here presenting 1000 Hz tones as standard stimuli (blue marks) and 1200 Hz tones as deviant stimuli (red marks). A similar stimulus sequence produced the ERPs and MMN differential potential represented in Figure 2. This is also similar to the stimulus sequence used for generating pure-tone MMN in Experiments 2, see Publications B in the Appendix.

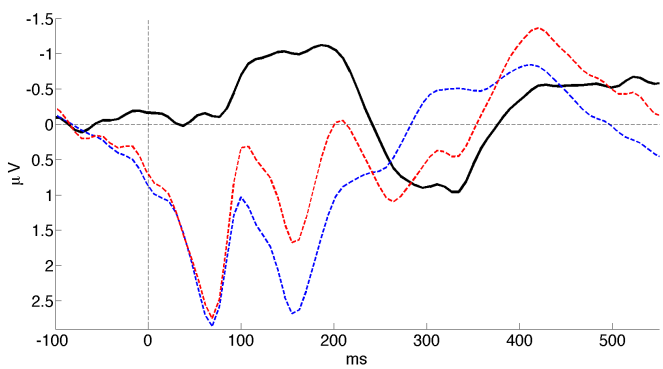


Fig. 2: Pure-tone MMN at electrode Fz produced by a participant in Experiment 2 (see Publication B in Appendix) exposed to a stimulation similar to the sequence represented in Figure 1. The blue line represents the average ERP due to standard stimuli, the red line represents the average ERP due to deviant stimuli, and the black line represents the differential potential presenting a distinct MMN response.

A key feature of the MMN response is that it is pre-attentive (Garrido et al., 2009; Näätänen et al., 1978). Multiple experiments have demonstrated that the differential potential is not influenced by attention. Thus, MMN is also evoked in sleeping subjects (Sallinen et al., 1994), and in newborns (Alho et al., 1990). It can also by produced by comatose patients, in which it has been proposed as a predictor of recovery (Kane et al., 2000). Some studies report that MMN can be modulated by attention (e.g. Arnott and Alain, 2002), however, attentional load does not seem to influence this modulation (Alho et al., 1992).

MMN is evidently generated by the relation between the deviant stimulus and preceding standard stimuli, and it does not depend on any isolated features of

the deviant stimulus. In this very basic sense, MMN has been interpreted as a a memory effect: it is only evoked if there is a stored "memory trace" of the standard stimulus (Näätänen et al., 2004), and the deviant stimulus must be presented within the interval in which this trace is still active to evoke a differential response. The duration of this trace is at least 10 s (Böttcher-Gandor and Ullsperger, 1992). However, different explanatory models for the MMN exist. Below, three dominant theories are briefly reviewed.

Historically, MMN has been interpreted in two competing theories. Following the basic idea of the memory trace mechanism, Näätänen proposed that the MMN was generated by a short-term auditory memory system, responsible for an automatic cortical change-detection process (Naatanen et al., 2005). As such, it generates a model of incoming stimuli based on the frequent standard stimulus. The model is updated or adjusted when the deviant occurs, which represents an error in comparison with the learned pattern (Naatanen et al., 2005). In an alternative interpretation, MMN is accounted for as a variation of the N1 component (Jaaskelainen et al., 2004), which it overlaps. In this view, MMN is a result of neuronal adaptation to the repeated standard stimuli: As a result of repeated stimulation, the neuron populations generating the N1 reduce their response (i.e. by adaptation). When the deviant stimulus occurs, a *fresh afferent* path is active, producing a higher N1 amplitude (Jaaskelainen et al., 2004). Recently, MMN was given a novel, third interpretation along the lines of predictive coding theory (Garrido et al., 2009). In this view, the perceptual system generates a prediction of incoming stimuli on basis of a pattern of repeated stimuli, such as in an auditory oddball sequence. When the deviant is presented, it results in a prediction error. This demands a short-term correction of the prediction model, which evokes the MMN differential potential. When the patterns continues with subsequent standard stimuli, the model is temporarily adjusted, due to which the MMN response is again suppressed (Garrido et al., 2009). In the present experimental work, however, the aim is not do discriminate between or verify any of these interpretations. MMN is exclusively employed as a tool, which enables registering of an early, pre-attentive response in auditory cortex.

Importantly, MMN responses may also be evoked by audiovisual speech stimuli. Using MEG, Sams and colleagues (1991) observed MMN (i.e. the MEG equivalent to MMN) responses to incongruent AV speech where the acoustic signal was kept constant and only the congruency of the visual input deviated. Thus, the deviant stimulus was generated by a change in the visual stimulus, evoking an illusory auditory percept by means of the McGurk illusion (see Figure 3 for a schematic representation of the stimulus sequence). In EEG-based MMN-paradigms, Colin and coworkers (2002), Saint-Amour and colleagues (2007), Stekelenburg and Vroomen (2012) and Ponton and coworkers (2009) have observed a similar MMN response to McGurk-type syllables as deviant stimuli.

Figure 3: A representation of a typical McGurk-MMN stimulus sequence as used in Experiment 1, 2 and 3; Publications A, B and C in the Appendix.

Auditory MMN paradigms produce a differential potential already from 90 ms onwards with amplitudes of several μV (see e.g. Näätänen et al., 2004). Usually the negative peak occurs in the 100-150 ms range and the negative potential decreases towards the 200-250 ms interval. The auditory MMN component is distributed over a wide central to frontal area, over both hemispheres. Many studies report that the differential potential starts at central electrodes with slightly longer latencies at more frontal sites (Garrido et al., 2009).

The McGurk-induced MMN component has a longer latency, often reported in the 150-250 ms range (see e.g. Colin, 2002; Sams et al., 1991), but the scalp distribution follows the same pattern. When calculating McGurk-driven MMN components, epoching and baselining conventions play a central role: Each modality provide relevant but temporally slightly offset events, which could be argued to be relevant as starting-points. Acoustic onset is easily located, but the beginning of articulatory movements usually precede acoustic onset by several ms in natural speech (Chandrasekaran et al., 2009). Epoching and baselining strategies may here render the resulting MMN potential different.

The McGurk-driven MMN is generated by using a deviance in the visual stimulus sequence to produce a deflection of the auditory ERP. A deviant stimulus in a

visual-only oddball sequence would possibly elicit a visual ERP in response. Such a purely visual response would contaminate any MMN due to bimodal interaction. To alleviate this, studies of the McGurk-driven MMN often record ERPs due to visual-only (VO) standard and deviant stimuli, in a similar oddball sequence. To correct for visual oddball responses, AV and VO ERPs are first epoched and baselined to onset of the visual stimulus. Then the VO ERPs due to standard and deviant stimuli are subtracted from the corresponding AV ERPs. Upon this, the resulting AV-VO ERPs are epoched and baselined to onset of the acoustic stimulus, and the MMN differential potential is calculated (see e.g Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012).

The AV-VO MMN is meaningful if looking at McGurk-driven MMN in isolated conditions, as in Experiment 3 (see Publication C, Appendix). However, in some instances, subtraction of visual potentials may not be relevant. In Publications 1 and 2 (see Appendix), multiple audiovisual conditions are directly compared, while only the structure of the visual stimulus is varied. Here, the visual signal across all conditions carry highly similar information, and in the comparison of such conditions, the visual oddball response contribution could be argued to be eliminated.

Interestingly, the McGurk illusion can also be used in a reverse stimulation pattern, *eliminating* the MMN response. In a novel paradigm, Kislyuk and colleagues (2008) demonstrated how an auditory MMN response evoked by a phonetic difference (/va/ as standard, /ba/ as deviant) was extinguished in an audiovisual condtion by presenting a visual syllable /va/, evoking an illusory auditory percept (/va/, see Figure 4). Here, phonetic integration altered phonetic perception of the acoustic deviant /ba/ and effectively abolished the phonetic difference, producing the MMN.

Figure 4: A schematic representation of the audiovisual stimulus sequence created by Kislyuk et al. (2008), which eliminated the MMN otherwise induced by the acoustic phonetic difference between /va/ and /ba/.

2.3.3 Phase-resetting as a measure of adaptation to intermodal lags

When assessing properties of temporal perception, as when investigating the tolerance towards asynchrony in audiovisual speech, it is natural to look for a specialized neural device or network capable of gauging the timing of stimuli. Findings of such neural structures have been reported in literature (Treisman et al., 1990; van Wassenhove, 2009; Wittmann and van Wassenhove, 2009).

It is known, however, that the encoding of temporal stimulus properties is dependent in part on oscillations in the neural circuitry (Buzsáki, 2006; Eagleman, 2008; Johnston and Nishida, 2001; Karmarkar and Buonomano, 2007). This dependence has been employed in studying audiovisual integration of asynchronous stimuli. In a study by Kösem and coworkers (2014), asynchronous audiovisual non-speech stimuli were presented in a 1 Hz rhythm, as to produce an entrained neural oscillation. The asynchrony was 200 ms, which without prior adaptation would be perceived as clearly non-simultaneous (Zampini et al., 2003). During adaptation to the repeated stimulus, phase coherence of the entrained neural oscillation changed systematically. This finding suggests, that the neural structures involved in perception of audiovisual events actively adjust to asynchrony. This is in concordance with previous research in animal models where phase-resetting to audiovisual stimuli in the auditory cortex of the macaque monkey has been observed (Kayser et al., 2008). Though being developed within non-speech audiovisual research, such findings may have great significance for our future understanding of audiovisual speech perception.

2.3.4 Other neurophysiological methods

Perception of audiovisual speech has been subjected to other varieties of neurophysiological experimentation. These will only be mentioned shortly here, as they are of less direct importance for work presented in this dissertation. MEG was already mentioned above, as an extension upon EEG, containing spatial information about the sources of audiovisual integration responses (Sams et al., 1991).

In a related fMRI study, Pekkola and colleagues (Pekkola et al., 2005) compared activity in auditory cortex when subjects were viewing a silent articulating face to activity due to viewing the same face overlaid with a blue oval (which, however, modulated its shape in correlation with lip area modulations in the original stimulus). The natural visual speech stimulus evoked a differential response in auditory cortex when compared with the manipulated stimulus. This may point to a significance of the natural face stimulus in speech perception. However, the experimental design did not control for other auditory factors, such as silent articulation during stimulation (Pekkola et al., 2005).

Experimental designs analogous to the ERP studies mentioned above, comparing responses to audiovisual speech with summed responses to unimodal stimuli have also been employed within fMRI. Here, congruent audiovisual speech has been observed producing higher (superadditive) activity in the superior temporal sulcus (STS), than the sum of STS activity due to the component acoustic and visual stimuli. Incongruent audiovisual stimuli, however, evoke a lower (subadditive) response (Calvert et al., 2000).

Neurobiological investigations of neural connections involved in perception of audiovisual events in animal models underpin such findings. If auditory cortex activity is evoked or modulated by visual stimuli, a neuronal connection must project from visual areas towards auditory cortex. As mentioned above, Kayser and colleagues (2008) studied how both audiovisual and visual stimuli may evoke auditory cortex activity. Interestingly, this effect was strongest for audiovisual stimuli where the visual stimulus preceded the auditory stimulus with 20-80 ms. Further, connections projecting from visual areas (V2) towards auditory cortex have been observed in the macaque monkey (Falchier et al., 2010). Interestingly, a reverse analogue to has also been observed: Visual regions (V1) receive projections from auditory cortex, the STS and other multisensory regions (Falchier et al., 2002). These auditory influences on early visual processing were interpreted as assisting spatial localization.

3 Face configuration and audiovisual integration in speech

3.1 Background

Face-to-face speech communication evidently relies on information from the articulating face. Interestingly, speech and faces both are examples of social stimuli, the encoding of which is learned in the earliest period of childhood (Meltzoff and Moore, 1983). However, research into the capacities of and constraints on perception of auditory speech and of faces seem to be distinct and separate (Rosenblum et al., 2000). Whereas the stimulus constraints on face processing are well-studied, questions regarding the significance of these constraints for integrated audiovisual speech perception remain largely unanswered.

3.1.1 When is a face a face?

Our abilities to read off emotional expressions and identity from faces are fast and trained on a vast number of facial confrontations during infancy and childhood (Bruce and Young, 2012). What requirements must a stimulus meet to invoke these resources? What are the boundaries of a normal facial structure? When is a face a face?

One method of gauging these mechanisms is to target when modulations along some stimulus dimension renders a face as odd, non-normal, or, clearly different from stimuli lacking the modulation. Here, Thompson (1980) provided a highly demonstrative stimulus set. In his study, four different stimuli were presented, crossing vertical inversions of the direction of the face (context) and of mouth and eyes segments.

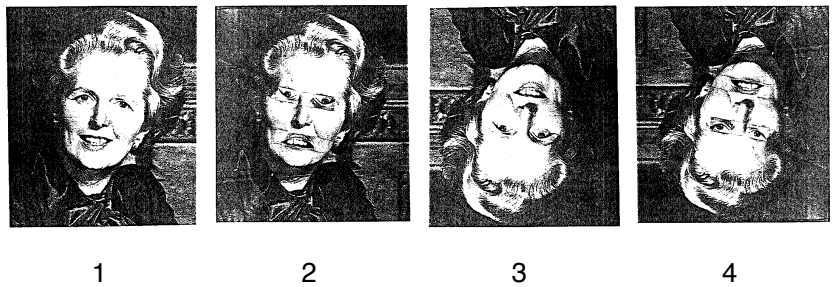


Fig. 5. Stimuli produced by Thompson (1980). The configuration of mouth and eyes are inverted in stimuli 2 and 4, face direction is upside-down in stimuli 3 and 4. The Thatcher illusion is present in stimulus 4, where the upside-down face direction renders the inverted configuration of mouth and eyes (which is the exact same as in stimulus 2) as quasi-normal. Excerpt from Thompson 1980.

Thompson's stimulus set thus presents a highly manipulated facial configuration, either in an upright facial context or in an upside-down context. The inverted segments render the face as strikingly grotesque and non-normal in the upright context, presenting a clear violation of facial structure (see Figure 5, stimulus 2). Interestingly, however, when inverting the direction of the facial context, the same manipulation at the segment level now is inconspicuous to the viewer (see Figure 5, stimulus 4). Where the facial manipulation in the upright facial context stands out, it is harder to discriminate the manipulated configuration in the inverted context from the natural configuration. This asymmetry in perception illustrates that face processing relies on the global configuration of features or segments in faces as they normally occur, and not only on processing of these features individually. Further, when inverting the orientation, the upside-down presentation seemingly eliminates processing of the relation between segments. This means, that perception of a specific aspect of faces – intersegment relations – can be modulated by inverting the orientation.

The quasi-normal appearance of the inverted, misconfigured stimulus was denoted the Thatcher illusion, and the misconfiguration was dubbed Thatcherization. It highlights how a face may be perceived as normal or grotesque depending on the context orientation. The Thatcher illusion stimulus set allows experimentation with stimuli, which retain information, either at context or segment level, while the global perception differs markedly. Thus, these stimuli are ideal for testing the influence of face perception upon audiovisual speech perception.

3.1.2 Levels of face perception

The Thatcher illusion indicates that perception of facial stimuli involves a multi-level hierarchical process. Here, information at the segment level is clearly distinct from the context level. Face perception is very sensitive to configurational violations, but only in an upright context. The illusion shows how information at the context level may interfere with processing of segment interrelations.

In one line of research, three different levels of face perception have been proposed (Maurer et al., 2002; for a different view, cf. Rakover, 2013). At one level, holistic properties are read off as to determine that the object is a face at all (i.e. the presence of the correct number of specific features, general shape, etc.). At a second level, facial segments (eyes, mouth, nose, etc.) are bound into a facial gestalt. A third level is concerned with the interrelations between these segments. When vertically inverting the direction of the facial context, perception of third level interrelations between facial segments is greatly reduced, while first level recognition of faces as such is intact, as is construction of the facial gestalt (Freire et al., 2000; Thompson, 1980). This suggests that inversion to some extent disengages perception of facial segment relations. While upside-down faces are still perceived as faces, context orientation inversion reduces the influence of manipulations at the segment level. Interestingly, vertical inversion mostly reduces the sensitivity to vertical alterations of segments, while horizontal manipulations are still detected (Goffaux and Rossion, 2007).

The dissociation of the segment interrelation level from the holistic and the segment levels in the Thatcher illusion has been investigated in a few neurophysiological studies. The Thatcherized face in upright orientation produces a higher amplitude in the N170 component than does unmanipulated faces or Thatcherized faces in other orientations (Carbon et al., 2005; Milivojevic et al., 2003). In contrast with these findings, Rothstein and colleagues (2001) found that both upright and inverted Thatcherized faces evoked a differential fMRI response compared to the unmanipulated face. Thus, current findings are

ambiguous in the interpretation of the reflection of the behavioral Thatcher illusion within the neural domain.

3.1.3 Visual speech and face perception

As visual speech perception is directed towards the same stimulus as face perception, similar sensitivity to manipulation of orientation could be expected. Vertical inversion of the facial context results in poorer identification of visual syllables (Massaro and Cohen, 1996), but the effect is greatly dependent on the specific syllable. For instance, identification of the labio-dental and vertically asymmetrical /va/ is reduced when the orientation is inverted, whereas the bilabial and vertically symmetrical /ba/ produces similar performance with upright and inverted faces (Rosenblum et al., 2000). Interestingly, identification of visual speech does not solely rely on the overall kinematics of the talking face. In a series of experiments, Jordan and colleagues (2000) investigated natural color gray-scale visual speech and point-light speech identification and found that color and gray-scaled visual speech resulted in similar performance, while performance for point-light speech was poorer. In this finding, chroma differences did not influence performance, whereas the luminance difference between the natural visual stimuli and point-light speech was detrimental to speech identification. This was interpreted as the luminance (shading) differences conveying important visual depth cues, revealing both segment detail (information e.g. from the mouth cavity) and holistic information (supporting perception of the stimulus as a face), both not present in the purely kinematic point-light stimulus.

Shading is also altered when vertically inverting and/or Thatcherizing faces. As shading differences reveal the orientation of the face in space, non-normal shading pattern may represent disturbing conflicts of gravitational information between segments. Talati and coworkers (2010) created two Thatcherized stimuli. One with conflicting shading due to a simple inversion of mouth and eye segments. In the other, mouth and eye segments sampled from a similar photograph, but with inverted lighting (i.e. from below the normal face). The latter stimulus was misconfigured as in normal Thatcherization, but the shading was aligned across all face segments. This reduced the grotesqueness rating,

compared to stimuli with conflicting shading, although a significant Thatcher illusion remained.

3.1.4 Audiovisual speech and face perception

These findings for visual speech suggest that face perception may also influence perception of AV speech to some degree. However, the role of face perception and violations of the face stimulus structure in encoding of audiovisual speech has only been devoted few studies thus far.

Hietanen and co-workers (2001) studied audiovisual integration with phonetically congruent and incongruent audiovisual VCV bisyllables (/iki/ and /ipi/), using the McGurk illusion as a measure of audiovisual integration. To investigate the impact of face perception processes on this, they varied the configuration of facial segments (eyes, mouth and nose). Vertical segment orientation was always intact, whereas the position of segments was varied. For all stimuli, including the natural face, audiovisual integration as indicated by the McGurk illusion was similar, except for a stimulus displaying an asymmetrical configuration of segments. For the latter stimulus, audiovisual integration was reduced. This was interpreted as being due to a violation of horizontal facial symmetry, being more severe than violating vertical positioning (Hietanen et al., 2001).

Vertical inversion of the facial context impedes perception of facial segment interrelations. But how does facial orientation as such influence audiovisual speech perception? Here, studies with vertically inverted faces report slight reductions of audiovisual integration in speech (Bertelson et al., 1994; Jordan and Bevan, 1997), however, Green reported a “significantly weaker” audiovisual integration response with inverted faces (1994). Massaro and Cohen (1996) combined four auditory and four visual syllables and found that the effect of inversion was highly dependent on the specific incongruent syllable combination. In their study, the combination of an auditory /ba/ and a visual /va/ yielded the largest difference between upright and inverted stimuli.

Combining manipulations of both facial direction and segment interrelations, Rosenblum and colleagues (2000) studied the impact of the Thatcher illusion

(Thompson, 1980) upon audiovisual speech processing. In their stimuli, Rosenblum and coworkers only inverted mouth segment and facial context, while leaving the eye segment unaltered in all stimuli. This produced audiovisual speech stimuli with four different facial configurations (see Figure 6).

In their Experiment 1, the authors targeted identification performance of congruent and incongruent (McGurk-type) audiovisual speech stimuli, using the four visual manipulations. McGurk responses were at a similar level for all facial configurations except when the lip segment was inverted in context of an upright face (see Figure 6, stimulus 2).

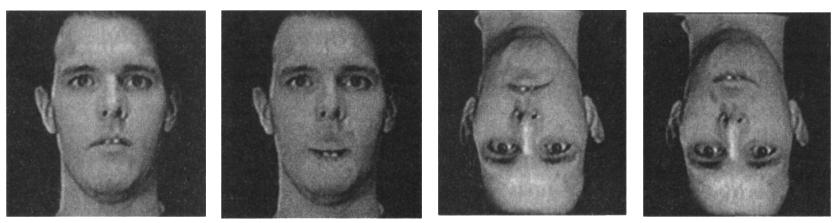


Fig. 6. Stimuli used by Rosenblum and coworkers (2000). The configuration of the mouth segment is inverted in stimuli 2 and 4, face direction is upside-down in stimuli 3 and 4. The Thatcher illusion is present in stimulus 4, where the upside-down face direction renders the inverted configuration of mouth and eyes as quasi-normal. In audiovisual speech tokens, Rosenblum and colleagues found a strong reduction of McGurk responses for the Thatcherized face in upright context (stimulus 2). Excerpt from Rosenblum et al. 2000.

The effect could perhaps in part be due to lip orientation alone. This question was addressed in a separate experiment (Experiment 3), presenting only the lip area segment of all four stimuli, while retaining the position of the segment on the visual display. Here, McGurk responses were on uniform high levels across all stimuli. On basis of these findings, the authors suggested that the reduction in McGurk responses with the Thatcherized face in upright facial context was due to a specific sensitivity to facial configuration. Due to its combination of the McGurk illusion and the Thatcher illusion, the authors colloquially dubbed this the McThatcher effect (Rosenblum, 2001).

Interestingly, perception of the talking face here interacts with perception of audiovisual speech. Rosenblum and colleagues (2000) further measured the impact of the visual speech tokens on visual speech recognition. Here, the

Thatcherized face in upright presentation again disrupted speech perception, while remaining facial stimuli had little impact on performance. Furthermore, stimuli with isolated lips segment showed near-perfect visual speech recognition, regardless of the vertical direction of the segment.

3.2 Experiment 1: Facial configuration and audiovisual integration of speech: a mismatch negativity study

3.2.1 Motivation

The findings of Rosenblum (2000) and Hietanen (2001) point to a possible role for global face processing in perception of speech. Here, Rosenblum's study directly shows an effect of obstructed face processing on integration of visual and acoustic speech. However, their findings strongly vary across stimulus material. Thatcherization reduces McGurk responses much more for the /ba/-/va/ combination, whereas results for the /ba/-/ga/ combination are less clear. This indicates that face configuration may be more influential for specific syllables.

The incongruent syllable /ba/-/va/ exhibited the greatest sensitivity towards face configuration in Rosenblum's study. When presented with an unmanipulated face, this stimulus evoked a visual dominance response (i.e. hearing a /va/ corresponding to the visual stimulus). Although indicating a visual influence, it leads to the question, whether what is observed actually is an effect of audiovisual integration. Fusion and combination responses would clearly indicate that the auditory percept was altered by the visual signal: As the perceived phoneme is present in neither modality, it would indicate that both speech signals were included in the processing. Visual dominance responses, however, can be difficult to distinguish from a bias towards the visual stimulus. Due to this, the causal mechanism producing the response is less clear – it might be either audiovisual integration or simply a disregard for the acoustic signal. Thus, when face configuration modulation changes the response towards the visual phoneme, it is unclear whether what is observed is actually a change in integration of the modalities. The Thatcherized face appears highly grotesque,

which in itself could divert resources towards the more familiar acoustic signal, eliminating the visual dominance response.

It is difficult to devise a behavioral method that alleviates these possible confounds. However, if facial configuration *does* alter speech perception as suggested by Rosenblum's findings, it would correspond to an alteration of speech processing of incongruent speech stimuli. The influence of incongruent visual speech on acoustic speech has previously been shown to evoke an MMN response (Colin, 2002; Ponton et al., 2009; Saint-Amour et al., 2007; Sams et al., 1991). Thus, if face configuration alters the visual influence on the speech percept, it should be observable in the MMN response produced by incongruent speech: the McGurk-driven MMN would vary with the level of visual dominance responses. However, if the effect rather rests on a response bias towards the normal, undistorted face and away from the grotesque Thatcherized face, auditory neural processing should not be affected. In this case, the visual signal would still exert an influence on the auditory percept, and an MMN response would be observed for Thatcherized stimuli as well. Another possibility is that visual responses to normal-face stimuli are not generated by AV integration but rather by response bias. In this case, no McGurk-driven MMN would be observed in this (and any other) condition at all.

Based on these hypotheses, we constructed a series of experiments (see Publications A and B in the Appendix), comparing behavioral responses to incongruent speech with different facial configurations with MMN responses to the same stimuli. The procedures of both these methods are straightforward. However, in each their way, they are also highly dependent on the stimulus material. As the production of an MMN response due to McGurk stimuli can be a challenge in itself, in our first experiment (Publication A in the Appendix), we exclusively focused on the McGurk MMN response and its sensitivity to the vertical orientation of the mouth segment. In other words, we looked for differences in McGurk-MMN response between upright normal faces and upright Thatcherized faces only.

3.2.2 Design

The behavioral effect of Thatcherization on speech perception was as of yet only reported by a single study (Rosenblum et al., 2000). Thus, to ensure that we were able to reproduce the effect, participants first performed a simple behavioral identification task, reporting the perceived phoneme when presented with congruent (/ba/+/ba/) and incongruent (/ba/+/va/) syllables, while varying the facial configuration between unaltered and Thatcherized. Incorrect identifications of the acoustic phoneme in incongruent stimuli were used as a measure of visual influence on the auditory percept. To avoid contamination of the subsequent MMN study by participants with generally low audiovisual integration response, an exclusion criterion of a minimum of 50% incorrect auditory identification of the incongruent syllable was used. No subjects were excluded on this criterion.

The McGurk-driven MMN response can be challenging to produce. As our target is the modulation of the MMN due to facial configuration and not the production of the McGurk-MMN in itself, the final analysis only included data from those subjects who produced an MMN-response to audiovisual speech with unaltered face configuration. The inclusion criterion was formulated as an AV MMN amplitude exceeding -1 μV. In the subsequent MMN measurement, eight of the 19 participants fulfilled this criterion.

Acoustic speech was delivered through ER-3A in-ear monitors, at an intensity of 65 dB(A) SPL. Visual speech was displayed on a 19" CRT screen at 1.2 meter distance.

3.2.3 Results

In the behavioral identification task, incongruent stimuli with normal face configuration produced a high level of incorrect responses (see Figure 7). In comparison, Thatcherization of the face produced a considerable reduction of this.

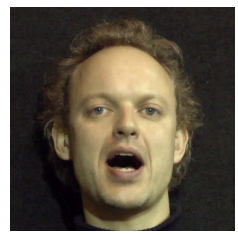
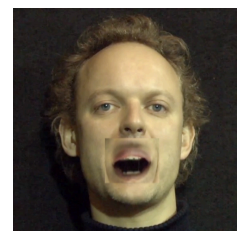
	Normal face configuration	Thatcherized configuration
		
Congruent /ba/ +/ba/	1.5 (0.7)	4.5 (1.5)
Incongruent /ba/ + /va/	93.0 (2.1)	27.0 (6.4)

Fig. 7. Incorrect behavioral identification responses to congruent and incongruent audiovisual phonemes presented with normal or Thatcherized face configuration. Numbers represent mean percent incorrect identifications, numbers in brackets represent standard error of mean. N = 8.

Data from eight participants revealed a MMN response at electrode Cz due to incongruent visual speech in the normal face condition. The grand average difference wave for these subjects is represented in Figure 8. The negative deflection starts at approx. 200 ms and reaches statistically significant levels in the interval 240 to 360 ms.

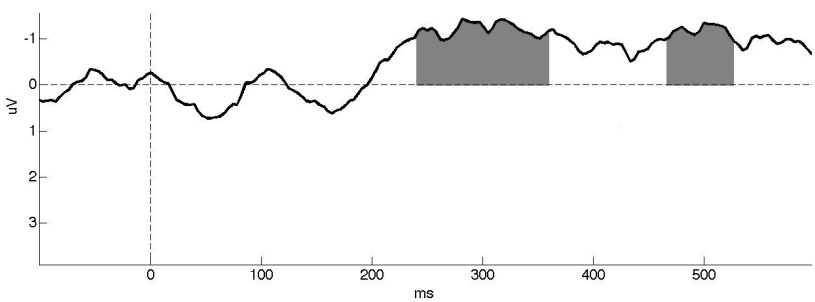


Fig. 8. Differential response due to incongruent phoneme /ba/+/va/ in the normal face condition, electrode site Cz. Shaded area represents the statistically significant part of the difference wave, when subjected to a repeated measures permutation test (for a detailed description see Groppe et al., 2011).

When proceeding to the MMN response in these select subjects in the Thatcherized condition, the differential potential is eliminated (see Figure 9).

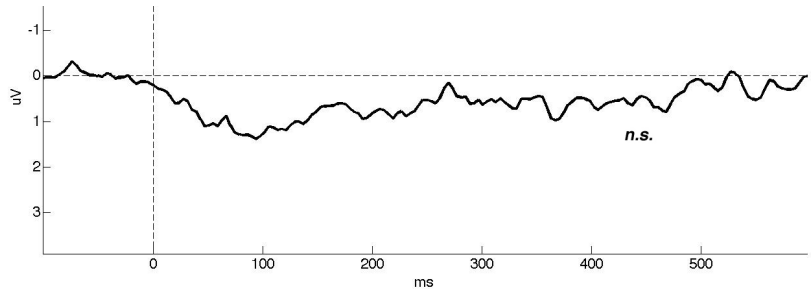


Fig. 9. Differential response due to incongruent phoneme /ba/+/va/ in the Thatcherized face condition, electrode site Cz. No part of the difference wave is significantly different from zero.

3.2.4 Summary of findings

Where the behavioral results represent a striking effect of Thatcherization upon speech perception, the MMN results are ambiguous. On one hand, in the selected group of participants, normal face stimuli evoked a negative differential potential. In the same subjects, this potential was eliminated in the Thatcherized condition. Operating under the assumption that the negative potential observed reflect a MMN response, the findings indicate that the Thatcherized face indeed alters auditory processing.

However, this conclusion should be viewed with caution. The differential potential observed does not follow the established pattern of an MMN response, as reported in previous research. As reviewed above, the auditory MMN response usually occurs in the 100-250 latency range, often starting earlier at 80-90 ms post stimulus. In the case of AV MMN, longer latencies are often reported, but the negative deflection usually covers the N1 response range. In the current findings, the difference starts at the P2-peak and reaches maximal levels even later. This makes it doubtful if the observed response reflects the pre-attentive, early response reported, which MMN is assumed to be.

3.3 Experiment 2: Face configuration affects speech perception: Evidence from a McGurk Mismatch Negativity study

Experiment 1 produced an ambiguous result, as the differential potential does not align well with previous MMN research. To address this, Experiment 2, which is given a full-length description in Publication B (see Appendix), was designed

with the intent of producing a clearer MMN response and testing more aspects of face perception.

3.3.1 Changes and extensions to Experiment 1

For this study, we produced a new stimulus set and introduced important changes to the experimental setup. First, acoustic speech was now delivered through a single speaker placed immediately below the visual display in stead of through in-ear monitors. Second, intensity of the acoustic stimulus was lowered from 65 to 60 dB. Third, new stimuli were produced. Here, we were inspired by Bertelson and colleagues (1994), who used an ambiguous acoustic stimulus (a consonant mixed between /m/ and /n/), which enabled a stronger influence of the visual stimulus. In our new stimulus set, we recorded speech with deliberately less clear articulation in both modalities, although still identifiable. These three changes to stimuli and stimulus delivery all affected the auditory and visual cues in reducing the speech information available. This was motivated by previous observations, that the McGurk illusion increases when the acoustic speech cues are reduced (by adding noise, see Sekiyama and Tohkura, 1991). By reducing the redundancy in both modalities, we aimed at optimizing the MMN illusion and further the McGurk-driven MMN.

Furthermore, to target and control the effect of face perception and not just mouth segment inversion, we constructed and presented newly generated equivalents to all four face configurations as presented by Rosenblum and colleagues (2000), see Figure 10.

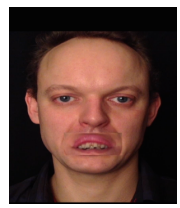
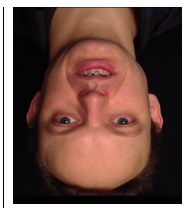
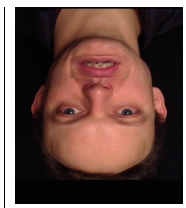
stimuli				
	1	2	3	4
Findings in Rosenblum et al., 2000	Strong McGurk response	Weak McGurk response	Moderate McGurk response	Moderate McGurk response
MMN hypotheses	MMN	No MMN	?	?

Fig. 10. Stills of the visual speech tokens produced as equivalents to stimuli used by Rosenblum et al. (2000). All images represent the visual syllable /va/ at maximal lip opening. In second row, the behavioral findings of Rosenblum and colleagues (2000) are described in brief. In third row, hypothesized MMN responses due to McGurk stimuli with the each of the four facial manipulations. The MMN response is expected to correlate with the behavioral response. Thus, the normal face would yield a strong McGurk-driven MMN, the upright Thatcherized face no such effect, whereas the MMN response is difficult to predict for the inverted faces, which produce moderate behavioral effects.

In Experiment 1, only a limited number of participants produced an MMN response in the normal face condition and were included in the analysis. This may be due to the specific propensity of the stimulus set or the presentation to generate a MMN response. But it could also be due to individual differences in the propensity to generate an MMN response. Auditory MMN is known to show substantial interindividual variability (Lang et al., 1995). If including subjects with a general low propensity to produce an MMN response, this would contaminate our findings of variations in the McGurk-driven MMN. Inspired by Colin and colleauges (2002), we introduced an inclusion criterion based on pure-tone MMN response. In a separate condition, MMN was measured in response to 1000 Hz tones as standard stimuli and 1200 Hz tones as deviant stimuli at 60 dB(A) SPL intensity over a total of 1200 presentations with a deviant rate of 15%. Data from subjects who produced a pure-tone MMN exceeding -1 μV were

included in the analysis. This criterion was introduced as a supplement to a criterion of at least 50% incorrect behavioral identifications of the auditory syllable in incongruent stimuli when presented in the normal face condition, as in Experiment 1.

3.3.2 Design

As in Experiment 1, we compared behavioral and MMN responses. Thus, participants first underwent a behavioral identification task, presenting congruent and incongruent audiovisual syllables with four face configurations (see Figure 10). Then, MMN responses were recorded with congruent and incongruent stimuli in the four face configuration conditions. Additionally, pure-tone MMN was recorded in a separate condition. Participants first completed the behavioral task. Subjects who showed sensitivity towards incongruent speech with normal faces was then included in the MMN experiment. Finally, if meeting the criterion of a clear pure-tone MMN, behavioral and MMN data from individual participants were included in the final analyses. 19 subjects participated in the experiment. All met the behavioral inclusion criterion, whereas eleven met the pure-tone MMN criterion. This is on level with previous studies (see e.g. Colin, 2002).

3.3.3 Results

In the behavioral task, the normal face condition produced a strong effect of incongruent speech (see Figure 11). Also, the Thatcherized condition produced a profound reduction of the visually influenced responses. For the inverted context stimuli, incongruent visual speech produced uniform and large incorrect auditory identification responses, though not on the level of the natural, upright face.

	Upright facial context		Inverted facial context	
	Upright mouth	Inverted mouth (Thatcherized)	Upright mouth	Inverted mouth
Incongruent /ba/ + /va/	92.3 (4.0)	32.7 (6.7)	54.1 (8.4)	65.4 (7.2)

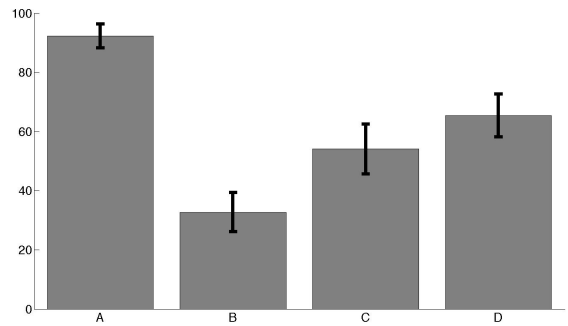


Fig. 11. Behavioral McGurk illusion strength represented as percent incorrect auditory identifications of incongruent syllable /ba/-/va/. A represents the normal, unmanipulated face. B represents the upright, Thatcherized face. C represents the vertically inverted face. D represents the vertically inverted face with inverted mouth segment (inverted Thatcherized). Bars represent mean proportion incorrect auditory identifications, error bars represent standard error of mean.

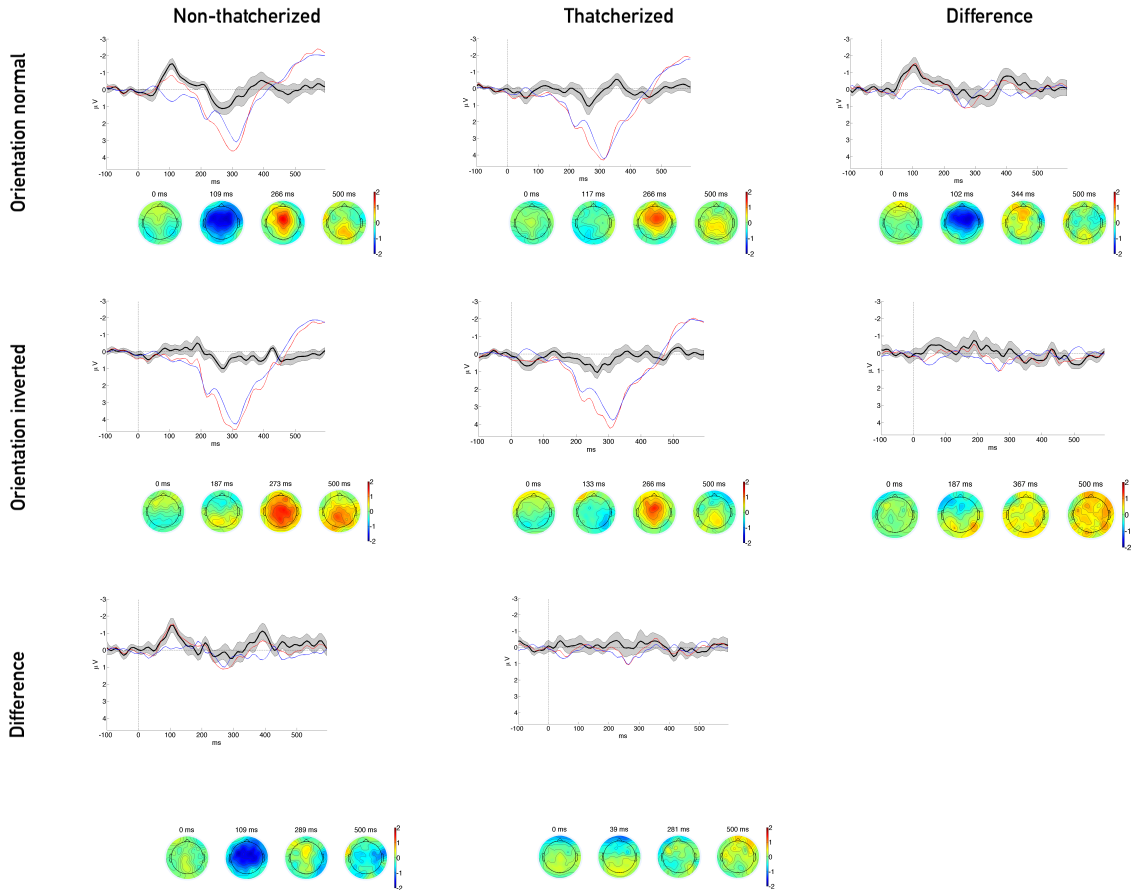


Fig.12. ERPs due to standard (red) and deviant (blue) stimuli, and the corresponding difference wave (black) with standard error of mean (grey shading) at electrode Fz. Scalp plots represent the distribution of potentials at baseline (0 ms), at the negative and positive peaks of the difference wave and at 500 ms.

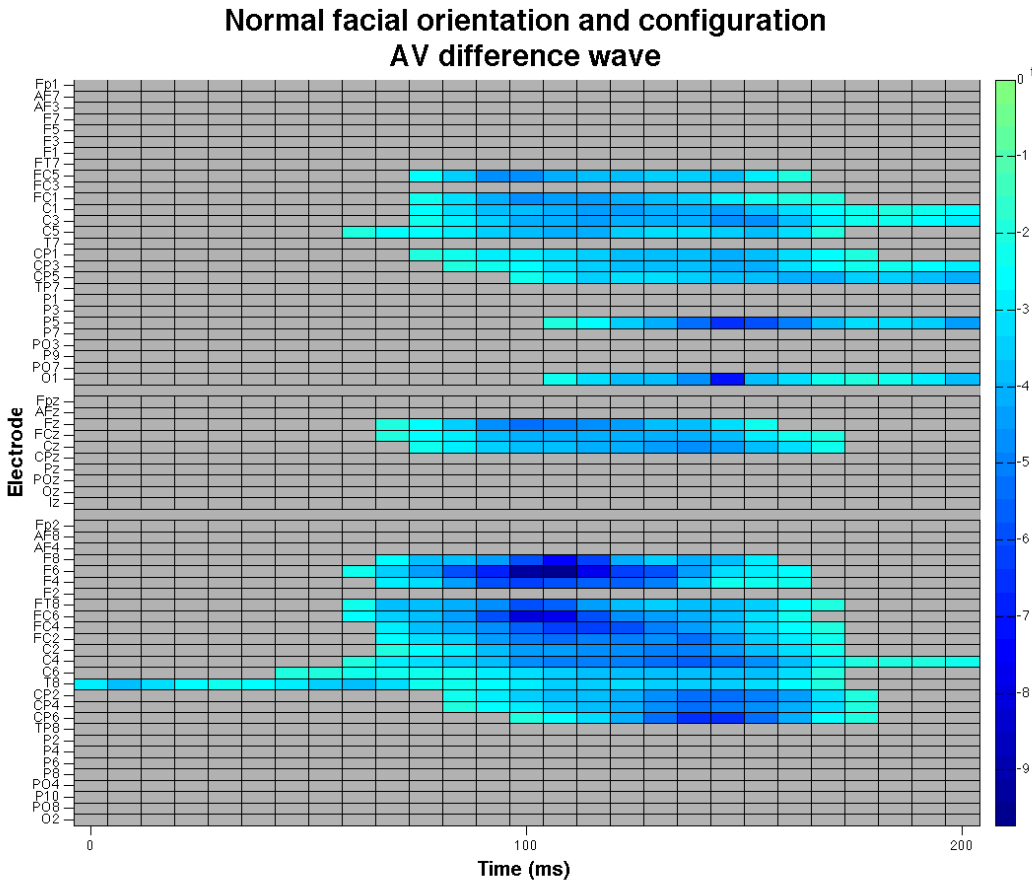


Fig. 13. Plot of t-scores for MMN due to upright face, upright mouth audiovisual stimuli resulting from a repeated-measures, one-tailed clustered permutation test (for a detailed description, see Publication B in the Appendix and Groppe et al., 2011). Colored areas represent t-scores exceeding the critical t-score of 1.83, thus being deemed statistically significant.

The behavioral response thus follows the pattern stated by Rosenblum and colleagues (2000). The main difference being that all Thatcherized and inverted stimuli produce even lower rates of incorrect auditory identifications.

As in Experiment 1, the necessary condition for using MMN as a measure of the strength of visual influence is that stimuli with an unmanipulated face – which resulted in a strong McGurk illusion in behavior - produce a convincing MMN component. For Experiment 2, we optimized stimuli and stimulus delivery set-up

for producing MMN. Results revealed a deep and consistent MMN response for upright stimuli with normal configuration (see Figure 12). The difference wave in this condition is characterized by a scalp distribution typical for both auditory and audiovisual MMN responses, covering a wide area from parietal over central to frontal electrodes. The latency of the response was also within the expected range for both auditory and audiovisual MMN, with a clear difference wave starting at 80 ms and ending at approx. 200 ms. The differential potential reached significant levels across a wide range of electrode in the 100 to 180 ms interval (see Figure 13).

Compared with results in Experiment 1, unmanipulated stimuli in Experiment 2 provided a clear and unambiguous MMN response in both scalp distribution and latency range.

Turning towards the Thatcherized stimulus in normal orientation, no MMN was observed. Here, inversion of the mouth segment effectively eliminated the differential potential. In comparison with the behavioral findings, this corresponds well to a reduction in influence of the visual stimulus on the auditory percept. Thus far, the effect of Thatcherization upon audiovisual integration is validated by the MMN response.

According to the McThatcher effect observed in the behavioral findings, responses to inverted face stimuli – with and without Thatcherization – should still reflect considerable levels of visual influence. However, MMN responses for both classes of inverted orientation stimuli were absent. Inversion of the face effectively eliminated any differential response. Comparing MMN responses for Thatcherized and non-Thatcherized inverted orientation stimuli revealed no differences (see Figure 12).

3.3.4 Summary of findings

Behavioral findings of Experiment 2 replicated the McThatcher effect as observed first by Rosenblum and colleagues (2000). Our aim here was to use McGurk-driven MMN as a means of verifying such behavioral findings in the neurophysiological domain. Looking at the effect of Thatcherization in the upright facial orientation in isolation, the elimination of the MMN corresponds to

the behavioral response. When considering responses to inverted orientation stimuli in isolation, Thatcherization did not yield any difference in MMN. This corresponds to the uniform behavioral responses with inverted orientation stimuli. However, in keeping with the McThatcher effect, inverted stimuli still supported a considerable visual influence on the auditory percept in the behavioral task. The MMN responses to inverted stimuli did not correspond to this.

Supplementary analysis on the first quarter of each condition

This lack of correspondence is surprising. In our efforts to understand this, we investigated possible causes for a reduced MMN response. Fatigue and adaptation may play a role in reducing amplitude of the MMN potential (Lang et al., 1995). The duration of the full EEG recording was long, which may possibly have led to fatigue in participants, in part because they were instructed to look at the screen during all audiovisual sessions. Adaptation to the auditory stimulus, which was the same in all audiovisual trials, may also have influenced recordings. As these factors could have reduced the MMN response in later parts of the recording, responses during the first periods of each condition may show a different pattern, than the result in the full recordings. To investigate this, we conducted a further analysis on the first quarter of each condition. We thus produced ERPs and generated differential potentials for the first 300 trials of each block (see Figure 14).

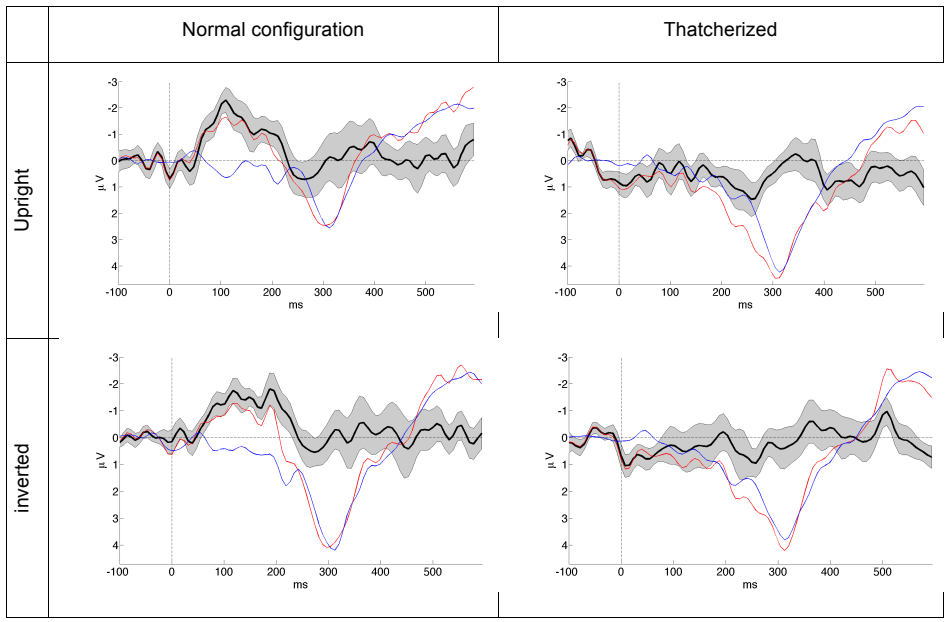


Fig. 14. ERPs and differential potentials generated from the first 300 trials, recorded at electrode Fz. Blue lines represent ERPs due to standard stimuli. Red lines represent ERPs due to deviant stimuli. Black lines represent the differential potential produced by subtracting standard ERPs from deviant ERPs. Shaded areas represent standard error of mean of the differential potential.

Here, upright stimuli present the same pattern as for the full dataset: normal configuration produces an MMN response, whereas Thatcherization eliminates this. Interestingly, in the inverted orientation stimuli, normal configuration stimuli produces a considerable MMN wave, which reaches the same level as for upright orientation, normal configuration stimuli. This MMN response is not present in the inverted orientation, Thatcherized condition. In other words, both normal configuration conditions produced an MMN response during the first 300 trials, which is eliminated in both Thatcherized conditions.

3.4 Discussion

Findings of Experiment 1 and 2 both suggest an influence of face perception upon behavioral and neural responses. In both experiments, behavioral responses replicate findings of Rosenblum and coworkers (2000). This indicates that facial configuration information indeed interacts with speech perception.

Although behavioral findings tightly repeat the pattern in the Rosenblum's behavioral study, the MMN responses to the same stimuli are ambiguous.

Whereas the eliminated MMN for Thatcherized stimuli in upright orientation are easily interpreted in terms of our hypotheses, in Experiment 2 we also expected some level of MMN responses to stimuli with inverted orientation. Here, analysis of the full dataset produced no MMN response for any of the inverted conditions.

McGurk-MMN as a direct measure of audiovisual binding

If using the McGurk-driven MMN in these findings as a yardstick of audiovisual integration, then the response to inverted orientation stimuli can only be interpreted as signifying that integration is eliminated. However, audiovisual integration in inverted orientation incongruent syllables has been found by many studies. Before accepting this strict interpretation, the nature of McGurk-driven MMN needs to be considered.

The production of auditory MMN is well-researched (Garrido et al., 2009; Lang et al., 1995; Näätänen and Alho, 1995). For instance, auditory MMN amplitude co-varies with the level of perceived stimulus difference in the deviant stimulus. A corresponding mechanism for generation AV MMN is unknown as of yet. Due to this, it is unclear how a reduction in McGurk illusion strength (as observed in Experiment 2 for both inverted conditions) would be reflected in the McGurk-MMN. If the McGurk-driven MMN responds non-linearly to differences in behavioral McGurk strength, this would explain the elimination of the MMN in inverted and/or Thatcherized conditions. As is represented in Publication B (see Appendix), when looking at data from subjects with strong behavioral McGurk responses, there is a correlation with McGurk-driven MMN. This may indicate that a strong McGurk illusion is needed to generate an MMN at all. This specific property of the audiovisual MMN has, to our knowledge, not yet been targeted in any experimental study. However, reviewing published studies based on this effect, subjects are often included in measurements of McGurk-MMN on basis of a strong behavioral McGurk response. Studies using McGurk-MMN, which also report the level of corresponding behavioral McGurk responses, report near 100% McGurk illusion with incongruent stimuli (Stekelenburg and Vroomen, 2012), or, “a strong McGurk illusion” (Saint-Amour et al., 2007). Kislyuk and colleagues (2008) exclude subjects with “a weak McGurk effect”.

Due to the concerns for experiment duration, only audiovisual conditions were tested in the MMN experiment. However, the processing of Thatcherized and/or inverted visual speech could perhaps also be better understood if including neural responses to visual-only versions of the same stimuli.

Alternative interpretations

The relation between the behavioral McGurk response and Mcgurk MMN is unknown also in another sense. For instance, the McGurk response with normal visual speech may use resources that generate an MMN when stimuli are incongruent, while manipulated visual speech stimuli, although generating the same behavior, may recruit other resources. Audiovisual integration with normal stimuli may be pre-attentive and early (thus evoking MMN), whereas degraded visual signals may use different, later resources, perhaps involving attention (thus producing no MMN, but potentially other, later responses).

When considering the supplementary finding on the first 300 trials of each condition, these first of all indicate, that the McGurk-MMN is sensitive to experiment duration or repeated stimulus presentation, be it due to fatigue or adaptation. Second, these results suggest a possible specific sensitivity to Thatcherization, or segment interrelation violation. This could indicate that the inverted, normal face and the inverted Thatcherized face recruit different neural processes, while the behavioral outcome is highly similar. This finding is also reflected in Rotshtein and colleagues' (2001) fMRI study of static visual faces using the same segment and orientation manipulations, where responses to normal faces contrast with responses to Thatcherized, while unaffected by orientation. One possible explanation here is that Thatcherized, misconfigured stimuli represent a degradation of the visual signal, which in turn reduces integration at an early stage (Garrido et al., 2009; Näätänen et al., 1978). Still, binding of the inverted, Thatcherized stimuli are observed in behavioral responses. This could, however, be the result of a later binding process, different from the one, which elicits the McGurk-MMN. If so, the behavioral integration responses register two different processes in the two inverted stimulus categories, of which only produce an MMN response.

Our supplementary findings here call for further investigation of the relation between face perception and audiovisual integration. While the MMN may validly register early, pre-attentive integration, other ERP or EEG measures of audiovisual integration in Thatcherized and/or inverted audiovisual speech may shed light onto integration at other stages.

4 The temporal window of audiovisual integration of speech

4.1 Background

Temporal coincidence supports integration of acoustic and visual signals (Stein and Meredith, 1993). This is also the case for binding of co-occurring acoustic and visual speech into a unified, audiovisual speech percept. If measuring temporal relations between acoustic and visual onsets in a multisensory stimulus, e.g. by means of an oscilloscope, temporal coincidence is easily defined. However, the human sensory system does not possess a similar device, capable of clocking or monitoring time on an absolute scale. Thus, establishing intersensory simultaneity is not a simple task of registering or clocking incoming signals.

Multiple factors influence the relative timing of vision and hearing. First, acoustic and visual speech stimuli differ in their physical, temporal properties before reaching our sensory organs. Second, acoustic and visual pathways from sensory periphery to primary sensory cortices have different processing speeds and behaviors. Third, perception of intersensory timing further requires another processing step, which may be more or less tolerant to temporal shifts between vision and hearing (Vroomen and Keetels, 2010).

Such properties of audiovisual stimuli and their perception may be involved in the relative tolerance towards audiovisual asynchrony in speech perception. This tolerance is often represented as a *temporal window* of simultaneity perception, of audiovisual integration, or, of specific crossmodal effects. The temporal window represents the interval of asynchronies (visual leads and acoustic leads) within which perception is unaffected by asynchrony.

4.1.1 Temporal properties of audiovisual speech

Acoustic and visual signals differ widely in physical properties. Under optimal listening and viewing conditions, the travelling speeds of sound and light differ to a degree (speed of sound 343 m/s, speed of light approx. 300,000,000 m/s), which may introduce considerable asynchrony dependent on distance. At a distance of e.g. 15 meters from a speaker, the travelling time from source to

sensory organ is negligible for the visual signal, whereas the acoustic signal is delayed by 44 ms.

Upon reaching the sensory periphery, processing times until the signal reaches the respective primary sensory cortices also differ between hearing and vision. Here, the auditory pathway is the faster and only approx. 10 ms are required for a signal to reach primary auditory cortex from the cochlea (Moore, 2012). In comparison, the signal that leaves the retina takes about 50 ms to reach primary visual cortex (Pöppel et al., 1990a). If taking only these measures into account and further assuming that intersensory synchrony is perceived at the level of the primary sensory cortices, synchrony would be perceived for the distance at which differences in travelling time of sound and light compensate for the differences in processing time along auditory and visual pathways. Thus, for audiovisual objects the so-called "horizon of simultaneity" would be at approx. 12 meters (Pöppel et al., 1990b). In addition, auditory and visual pathways may exhibit different temporal resolutions, or, metaphorically speaking "sampling rates", depending on oscillatory firing patterns in their respective neural circuitry (Kösem et al., 2014; van Wassenhove, 2009).

However, humans perceive audiovisual objects as simultaneous across a great interval of distances. For stimuli occurring in our natural environment, only extreme distances introduce an intersensory lag sufficient to *not* support binding of sight and hearing (e.g. observation at sea level of passenger planes passing at cruising height). Thus, our perceptual system has a certain tolerance towards asynchronous audiovisual stimuli. Or, audiovisual binding occurs not only at perfect synchrony, it is rather effective within a temporal window of intersensory delays.

This ability to tolerate asynchronous stimuli could be given different explanations. It could be a matter of a sheer tolerance towards delays, a relative temporal insensitivity of the nervous system. Another possible explanatory factor would be that perception adapts to asynchrony. This could either be due to information about distance to the stimulus. Or, it could be adaptation to intersensory lag based on prior experience with a specific stimulus or perception at different distances (Vroomen and Keetels, 2010). The causal explanation of

audiovisual integration across intersensory delays, however, is outside the scope of this dissertation. But the discussed asynchronies inherent in the steps from multimodal source to unified percept all underline that perception must have some tolerance, or, compensation for temporal shifts. The question is then, if this tolerance can be given a characteristic.

4.1.2 Behavioral methods in estimating asynchrony tolerance in speech perception

The study of perception of asynchronous stimuli has employed a variety of measures of audiovisual integration. These methods mainly divide into direct measures of temporal perception and indirect measures e.g. the sensitivity of other markers of audiovisual binding to asynchrony.

The simplest direct measure of asynchrony perception is behavioral simultaneity judgment (SJ), In such tasks, subjects are basically required to respond to whether a bimodal stimulus was perceived as simultaneous (or, non-simultaneous) (van Wassenhove et al., 2007). The method of matching for simultaneous perception could be understood as an inverted SJ task (Dixon and Spitz, 1980). In such tasks, subjects are asked to shift the onset of strongly asynchronous bimodal speech to the point where the stimulus is perceived as simultaneous.

Temporal order judgment (TOJ) tasks ask subjects which modality preceded in stimulus presentation. This task is more difficult than SJ, and it could be argued to contain a dual task: It requires a judgment of simultaneity and further a judgment of modality sequence. Due to this added difficulty, TOJ tasks often require training of subjects (Vroomen and Keetels, 2010).

As an alternative to these direct measures of temporal perception, the sensitivity to audiovisual asynchrony can be revealed by behavioral responses that indicate bimodal integration and thus indirectly, tolerance to asynchrony. The strength of McGurk responses to incongruent speech stimuli with different levels of asynchrony can be used as such a measure of temporal tolerance (Munhall et al., 1996; van Wassenhove et al., 2007).

A further effect, which also have been used in this way is the intelligibility advantage observed when a coinciding visual signal benefits encoding of a weak

auditory speech signal. In this way, Grant and Greenberg (2001) tested intelligibility of acoustic speech with a strongly depleted spectral content, while presenting visual speech at different onset lags and leads with respect to the acoustic stimulus.

An important factor in studying temporal integration of asynchronous audiovisual speech is adaptation. If being presented with a general asynchrony between acoustic and visual signals, adaptation may shift the temporal window of integration towards the direction of the asynchrony adapted to. Navarra and colleagues (2005) exposed subjects to asynchronous AV speech before performing SJ tasks of AV speech stimuli. Exposure to asynchronous speech, with SOA within the window of tolerable asynchronies extended the tolerance to asynchrony. Prior exposure to asynchronous speech with an SOA outside this window, however, did not elicit this response.

4.1.3 Behavioral estimates of the temporal window for audiovisual integration in speech perception

Tolerance towards audiovisual asynchrony has been studied across the entire span of modern psychology (if counting Wundt's complication experiment as a variety of temporal audiovisual integration task, cf. Wundt, 1862). The evaluation of this tolerance depends on its definition and the methods applied to its measurement.

In an early experiment, Hirsch and Sherrick (1961) investigated the ability to detect asynchrony between non-speech stimuli across pairings of modalities. Their findings suggested that bimodal events with more than 20 ms onset asynchrony were detected as asynchronous, though much longer asynchronies were required for subjects to detect the order of onsets in the compared modalities. At another extreme, Dixon and Spitz (1980) asked participants to temporally match acoustic and visual speech signals to the point of perceived simultaneity. Here, matching for audiovisual simultaneity occurred with acoustic onset delay of up to 250 ms.

The different outcomes of Dixon's and Hirsch's experiments highlight at least two central variables of bimodal temporal perception: First, the measure of

simultaneous perception may alter the tolerance window estimate widely. Second, estimates may differ between non-speech and speech stimuli. Since this dissertation investigates audiovisual speech perception exclusively, findings on tolerance to audiovisual asynchrony in non-speech stimuli will not be covered in detail.

Most studies have observed a higher sensitivity to visual lags than acoustic lags. This is expressed in temporal window estimates as an asymmetrical arrangement of the window around the point of synchronous presentation. While the shape of the window varies across studies (dependent e.g. on temporal resolution in experimental designs), a steeper fall of audiovisual responses is often observed for visual lags, whereas the function for visual lags usually produces a lower slope (see Figure 15).

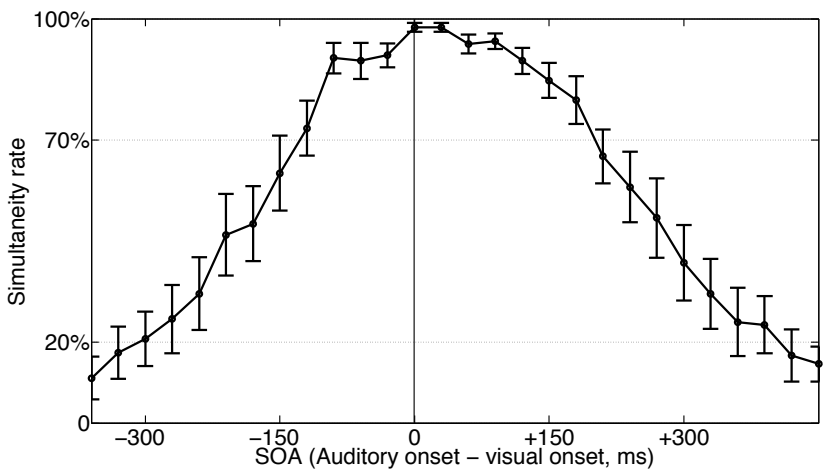


Fig.15. Data exemplifying a typical behavioral response pattern in an SJ task with an audiovisual syllable. The sensitivity to visual lag (negative SOAs) is higher than to acoustic lags (positive SOAs). The response fall-off is steeper on the visual-lag side than on the acoustic-lag side.

Using SJ as measure of perceived simultaneity, Conrey and Pisoni (2006) estimated a temporal window of SOAs between -144 and +254 ms. Van Wassenhove and colleagues (2007) applied a similar measure in their Experiment 2, findings suggesting a window of perceived simultaneity in the interval of approx. -80 to +130 ms.

Studies using the McGurk illusion as a measure of audiovisual integration have yielded similar, wide temporal window estimates. Here, Munhall and colleagues (1996) found that the McGurk illusion can be evoked for SOAs ranging from -60 to +240 ms, while the form of the response curve was v-shaped towards a point of maximal McGurk responses at +60 ms. In comparison, van Wassenhove and colleagues (2007) also tested McGurk responses to incongruent stimuli and found that a narrower window of SOAs ranging from -34 to +174 ms. The response curve, however, had the shape of an inverted u. Thus, the strength of the illusion reached a plateau with a uniform level of McGurk responses within this window.

How does the temporal window as estimated with SJ tasks relate to the estimate based on McGurk response sensitivity to asynchrony? Van Wassenhove and coworkers (2007) collected both types of data, and added a further SJ task with the incongruent phonemes used for the McGurk-based window estimate. Interestingly, SJ performance with incongruent stimuli for one of two incongruent syllables resulted in a narrower window estimate (ranging from SOA -37 to +122). This may indicate that the tolerance towards asynchrony in audiovisual speech partly relies on correlation between acoustic and visual (e.g. lip opening) envelopes (van Wassenhove et al., 2007). When acoustic and visual articulatory cues are mismatched, audiovisual binding may be weaker due to weaker articulatory matching, resulting in a greater sensitivity to temporal misalignment.

Using the intelligibility benefit associated with audiovisual over auditory speech, Grant and Greenberg (2001) presented spectrally depleted acoustic speech with asynchronous visual speech. In their findings, auditory intelligibility is facilitated for SOAs in the range +/-400 ms. For negative SOAs (visual lag), the intelligibility benefit steadily increases from -400 ms to 0 ms SOA. However, for positive SOAs (acoustic lag), a plateau of a uniform intelligibility advantage is upheld from 0 ms SOA to +200 ms. After this point, the advantage steadily decreases until it vanishes at the +400 ms SOA. Interestingly, audiovisual intelligibility here exhibits an asymmetrical shape around synchronuous presentation, with a

plateau of uniform performance as observed by van Wassenhove and colleagues (2007).

4.1.4 Neural estimates of the temporal window for audiovisual integration in speech perception

Tolerance to asynchronous audiovisual speech has also been targeted in the neural domain. A variety of methods have been used, including PET, fMRI, ERP and EEG phase entrainment measures.

Recording PET responses to audiovisual speech in synchrony or with a SOA of -240 ms (visual lag), Macaluso and colleagues (2004) targeted responses to stimuli which would either support or clearly not support audiovisual integration. The authors found a differential response to synchronous stimuli in the left superior temporal sulcus and in the fusiform gyri. Stevenson and coworkers (2010) measured BOLD response in fMRi to a series of audiovisual speech samples with SOAs ranging from -400 to +400 ms in 100 ms steps. Findings suggest that two distinct subregions of superior temporal cortex were specifically responsive to synchronous stimuli.

ERP studies of audiovisual speech perception have revealed that audiovisual responses differ from both responses to auditory and visual speech respectively, and from the aggregate of the unimodal responses. Van Wassenhove and coworkers (2005) found that amplitude and latency of the N1 component was reduced with audiovisual speech compared with auditory-only speech, a finding that was consistent with a study by Besle and colleagues (2004). Presenting audiovisual speech at synchrony and at an SOA of -200 (visual lag), Pilling (2009) found that only synchronous speech exhibited suppression of N1 amplitude.

Changes in the phase of the electrophysiological response may indicate alterations of processing. Kösem and colleagues (2014) used this approach and asked how phase patterns evoked by synchronous and asynchronous non-speech would relate. Interestingly, oscillations entrained by repeated asynchronous stimuli (with SOAs of +200 and -200 ms) exhibited adaptation towards the phase pattern evoked by synchronous stimuli. The adaptive pattern mirrored perceived simultaneity in behavioral measurements, and thus

interpreted as indicating a neural compensation for temporal offsets in stimuli. This method, however, has not yet been applied to audiovisual speech stimuli.

4.2 Experiment 3: Electrophysiological correlates of the temporal window of audiovisual integration in speech perception

4.2.1 Motivation

As reviewed above, the temporal window of audiovisual integration has been targeted by a series of behavioral studies. The relative ease of participation in behavioral paradigms and their lower general duration allow presentation of stimuli at many levels of asynchrony before fatigue impedes performance. Behavioral studies thus may enjoy high temporal resolution (cf. e.g. van Wassenhove et al., 2007), enabling fine-grained estimates of asynchrony tolerance. In comparison, neurophysiological methods generally are limited to testing at few asynchronies (in many studies only three SOAs, cf. Kösem et al., 2014; Macaluso et al., 2004; Pilling, 2009; whereas nine SOAs are presented in Stevenson et al., 2010). This introduces a discrepancy between behavioral and neural experimentation: Many studies of neural processing present a synchronous stimulus and a limited number of asynchronous stimuli at extreme asynchronies where audiovisual integration is not supported in behavior.

Due to this, the precisely drawn temporal window of audiovisual integration based on behavioral data has not yet been paralleled or verified in neurophysiological research. Thus, the neural temporal window of integration is only known at its central and extreme points. The aim of Experiment 3 (which is described in greater detail in Publication C in the Appendix) is to get a step closer to this by asking if the behaviorally estimated integration window is reflected in neural responses. Specifically, we ask how neural responses correlate with integration responses at select asynchronies on the behaviorally estimated function.

Previous studies have revealed that the tolerance to asynchrony may vary across subjects. Speech presented at a specific SOA may be perceived and integrated differently among individual subjects. To circumvent this problem and meaningfully target the correlation between a neural response and behavior, we

chose to select individual SOAs for the neurophysiological measurement on basis of individual behavioral data.

4.2.2 Experimental design

As the aim is the correlation between a behavioral estimate and neural responses, our study takes off from a behavioral estimate of the temporal window of audiovisual integration. Here, we choose the McGurk illusion as our indicator of integration. Thus, subjects first identify an audiovisual incongruent bisyllable (acoustic /tabi/ + visual /tagi/), presented at SOAs ranging from -600 to +600 in 40 ms steps. Each asynchrony was represented by 20 presentations in the task, totaling 620 trials in randomized order. On basis of the proportion incorrect auditory responses (i.e. McGurk responses) at each asynchrony level, individual temporal window estimates were generated by fitting a asymmetrical double sigmoidal curve (cf. van Wassenhove et al., 2007). However, for the comparison of behavioral and neurophysiological measures to be meaningful, the presence of a behavioral response pattern allowing the estimation of an asymmetrical double sigmoidal curve was required. Subjects whose response pattern did not lend itself to such an estimation procedure (i.e. no or very limited McGurk responses, or, a random distribution of McGurk responses over asynchrony levels) were excluded from participating in the neurophysiological part of the experiment. Their behavioral data was also excluded from the analysis. Twelve subjects were excluded on this criterion.

As a neural indicator of audiovisual integration, we again chose the McGurk-driven MMN. Here, we hypothesized that the McGurk-driven MMN would mirror the pattern of behavioral McGurk responses at the different SOAs. Thus, for each chosen SOA to enter the MMN measurement part of the experiment, standard stimuli were congruent bisyllables (audiovisual /tabi/), whereas deviant stimuli were incongruent (acoustic /tabi/+visual /tagi/). Since individual behaviorally estimated temporal window functions vary, the SOAs at which we measured McGurk-MMN was chosen individually. Due to this, the audiovisual MMN responses were not aligned at specific SOAs, but rather at specific points on the individual temporal window functions.

Due to the nature of MMN experimentation, duration was a concern. Generation of a clean MMN response requires hundreds or thousands of trials. To keep duration as low as possible while still collecting meaningful data, we chose to only test on one side of the temporal window, namely the acoustic lag side, primarily due to its lower slope in typical temporal window estimates. The lower slope allows a longer interval between chosen SOAs, resulting in a more pronounced stimulus difference between SOA conditions.

We further chose only three SOAs, at the point of maximal behavioral McGurk response level, at the 70% level and at the 20% level. In recordings of auditory MMN, strength of the MMN differential potential correlates with the level of perceived difference in the deviant stimulus (Näätänen et al., 2007). Thus, we hypothesized that the MMN response would reflect the level of McGurk responses at the chosen SOAs.

However, individual propensity to generate a MMN response varies (Lang et al., 1995). To avoid contamination of the result with data from subjects generally producing a weak or no MMN response, we ran a separate control condition, measuring pure-tone MMN due to 1000 Hz and 1200 Hz tones of 100 ms duration presented at 65 dB(A) SPL. We defined an exclusion criterion as a pure-tone MMN of lower amplitude than -1 μV. However, no subjects were excluded on this criterion.

A well-known complication when using ERPs as a measure of visual influence on the auditory percept is that a visual stimulus difference in itself may evoke a differential response. On basis of this, a visual MMN analogue has even been proposed, generated by presenting deviating stimuli in an oddball pattern similar to auditory MMN paradigms (cf. e.g. Czigler, 2007). A McGurk-driven MMN response would be hard to discriminate from an MMN-response due to the visual stimulus deviance alone. To alleviate this, we included a visual-only (VO) condition, presenting a similar oddball sequence of congruent and incongruent speech. The potentials evoked by these stimuli were then used to correct the audiovisual ERPs for purely visual responses.

When studying the influence of vision upon the auditory percept, i.e. in auditory identification tasks, or, with auditory ERP components such as MMN, the

direction of eye gaze is a central concern. If participants disregard the visual stimulus, i.e. by closing their eyes or by directing gaze away from the stimulus display, auditory identification is still possible and an auditory ERP would still be evoked. Such behavior would produce null effects. To limit the influence of this confounding factor, we introduced a secondary visual detection task. In this task, subjects had to detect a blink in a white dot overlaid the nose of the talking face. These visual catch trials made up 5% of the total number of trials in the MMN tasks. Data from catch trials were omitted from both datasets.

4.2.3 Results

Behavioral results

Behavioral identification performance with incongruent speech (see Figure 16) largely followed the pattern of previous studies (Munhall et al., 1996; van Wassenhove et al., 2007).

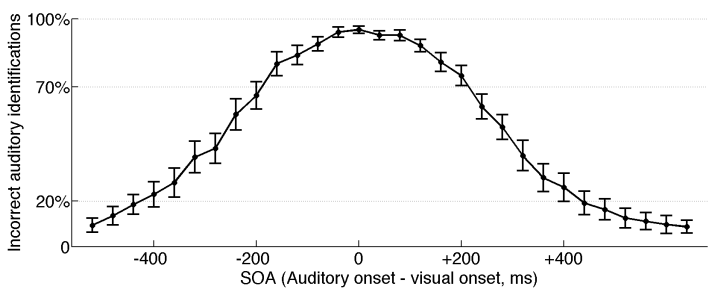


Fig. 16. Behavioral McGurk illusion strength, as represented by average proportion incorrect auditory identifications indicating audiovisual integration at different audiovisual SOAs. Points represent average proportion incorrect responses, error bars represent standard errors of the mean.

In comparison with prior research, responses follow the pattern observed by van Wassenhove (2007) with a wide temporal window, slightly asymmetrically arranged over the synchronous condition. A bonferroni-corrected multiple comparisons test revealed that the rate of incorrect identifications was not significantly different between SOAs of -120 and +200 ms. The bell-shaped response curve thus contains a large plateau of uniform responses, asymmetrically arranged across over the point of synchronous presentation. Here, responses have greater similarity with findings of van Wassenhove and

colleagues (2007) than with the v-shaped response curve found by Munhall and coworkers (1996).

Mismatch negativity measurements

Performance in the secondary, visual detection task was at a high level across all parts of the experiment (mean detection rate 97.4%, SD = 1.7%). This indicates that participants directed their gaze towards the visual stimulus as instructed to.

Figure 17 shows the average ERPs due to audiovisual, visual-only and audiovisual subtracted visual-only responses. VO and AV responses are epoched and baselined to onset of the visual stimulus. The VO standard and deviant ERPs are then subtracted from the AV standard and deviant ERPs. The resulting AV-VO ERPs are then baselined and epoched to auditory onset before calculation of the final differential potential, representing the AV-VO MMN response. Note here, that due to SOAs in the AV condition being chosen individually, that individual offsets between visual onset and acoustic onset differ. Thus, the average ERPs in AV and VO do not align with AV-VO ERPs.

As can be seen in Figure 17, an AV-VO MMN potential was generated in the maximal McGurk response condition with the maximal amplitude approaching -1 μV. This MMN response reached significant levels at centro-parietal and parieto-occipital electrodes in the interval 350-584 ms (see Publication C in the Appendix for statistics). AV-VO MMN recorded at the 70% McGurk response point produced a less clear MMN response.

It did, however, reach significant levels in the interval 100-350 ms at parietal, parieto-occipital and occipital electrodes. AV-VO MMN responses recorded with the SOA producing 20% behavioral McGurk responses did not yield any recognizable MMN wave.

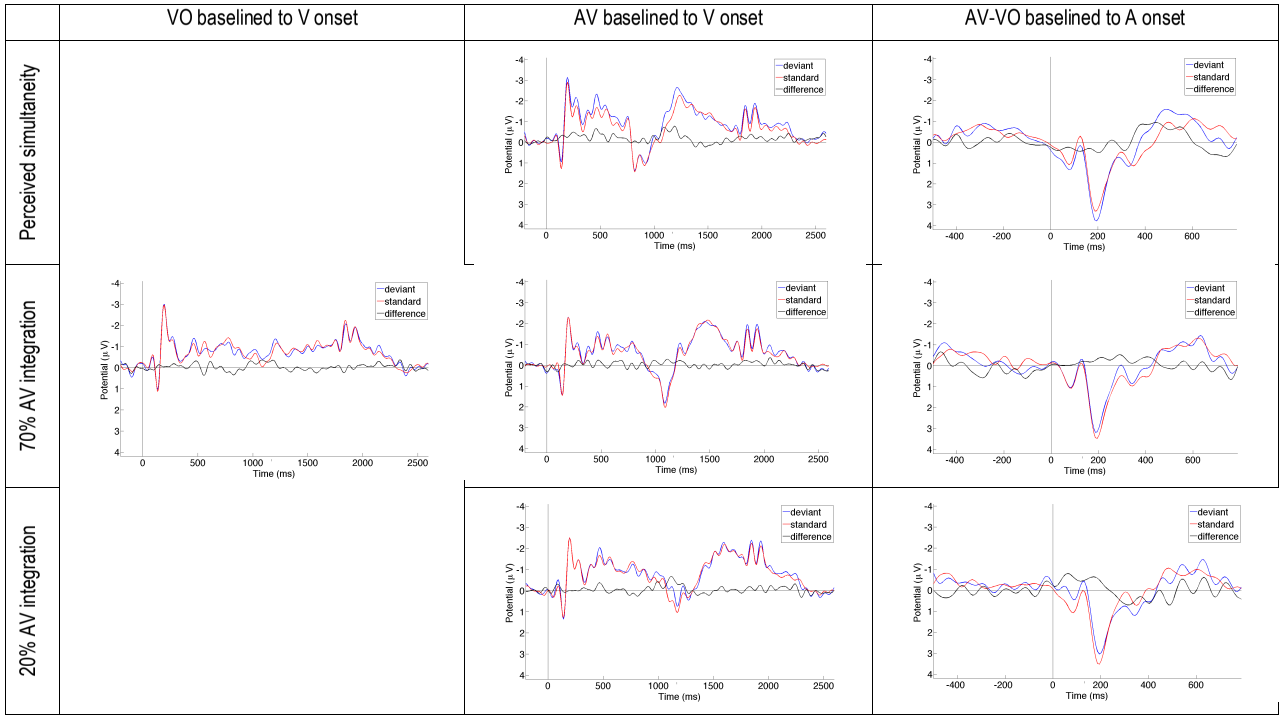


Figure 17: ERPs and difference waves at a parieto-occipital electrode, from the VO and AV conditions, and the VO-corrected AV condition (AV-VO). The blue line represents average ERP due to deviants, the red line represents average ERP due to standards, and the black line represents the MMN differential potential. VO and AV potentials are epoched and baselined to the visual onset. The AV-VO potentials are epoched baselined to the auditory onset. Note that the interval between visual and auditory onset is subject to interindividual variability. Due to this, the average ERPs vary not just in baseline but also in waveform between VO, AV and AV-VO.

4.3 Summary of findings and discussion

The behavioral sensitivity of the McGurk illusion to audiovisual asynchrony largely mirrors findings in previous studies: The response curve is bell-shaped with a plateau representing a uniform level of audiovisual integration responses. This may indicate that integration as seen in the McGurk illusion is insensitive to asynchrony when within this interval. The width of the plateau, however, is wider than reported in similar studies (van Wassenhove et al., 2007). The curve is asymmetrically arranged around the SOA representing bimodal synchrony. Furthermore, the visual-lag side of the curve displays a slightly steeper slope than the acoustic-lag side. This is aligned with previous studies and is usually interpreted as a higher sensitivity towards visual lags than acoustic lags (Conrey and Pisoni, 2006; Munhall et al., 1996; van Wassenhove et al., 2007).

The MMN findings in comparison are inconclusive. Although an AV-VO MMN was generated in the condition corresponding to maximal behavioral McGurk response, the distribution of the potential over the scalp does not follow a standard pattern for MMN responses. For both AV MMN and auditory MMN, a wide distribution centered on central and fronto-central electrodes would be expected (see e.g. Experiment 2, AV MMN for normal-face stimuli). Here, the scalp distribution of the AV-VO MMN is concentrated in occipital and parietal regions, which makes it hard to interpret as an auditory response.

AV-VO MMN for stimuli with SOA corresponding to the 70% behavioral McGurk response point shares this problem. This condition furthermore shows a remarkably early MMN response, which begins before acoustic onset. This could be interpreted in two ways: First, it may signify an artifact, not generated by any perceptual response but rather e.g. by the subtraction of the visual-only ERP. Second, it may indicate a modulation of auditory activity, even though the corresponding acoustic syllable is not yet presented. The latter interpretation may find some support in findings of Möttönen (2002), suggesting that silent visual speech modulates auditory cortex. However, the distribution of the AV-VO MMN response is still concentrated over parietal and occipital regions, which makes the latter interpretation questionable.

Our main question, whether the asynchrony sensitivity function can be seen in AV-VO MMN responses at select SOAs chosen on the function, thus is hard to answer on basis of the present findings. However, the difficulty in producing meaningful MMN responses leads to some secondary, methodological considerations.

Like in Experiment 1 and 2, the use of the AV and AV-VO MMN in representing a graded audiovisual integration response comes into question. As in the previous experiments, the relation of MMN response strength to behavioral audiovisual integration response levels is unknown. The problems in using MMN as a gauge of audiovisual integration described in the discussion section of Experiments 1 and 2 also apply here.

5 Conclusion

The experimental findings presented above all bear on the methodological question regarding the use of McGurk-driven MMN as a measure of the strength of audiovisual integration. But the experiments also shed light on two aspects of audiovisual speech processing.

In Experiments 1 and 2, behavioral findings replicated previous findings of Rosenblum and colleagues (2000), that face processing does indeed influence the integration of the visual signal into the auditory speech percept. The MMN effects were not congruent with behavioral findings. This calls for more complex explanations. When presenting AV speech with unmanipulated face configuration, a clear AV MMN response was generated, highly similar to both auditory and AV MMN described in prior studies. This potential was eliminated for all other face configurations. Thatcherized and/or inverted stimuli yielded reduced behavioral effects, which provided two possible interpretations: Either that AV integration (as indicated by the AV MMN) is effectively eliminated by any face manipulation. But this would contradict a number of behavioral studies, finding moderate to high McGurk illusion levels for rotated faces (e.g. Bertelson et al., 1994; Jordan and Bevan, 1997; Massaro and Cohen, 1996). Or, that AV MMN requires a very strong perceptual McGurk illusion for being evoked.

The supplementary analysis of Experiment 2 EEG data showed that for shorter experiment durations or fewer stimulus exposures, the MMN was only eliminated with misconfigured, Thatcherized faces. These stimuli represent a strongly degraded visual signal, which potentially could require a different, processing for being integrated into the speech percept. While integration may still be possible (as observed in behavior), the processes involved in producing it may be different and later than for normal faces, and thus not evoking an MMN potential. Thus, the variability in McGurk-MMN across face manipulations may here discriminate between sensory integration at different latencies, rather than signify that integration is eliminated. However, for reaching any conclusion on this, our findings of McGurk MMN must be supplemented with other measures, such as ERP measurements.

Furthermore, duration seems to play a role, whether introducing fatigue or adaptation. In future research based on McGurk-MMN, this factor should be considered. The recommendable duration for a block of McGurk-MMN stimulation may be shorter than the maximum 10 minutes often suggested for auditory MMN generation (Lang et al., 1995).

In Experiment 3, we attempted at verifying individual asynchrony tolerance functions by means of the AV-VO MMN. The behavioral data mirror previous similar studies closely. The AV-VO MMN results, however, were inconclusive. The question of how MMN strength relates to strength of the McGurk illusion can also be asked here. But two further specific factors may also have influenced Experiment 3:

The AV MMN measurements were meant to verify behavioral findings at individual SOAs. However, even though stimuli and apparatus were kept unchanged between measurements of behavioral and MMN responses, the two parts differed strongly in the stimulus sequence. Here, the behavioral part employed a random sequence, each SOA being represented in 20 trials. The random sequence eliminated any adaptation effects. The AV MMN oddball sequence, on the other hand, repeated a single SOA in standard and deviant stimuli for a full block before changing to a different SOA. This of course is inherent in any MMN paradigm. But when experimenting with temporal effects, adaptation to a repeated asynchronous stimulus may influence the effect. Here, findings of e.g. Navarra and colleagues (2005) suggest that exposure to desynchronized AV speech may widen the interval of SOAs within which audiovisual integration (as observed in TOJ) is supported. An extension of the window of integration, however, should in theory confound our AV-VO MMN results by inducing a MMN response at SOAs where the non-adaptive behavioral paradigm would predict no MMN response.

Another factor, which we may speculate influenced our weak AV-VO MMN findings is that we chose SOAs on basis of individual behavioral findings. But are these reliable? Our specific design rests on the assumption that differences in individual responses to SOAs represent individual differences in neural

processing. But differences in behavioral responses could very well be due to other factors, such as e.g. attention.

Turning towards methodological conclusions, Experiments 1-3 gathered important insights into the use of McGurk-driven MMN for verifying behavioral measures of AV integration.

MMN is known as a powerful tool within cognitive and auditory research (Garrido et al., 2009; Näätänen, 2003). As reviewed above, it may also gauge audiovisual integration as evoked by the McGurk illusion. In the present experiments, we attempted to use the McGurk-driven MMN as a graded measure of the strength of bimodal integration. This showed some unforeseen properties of the audiovisual MMN. An important lesson from these efforts is the requirement of a better understanding of the relation between the strength of the McGurk illusion in behavior and the strength of the McGurk-driven MMN. Previous studies using the McGurk-MMN typically report strong corresponding behavioral effects. But the relation between the two domains has to our knowledge not yet been investigated. Future research using the McGurk-MMN would benefit strongly from a better understanding of this relation.

Future studies of neural correlates of audiovisual integration phenomena would benefit from supplementing McGurk-MMN with alternative methods, which measure bimodal integration in other ways. For instance, ERP studies could target later integration stages than the pre-attentive, early process reflected in MNN. On basis of behavioral findings, audiovisual integration has previously been proposed to take place in multiple stages (see Publication D below, cf. Schwartz et al., 2004). If so, McGurk-MMN may track early integration stages, whereas other methods could shed light on integration at other latencies.

6 References

Alho, K., Sainio, K., Sajaniemi, N., Reinikainen, K., and Näätänen, R. (1990). Event-related brain potential of human newborns to pitch change of an acoustic stimulus. Electroencephalogr. Clin. Neurophysiol. Potentials Sect. *77*, 151–155.

Alho, K., Woods, D.L., Algazi, A., and Näätänen, R. (1992). Intermodal selective attention. II. Effects of attentional load on processing of auditory and visual stimuli in central space. Electroencephalogr. Clin. Neurophysiol. *82*, 356–368.

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual Integration of Speech Falters under High Attention Demands. Curr. Biol. *15*, 839–843.

Arnott, S.R., and Alain, C. (2002). Stepping out of the spotlight: MMN attenuation as a function of distance from the attended location: NeuroReport *13*, 2209–2212.

Bernstein, L., Auer, E.T.J., and Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading.

Bertelson, P., Vroomen, J., Wiegeraad, G., and de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. In ICSLP 94, (Yokohama), pp. 559–562.

Bertelson, P., Vroomen, J., Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. *62*, 321–332.

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.

Böttcher-Gandor, C., and Ullsperger, P. (1992). Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval. Psychophysiology *29*, 546–550.

Bruce, V., and Young, A.W. (2012). Face perception (London ; New York: Psychology Press).

Buzsáki, G. (2006). Rhythms of the brain (Oxford ; New York: Oxford University Press).

Calvert, G.A., Campbell, R., and Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. *10*, 649–657.

Carbon, C.-C., Schweinberger, S.R., Kaufmann, J.M., and Leder, H. (2005). The Thatcher illusion seen by the brain: an event-related brain potentials study. Cogn. Brain Res. *24*, 544–555.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The Natural Statistics of Audiovisual Speech. PLoS Comput. Biol. *5*, e1000436.

Colin, C. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. *113*, 495–506.

Colin, C., Radeau, M., Deltentre, P., and Morais, J. (2001). Rules of intersensory integration in spatial scene analysis and speechreading. Psychol. Belg. *41*, 131–144.

Conrey, B., and Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J. Acoust. Soc. Am. *119*, 4065.

Czigler, I. (2007). Visual Mismatch Negativity: Violation of Nonattended Environmental Regularities. J. Psychophysiol. *21*, 224–230.

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

Eagleman, D.M. (2008). Human time perception and its illusions. Curr. Opin. Neurobiol. *18*, 131–136.

Eskelund, K., Tuomainen, J., and Andersen, T.S. (2010). Multistage audiovisual integration of speech: dissociating identification and detection. Exp. Brain Res. *208*, 447–457.

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. J. Neurosci. Off. J. Soc. Neurosci. *22*, 5749–5759.

Falchier, A., Schroeder, C.E., Hackett, T.A., Lakatos, P., Nascimento-Silva, S., Ulbert, I., Karmos, G., and Smiley, J.F. (2010). Projection from Visual Areas V2 and Prostriata to Caudal Auditory Cortex in the Monkey. Cereb. Cortex *20*, 1529–1538.

Fogassi, L., and Gallese, V. (2004). Action as a binding key to multisensory integration. In He Handboock of Multisensory Processes, (Cambridge (Mass.): MIT Press),.

Freire, A., Lee, K., and Symons, L.A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. Perception *29*, 159–170.

Garrido, M.I., Kilner, J.M., Stephan, K.E., and Friston, K.J. (2009). The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol. *120*, 453–463.

Goffaux, V., and Rossion, B. (2007). Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. J. Exp. Psychol. Hum. Percept. Perform. *33*, 995–1002.

Grant, and Greenberg (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In AVSP,.

Grant, K.W., and Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. *108*, 1197.

Grant, K.W., Wassenhove, V. van, and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. Speech Commun. *44*, 43–53.

Green, K.P. (1994). The influence of an inverted face on the McGurk effect. J. Acoust. Soc. Am. *95*, 3014.

Groppe, D.M., Urbach, T.P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. Psychophysiology *48*, 1711–1725.

Hietanen, J.K., Manninen, P., Sams, M., and Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? Eur. J. Cogn. Psychol. *13*, 395–407.

Hirsh, I.J., and Sherrick, C.E., Jr. (1961). Perceived order in different sense modalities. J. Exp. Psychol. *62*, 423–432.

Jaaskelainen, I.P., Ahveninen, J., Bonmassar, G., Dale, A.M., Ilmoniemi, R.J., Levanen, S., Lin, F.-H., May, P., Melcher, J., Stufflebeam, S., et al. (2004). Human posterior auditory cortex gates novel sounds to consciousness. Proc. Natl. Acad. Sci. *101*, 6809–6814.

Jack, C.E., and Thurlow, W.R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept. Mot. Skills *37*, 967–979.

John MacDonald, and McGurk, H. (1978). Visual influences on speech perception processes. Percept. Psychophys. *24*.

Johnston, A., and Nishida, S. (2001). Time perception: Brain time or event time? Curr. Biol. *11*, R427–R430.

Jordan, T.R., and Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. J. Exp. Psychol. Hum. Percept. Perform. *23*, 388–403.

Jordan, T.R., Mccotter, M.V., and Thomas, S.M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. Percept. Psychophys. *62*, 1394–1404.

Kane, N.M., Butler, S.R., and Simpson, T. (2000). Coma Outcome Prediction Using Event-Related Potentials: P₃ and Mismatch Negativity. Audiol. Neurootol. *5*, 186–191.

Kanwisher, N., McDermott, J., and Chun, M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. J. Neurosci. *17*, 4302–4311.

Karmarkar, U.R., and Buonomano, D.V. (2007). Timing in the Absence of Clocks: Encoding Time in Neural Network States. Neuron *53*, 427–438.

Kayser, C., Petkov, C.I., and Logothetis, N.K. (2008). Visual Modulation of Neurons in Auditory Cortex. Cereb. Cortex *18*, 1560–1574.

Kidd, G., Favrot, S., Desloge, J.G., Streeter, T.M., and Mason, C.R. (2013). Design and preliminary testing of a visually guided hearing aid. J. Acoust. Soc. Am. *133*, EL202.

Kislyuk, D.S., Möttönen, R., and Sams, M. (2008). Visual Processing Affects the Neural Basis of Auditory Discrimination. J. Cogn. Neurosci. *20*, 2175–2184.

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

Kösem, A., Gramfort, A., and van Wassenhove, V. (2014). Encoding of event timing in the phase of neural oscillations. NeuroImage.

Lang, A.H., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S., and Aaltonen, O. (1995). Practical issues in the clinical application of mismatch negativity. Ear Hear. *16*, 118–130.

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. NeuroImage *21*, 725–732.

Massaro, D.W., and Cohen, M.M. (1996). Perceiving speech from inverted faces. Percept. Psychophys. *58*, 1047–1065.

Massaro, D.W., Cohen, M.M., and Smeele, P.M.T. (1996). Perception of asynchronous and conflicting visual and auditory speech. J. Acoust. Soc. Am. *100*, 1777.

Maurer, D., Grand, R.L., and Mondloch, C.J. (2002). The many faces of configural processing. Trends Cogn. Sci. *6*, 255–260.

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

Meltzoff, A.N., and Moore, M.K. (1983). Newborn Infants Imitate Adult Facial Gestures. Child Dev. *54*, 702.

Milivojevic, B., Clapp, W.C., Johnson, B.W., and Corballis, M.C. (2003). Turn that frown upside down: ERP effects of thatcherization of misorientated faces. Psychophysiology *40*, 967–978.

Moore, B.C.J. (2012). An introduction to the psychology of hearing (Bingley: Emerald).

Möttönen, R., Krause, C.M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. Brain Res. Cogn. Brain Res. *13*, 417–425.

Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. Percept. Psychophys. *58*, 351–362.

Naatanen, R., Jacobsen, T., and Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. Psychophysiology *42*, 25–32.

Näätänen, R. (2003). Mismatch negativity: clinical research and possible applications. Int. J. Psychophysiol. *48*, 179–188.

Näätänen, R., and Alho, K. (1995). Mismatch negativity-a unique measure of sensory processing in audition. Int. J. Neurosci. *80*, 317–337.

Näätänen, R., Gaillard, A.W.K., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst.) *42*, 313–329.

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. Clin. Neurophysiol. *115*, 140–144.

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.

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

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A., and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport *16*, 125–128.

Pilling, M. (2009). Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. J. Speech Lang. Hear. Res. *52*, 1073–1081.

Pisoni, D. and Remez, R. (eds.) (2005). The handbook of speech perception (Malden, MA: Blackwell Pub).

Ponton, C.W., Bernstein, L.E., and Auer, E.T. (2009). Mismatch Negativity with Visual-only and Audiovisual Speech. Brain Topogr. *21*, 207–215.

Pöppel, E., Ruhnau, E., Schill, K., and Steinbüchel, N. (1990a). A Hypothesis Concerning Timing in the Brain. In Synergetics of Cognition, H. Haken, and M. Stadler, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 144–149.

Pöppel, E., Schill, K., and Steinbüchel, N. (1990b). Sensory integration within temporally neutral systems states: A hypothesis. Naturwissenschaften *77*, 89–91.

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., Alho, K., Martinkauppi, S., Ilmoniemi, R.J., and Näätänen, R. (2001). Memory Traces for Words as Revealed by the Mismatch Negativity. NeuroImage *14*, 607–616.

Rakover, S.S. (2013). Explaining the face-inversion effect: the face–scheme incompatibility (FSI) model. Psychon. Bull. Rev. *20*, 665–692.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (1981). Speech perception without traditional speech cues. Science *212*, 947–949.

Rosenblum, L.D. (2001). Reading Upside-down Lips, http://www.faculty.ucr.edu/~rosenblu/VSinvertedspeech.html.

Rosenblum, L.D., Yakel, D.A., and Green, K.P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. J. Exp. Psychol. Hum. Percept. Perform. *26*, 806–819.

Rotshtein, P., Malach, R., Hadar, U., Graif, M., and Hendler, T. (2001). Feeling or Features - Different Sensitivity to Emotion in High-Order Visual Cortex and Amygdala. Neuron *32*, 747–757.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., and Foxe, J.J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia *45*, 587–597.

Saldaña, H.M., and Rosenblum, L.D. (1993). Visual influences on auditory pluck and bow judgments. Percept. Psychophys. *54*, 406–416.

Sallinen, M., Kaartinen, J., and Lyytinen, H. (1994). Is the appearance of mismatch negativity during stage 2 sleep related to the elicitation of K-complex? Electroencephalogr. Clin. Neurophysiol. *91*, 140–148.

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

Sandmann, P., Dillier, N., Eichele, T., Meyer, M., Kegel, A., Pascual-Marqui, R.D., Marcar, V.L., Jancke, L., and Debener, S. (2012). Visual activation of auditory cortex reflects maladaptive plasticity in cochlear implant users. Brain *135*, 555–568.

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

Sekiyama, K., and Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. J. Acoust. Soc. Am. *90*, 1797.

Soto-Faraco, S., and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. J. Exp. Psychol. Hum. Percept. Perform. *35*, 580–587.

Stein, B.E., and Meredith, M.A. (1993). The merging of the senses (Cambridge, Mass.: MIT Press).

Stekelenburg, J.J., and Vroomen, J. (2007). Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events. J. Cogn. Neurosci. *19*, 1964–1973.

Stekelenburg, J.J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia.

Stevenson, R.A., Altieri, N.A., Kim, S., Pisoni, D.B., and James, T.W. (2010). Neural processing of asynchronous audiovisual speech perception. NeuroImage *49*, 3308–3318.

Sumby, W.H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Am. *26*, 212.

Talati, Z., Rhodes, G., and Jeffery, L. (2010). Now You See It, Now You Don't: Shedding Light on the Thatcher Illusion. Psychol. Sci. *21*, 219–221.

Thompson, P. (1980). Margaret Thatcher: a new illusion. Perception *9*, 483–484.

Thurlow, W.R., and Jack, C.E. (1973). Certain determinants of the " ventriloquism effect."Percept. Mot. Skills *36*.

Tiippana, K., Andersen, T.S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. Eur. J. Cogn. Psychol. *16*, 457.

Treisman, M., Faulkner, A., Naish, P.L.N., and Brogan, D. (1990). The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. Perception *19*, 705–743.

Tuomainen, J., Andersen, T.S., Tiippana, K., and Sams, M. (2005). Audio-visual speech perception is special. Cognition *96*, B13–B22.

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. Atten. Percept. Psychophys. *72*, 871–884.

Vroomen, J., and Stekelenburg, J.J. (2010). Visual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual Stimuli. J. Cogn. Neurosci. *22*, 1583–1596.

Van Wassenhove, V. (2005). Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci. *102*, 1181–1186.

Van Wassenhove, V. (2009). Minding time in an amodal representational space. Philos. Trans. R. Soc. B Biol. Sci. *364*, 1815–1830.

Van Wassenhove, V., Grant, K.W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. Neuropsychologia *45*, 598–607.

Witkin, H.A., Wapner, S., and Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. J. Exp. Psychol. *43*, 58–67.

Wittmann, M., and van Wassenhove, V. (2009). The experience of time: neural mechanisms and the interplay of emotion, cognition and embodiment. Philos. Trans. R. Soc. B Biol. Sci. *364*, 1809–1813.

Wundt, W. (1862). Die Geschwindigkeit des Gedankens. Gartenlaube *1862*, 263–265.

Zampini, M., Shore, D.I., and Spence, C. (2003). Audiovisual temporal order judgments. Exp. Brain Res. *152*, 198–210.

7 Appendix: Publications

A Facial configuration and audiovisual integration of speech: a mismatch negativity study (published in proceedings of ISAAR 2013)

Facial configuration and audiovisual integration of speech: a mismatch negativity study

Kasper Eskelund[1,2], Laura Frølich[1], Tobias S. Andersen[1,2]

[1]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark

[2]CHeSS, Oticon Centre for Hearing and Speech Sciences, Technical University of Denmark.

Author contact: kaes@dtu.dk

Abstract

Audiovisual integration plays a central role for speech perception in face-to-face communication. Visual speech may facilitate auditory detection and identification in noisy conditions. Further, a visual syllable may alter the auditory phonetic percept, as observed in the McGurk illusion.

In this study, we investigate the role of the configuration of facial features in perception of audiovisual speech. Face perception is known to be highly sensitive to specific arrangements of facial features. By nature, visual speech perception – and thus bimodal integration of audiovisual speech - relies on information from the talking face. However, the influence of face perception processes upon speech perception was only the subject of very few studies in prior research. Previous behavioral findings have shown that for some speech tokens, audiovisual speech perception is altered when the facial configuration is manipulated, even though the constituent features are unchanged. This suggests a dependency of the encoding of audiovisual speech and face perception. Here, we investigate the effect by means of electrophysiology in a mismatch negativity paradigm. Specifically, we present stimuli that support normal face perception and stimuli that disturb these processes, but only find mismatch negativity indicating audiovisual integration with the former.

Introduction

The integration of acoustic and visual speech signals is known to be beneficial for speech reception in many ways. Acoustic speech is detected at lower intensities

if accompanied by a corresponding talking face, and visual speech can facilitate speech comprehension (Grant and Seitz, 2000; Schwartz et al., 2004; Sumby and Pollack, 1954). When visual and auditory speech signals are phonetically incongruent, an illusory alteration of the perceived auditory phoneme may occur. This is known as the McGurk illusion (McGurk and MacDonald, 1976).

Perception of natural audiovisual speech evidently relies on visual speech information emanating from the talking face. Here, the signal emanating from the lip area plays a central role. Silent, visual speech is e.g. known to modulate activity in auditory cortex (Calvert and Campbell, 2003). When considering visual perception of the talker in general, face perception directed towards the configuration of facial features may also be involved. Interestingly, the impact of face perception upon speech perception has only been subject of a limited number of studies. Here we thus ask whether facial configurational information influence audiovisual perception of speech.

The importance of configurational information for face perception is demonstrated by the so-called Thatcher illusion (Thompson, 1980). This striking illusion is based on four manipulations of a face stimulus, a) a normal face with upright facial context and upright mouth (UF-UM), b) facial context kept upright, but mouth area inverted vertically (UF-IM), c) facial context inverted vertically but mouth area kept upright (IF-UM), d) facial context and mouth area both inverted vertically (IF-IM). When presenting these stimuli, Thompson (1980) observed that they were all perceived as normal faces, except for stimulus UF-IM, which was perceived as strikingly grotesque. Although the relation between directions of facial context and mouth area in stimuli UF-IM and IF-UM are identical, configurational mismatch is only perceived in the upright facial context. Thus, facial configuration information is encoded for stimuli with upright facial context (UF) only.

To investigate the influence of facial configuration on audiovisual speech perception, Rosenblum and colleagues used video stimuli based on the Thatcher illusion (2000). The four visual stimulus modifications were combined with audio, forming congruent and incongruent (McGurk-type) audiovisual speech tokens, which according to direction of facial context supported or did not

support perception of facial configuration. For the incongruent audiovisual syllable consisting of an auditory /ba/ and a visual /va/, Rosenblum reported 90% visually driven (McGurk) responses for UF-UM stimuli, while this tendency was reduced to 45% for UF-IM stimuli. Thus, audiovisual integration was reduced when perception of facial configuration was obstructed. This finding suggests a role for face perception in audiovisual speech perception.

The hypothesis of some degree of dependency of audiovisual integration on configurational properties of the talking face is intriguing. In the present study, we investigate if the behavioral findings are mirrored in a neural differential response.

Specifically, we attempt at testing the influence of configurational face information by electrophysiological means, using the mismatch negativity (MMN) paradigm developed by Näätänen (1978). MMN is a component in the auditory event-related potential (ERP), generated by presenting an oddball sequence of standard and deviant auditory stimuli at a constant inter-trial interval. Deviant stimuli usually represent 9-15% of the sequence, and a deviant trial is always followed by at least one standard trial. Deviant stimuli may deviate in any basic stimulus dimension, e.g. intensity, pitch, modulation frequency, spatial location, or, in the case of speech stimuli, phoneme. When averaging ERPs due to standard and deviant stimuli, a negative deflection of the deviant ERP is observed, reflected also in a negative differential potential in the 100-250 ms interval post-stimulus for auditory stimuli. The amplitude of the difference varies with the perceived stimulus difference in deviant stimuli: to elicit an MMN, the difference must be at least at the level of the just noticeable difference (Garrido et al., 2009). The MMN has been hypothesized to be produced by a memory process, comparing each incoming stimulus with the established trace of standard stimuli (Näätänen, 2003).

Sams and colleagues (1991) showed that the McGurk illusion can elicit MMN without any acoustic difference between standard and deviant stimuli. In such McGurk-MMN paradigms, standard trials are congruent combinations of e.g. an audiovisual /ba/. Phonetic deviance is then induced by McGurk-type audiovisual integration with incongruent audiovisual stimuli, e.g. /ba/ + /va/ (also cf. Colin,

2002; Ponton et al., 2009; Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012). Thus, only the visual phoneme is altered in deviant trials (for a different approach, cf. Kislyuk et al., 2008).

We chose phonemes /ba/ and /va/ as in Rosenblum's study (2000) and generated new stimuli for use with native Danish-speaking subjects. To keep the duration of the MMN paradigm within practical limits for EEG recordings, only two visual stimulus types were used, i.e. UF-UM and UF-IM, which yielded normal audiovisual integration and reduced audiovisual integration responses in Rosenblum's study, respectively. For the UF-UM stimuli, we would expect normal bimodal integration, resulting in a McGurk-type percept with deviant stimuli, and thus an MMN signature in the ERP. UF-IM stimuli, on the other hand, are expected not to support audiovisual integration due to their disruption of normal face perception. Thus, deviant stimuli should not induce an MMN response with UF-IM stimuli.

To ensure that audiovisual integration was present in all subjects, a behavioral task was devised after the EEG recordings. In the behavioral task, subjects were asked to identify the same stimuli as presented in the EEG experiment.

Methods

Subjects

24 engineering students and university faculty members participated, 11 female. Mean age 29 years, age range 21-59. Five subjects were excluded due to electrode failure or movement artifacts.

Stimuli

Stimulus material was generated from a video recording of syllables /ba/ and /va/. Each video was recorded at 30 fps and lasted 31 frames. Sound was recorded at 44.1 Hz sampling rate and 16 bit depth. The single auditory /ba/ was combined with four different visual stimuli: a visual /ba/ with upright face and upright lips and a visual /va/ with upright face and vertically inverted lips. This yielded congruent and incongruent UF-UM syllables, and congruent and incongruent UF-IM syllables.

Stimuli were presented on a 19" CRT screen and with Etymotic Research ER-2 ear probes at an intensity of 60 dB SPL. Subjects were seated in a comfortable armchair in a dimly lit, shielded EEG booth at a distance of 1.2 meters from the visual display.

Behavioral task

The behavioral task consisted of a random presentation of 25 trials of each of the four audiovisual stimuli. After each trial, subjects were prompted to identify what they just heard in response categories "ba", "da", "fa", or, "va".

EEG recordings

EEG was recorded on a BioSemi ActiveTwo 64-channel system with six EOG and two mastoid electrodes. Data were sampled at 512 Hz.

The four stimuli were presented in the following sequence: Two conditions were constructed, consisting of UF-UM and UF-IM audiovisual stimuli, respectively. In each of these conditions, a congruent /ba/+/ba/ combination was used as standard, while a /ba/+/va/ was used as deviant stimulus. Each grand condition was presented in two blocks, consisting of a total of 550 trials each. In each block, 15% of trials were deviant stimuli, which were distributed randomly in the sequence, with the condition that at least 2 and maximally 9 standards followed each deviant. 30 standard stimuli preceded each block as a training sequence so that the memory trace for the standard stimulus could be established. To counter movement artifacts, the stimulus sequence was paused every two minutes to allow for a 20-seconds break where subjects were instructed to relax. In total, 1100 stimuli were presented in each condition, of which 165 were deviants. Duration of each EEG recording (both conditions) was approx. 1 hour and 30 minutes, including breaks between blocks.

Results

EEG recordings

All analyses were performed with the EEGLAB toolbox developed for MATLAB (Delorme and Makeig, 2004). Continuous data from the EEG recordings was bandpass filtered between 1 and 30 Hz (Näätänen, 2003) and referenced to

averaged mastoids. Noisy electrodes were detected by a measure of kurtosis, and if any were found, their original channel data were replaced with data interpolated from surrounding electrodes. Data was segmented to epochs from 100 ms before to 600 ms after auditory onset and baselined to the 100 ms period preceding auditory onset. As a means of artifact rejection, an independent component analysis was used to reveal activity distributions and time-series attributable to non-neural sources such as eye-blinks, muscular artifacts, loose electrodes, etc. Only data from non-interpolated electrodes were entered into the independent component analysis. After decomposition, artifactual components were selected and removed upon visual inspection of spatial distributions and time-series. Residual artifacts were removed by applying a simple threshold of -100/+100 μV on all electrodes.

Pre-selection of subjects

The audiovisual MMN paradigm of the experiment relies on both perceptual (the McGurk illusion) and neurophysiological effects (the MMN response). These are well-known effects, but do not occur uniformly in a given population. The prevalence of acoustic MMN is high, but not universal (Lang et al., 1995). This is also the case for the McGurk effect, which is the auditory illusion that drives the audiovisual MMN. In the present experiment, we look for changes in audiovisual MMN when the facial configuration is altered. To be able to securely observe this, we pre-selected subjects that display an audiovisual MMN driven by the McGurk effect with a normal face (the UF-UM condition). Eight subjects were pre-selected on the criterion of an audiovisual MMN with UF-UM stimuli of > 1 μV 200-400 ms post-stimulus.

ERPs from the UF-UM condition recorded at electrode Cz are presented in Figure 1. Standard and deviant ERPs follow similar same pattern until approx. 200 ms post stimulus, where a negative deflection of the deviant ERP starts.

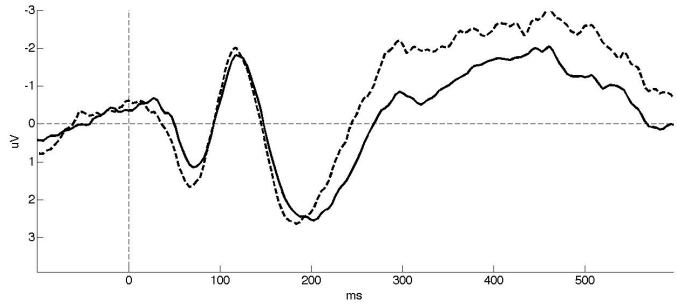


Fig.1.: Average ERPs recorded from UF-UM stimuli at electrode Cz. Auditory onset at 0 ms. Full line represents ERPs due to standard stimuli. Dashed line represents ERPs due to deviant stimuli.

Interestingly, ERPs from the UF-IM condition displayed in Figure 2 do not show the same tendency. Here, deviant ERPs show a general, but less articulate positive shift, which starts at the beginning of the auditory stimulus.

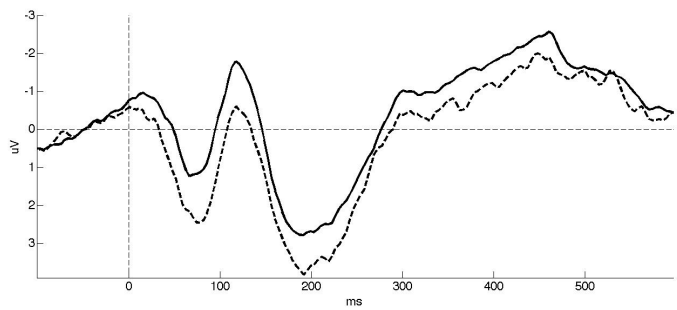


Fig.2.: Average ERPs recorded from UF-IM stimuli at electrode Cz. Auditory onset at 0 ms. Full line represents ERPs due to standard stimuli. Dashed line represents ERPs due to deviant stimuli.

In the UF-UM condition, a mismatch negativity pattern is easily seen in the difference between deviant and standard ERPs. As is evident in Figure 3, the UF-UM condition generates an MMN response beginning at approx. 200 ms and culminating with an amplitude of -1.43 μV at 280 ms. To detect reliable differences in the MMN from zero, we submitted the ERPs producing the difference wave to a repeated measures, two-tailed permutation test based on the tmax statistic (Blair and Karniski, 1993), using a family-wise alpha of 0.05. All time-points between 200 and 600 ms were included in the test. 2500 random within-subject permutations of the data were used to estimate the distribution of the null hypothesis (i.e. no difference between ERPs, or, difference wave at zero).

Based on this estimate, a critical t-score of +/- 4.31 was derived, i.e. any differences between the ERPs that exceeded this t-score were deemed statistically significant. This was the case for portions from 240 to 360 ms and 460 to 530 ms. Maximal t-score was -10.8 at 290 ms.

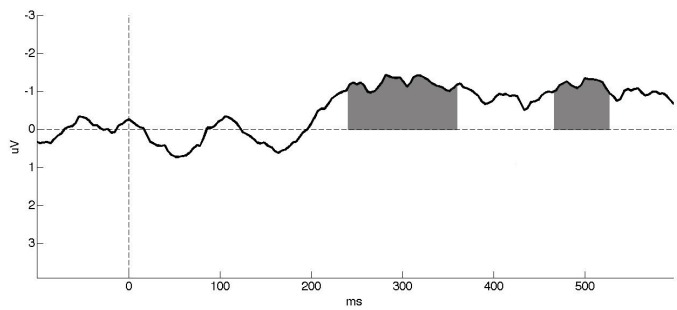


Fig.3.: Difference wave representing the difference between deviants and standards in the UF-UM condition at electrode Cz. Auditory onset at 0 ms. Shaded area marks statistically significant portions of the difference wave (exceeding the critical t-score of +/- 4.31).

As can be seen in Figure 4, the UF-IM condition generated a differential response (deviant ERP minus standard ERP) with less amplitude and reverse polarity. In this case, a permutation test identical to the one used for UF-UM data above revealed no portions of the UF-IM standard and deviant ERPs (see Figure 2) to differ significantly (critical t-score +/- 3.54, maximal t-score in the window 200 to 600 ms was +1.38 at 450 ms).

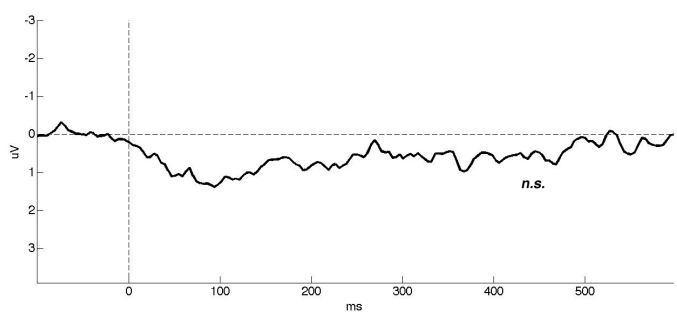


Fig.4.: Difference waves representing the difference between standards and deviants in the UF-IM condition at electrode Cz. Auditory onset at 0 ms.

Behavioral task

Observers' responses in the behavioral task were re-categorized as correct ("ba") and incorrect (all other responses). Here, we consider the mean percentage incorrect identifications as a measure of the strength of the McGurk illusion (listed in Table 1).

Table 1: Percentage incorrect identifications of the acoustic phoneme /ba/ in the behavioral task after EEG recordings. First value is mean proportion incorrect identifications, numbers in brackets represent standard error of mean.

	UF-UM	UF-IM
Congruent /ba/ +/ba/	1.5 (0.7)	4.5 (1.5)
Incongruent /ba/ + /va/	93.0 (2.1)	27.0 (6.4)

As can be seen in Table 1, incongruent UF-UM stimuli produced clear audiovisual integration responses, whereas incongruent UF-IM stimuli produced a less clear result, suggesting reduced bimodal integration. Responses were arcsine-transformed to correct for the heterogeneity of variances and analyzed using a two-way (Syllable × Mouth Direction) repeated-measures ANOVA. Arcsine-transformation did not change the outcome of any of the hypothesis tests. Factor Syllable had two levels (congruent and incongruent). Factor Mouth Direction had two levels (upright mouth and inverted mouth). *P*-values were Greenhouse-Geisser-corrected when appropriate.

The results showed that the interaction between Syllable and Mouth Direction was significant ($F(3, 21) = 120.1$, $P < 0.001$), indicating an effect of Mouth Direction on Syllable identification. We further performed repeated measures ANOVAs to compare identification performance pairwise between syllables and between mouth directions. Performance differences between congruent and incongruent syllables was significant for UF-UM ($F(1, 7) = 238.1$, $P < 0.001$) and UF-IM stimuli ($F(1, 7) = 14.5$, $P < 0.01$). The difference in congruent syllable identification between UF-UM and UF-IM stimuli were not significant ($F(1, 7) = 2.3$, $P > 0.1$). And finally, the difference in incongruent syllable identification between UF-UM and UF-IM stimuli - i.e. the difference in audiovisual integration responses between the two facial configurations - was significant ($F(1, 7) = 69.0$, $P < 0.001$).

Discussion

Results from the behavioral task match the findings of Rosenblum (2000). In the present results, the difference in audiovisual integration responses was even slightly more articulate, with 93% in the UF-UM condition vs. 27% in the UF-IM condition.

MMN results mirrored the behavioral findings. Here, the MMN response generated by visual phonetic deviance with UF-UM stimuli effectively vanished with UF-IM versions of the same stimuli. The minor, positive deflection observed was not found to reliably differ from zero, and it is hypothesized to be due to random fluctuations. Thus, the manipulated facial configuration had a significant impact on McGurk-MMN.

It is worth noting, that subjects were pre-selected for analysis on basis of their MMN in the UF-UM condition. However, the object of the present study was the change in audiovisual integration between UF-UM and UF-IM conditions and not audiovisual MMN in isolation. Because the McGurk-driven audiovisual MMN per se is not universally present in subjects, a pre-selection was necessary. The pre-selection in the present study, however, does not differ much from selection rates in other audiovisual MMN studies (cf. Colin, 2002).

Our behavioral and neurophysiological findings support the findings of Rosenblum and colleagues (2000) in suggesting that facial configuration information influences audiovisual integration in speech perception.

References

Blair, R.C., and Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. Psychophysiology *30*, 518–524.

Calvert, G.A., and Campbell, R. (2003). Reading Speech from Still and Moving Faces: The Neural Substrates of Visible Speech. J. Cogn. Neurosci. *15*, 57–70.

Colin, C. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. *113*, 495–506.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods *134*, 9–21.

Garrido, M.I., Kilner, J.M., Stephan, K.E., and Friston, K.J. (2009). The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol. *120*, 453–463.

Grant, K.W., and Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. *108*, 1197.

Kislyuk, D.S., Möttönen, R., and Sams, M. (2008). Visual Processing Affects the Neural Basis of Auditory Discrimination. J. Cogn. Neurosci. *20*, 2175–2184.

Lang, A.H., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S., and Aaltonen, O. (1995). Practical issues in the clinical application of mismatch negativity. Ear Hear. *16*, 118–130.

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

Näätänen, R. (2003). Mismatch negativity: clinical research and possible applications. Int. J. Psychophysiol. *48*, 179–188.

Näätänen, R., Gaillard, A.W.K., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst.) *42*, 313–329.

Ponton, C.W., Bernstein, L.E., and Auer, E.T. (2009). Mismatch Negativity with Visual-only and Audiovisual Speech. Brain Topogr. *21*, 207–215.

Rosenblum, L.D., Yakel, D.A., and Green, K.P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. J. Exp. Psychol. Hum. Percept. Perform. *26*, 806–819.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., and Foxe, J.J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia *45*, 587–597.

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

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

Stekelenburg, J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia.

Sumby, W.H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Am. *26*, 212.

Thompson, P. (1980). Margaret Thatcher: a new illusion. Perception *9*, 483–484.

B Face configuration affects speech perception: Evidence from a McGurk Mismatch Negativity study (published in Neuropsychologia)

Neuropsychologia 66 (2015) 48–54

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Research Report

Face configuration affects speech perception: Evidence from a McGurk mismatch negativity study

Kasper Eskelund [a,b,*,1], Ewen N. MacDonald [b,c], Tobias S. Andersen [a,b]

[a] Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[b] CHeSS, Oticon Centre of Excellence for Hearing and Speech Sciences, Technical University of Denmark, Denmark
[c] Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Denmark

ARTICLE INFO

Article history:
Received 25 May 2014
Received in revised form
23 September 2014
Accepted 14 October 2014
Available online 24 October 2014

Keywords:
Multisensory
Audiovisual
Speech perception
Face perception
Mismatch negativity
EEG

ABSTRACT

We perceive identity, expression and speech from faces. While perception of identity and expression depends crucially on the configuration of facial features it is less clear whether this holds for visual speech perception.

Facial configuration is poorly perceived for upside-down faces as demonstrated by the Thatcher illusion in which the orientation of the eyes and mouth with respect to the face is inverted (Thatcherization). This gives the face a grotesque appearance but this is only seen when the face is upright.

Thatcherization can likewise disrupt visual speech perception but only when the face is upright indicating that facial configuration can be important for visual speech perception. This effect can propagate to auditory speech perception through audiovisual integration so that Thatcherization disrupts the McGurk illusion in which visual speech perception alters perception of an incongruent acoustic phoneme. This is known as the McThatcher effect.

Here we show that the McThatcher effect is reflected in the McGurk mismatch negativity (MMN). The MMN is an event-related potential elicited by a change in auditory perception. The McGurk-MMN can be elicited by a change in auditory perception due to the McGurk illusion without any change in the acoustic stimulus.

We found that Thatcherization disrupted a strong McGurk illusion and a correspondingly strong McGurk-MMN only for upright faces. This confirms that facial configuration can be important for audiovisual speech perception. For inverted faces we found a weaker McGurk illusion but, surprisingly, no MMN. We also found no correlation between the strength of the McGurk illusion and the amplitude of the McGurk-MMN. We suggest that this may be due to a threshold effect so that a strong McGurk illusion is required to elicit the McGurk-MMN.

1. Introduction

Face perception has three important functions: face recognition, perception of facial expression and visual speech perception (cf. Bruce and Young, 2012). Face perception is special, differing from perception of other objects in a number of ways. Perhaps the most notable of these is the strong dependence of face recognition and perception of facial expression not only on features such as the mouth, eyes and nose but also, to a larger degree, on their configuration (Farah et al., 1998; Valentine, 1988).

Whether visual speech perception, as the third major function of face perception, is also dependent on configuration information is less clear. Understanding visual speech perception is particularly interesting because of the effect that automatic, subconscious speech reading has on auditory speech perception in face-to-face conversation. Evidence for this effect comes from studies showing that seeing the interlocutor's face facilitates speech perception (Sumby and Pollack, 1954) and from studies of the McGurk illusion. In the McGurk illusion (McGurk and MacDonald, 1976), an auditory phonetic percept is altered by seeing an incongruent visual phoneme. The resulting, illusory auditory percept may represent a combination of the incongruent acoustic and visual stimuli (e.g. acoustic /ga/+visual /ba/ producing an illusory percept /bga/). Or, it may produce a fusion percept, a third phoneme absent in either stimulus (e.g. acoustic /ba/ +visual /ga/ producing an illusory percept /da/). Finally, the visual phoneme may dominate the auditory percept (e.g. acoustic /ba/ +visual /ga/ producing an illusory percept /ga/). The automaticity and robustness of the McGurk effect is in stark contrast to the difficulty with which untrained observers speech read (Walden et al., 1977). This indicates that audiovisual speech perception can

* Corresponding author.
E-mail address: kaspereskelund@gmail.com (K. Eskelund).
[1] Kasper Eskelund was funded by the Oticon Foundation.

http://dx.doi.org/10.1016/j.neuropsychologia.2014.10.021

be based on visual cues that are not directly accessible to most observers. Therefore the strength of the McGurk illusion is a good measure for the accuracy of perception of visual speech—perhaps even better than direct measures of speech reading ability. This has been the reason for several studies of configuration information in speech reading to study audiovisual, in addition to, visual speech perception (e.g. Rosenblum et al., 2000).

It is clear that visual and audiovisual speech perception rely heavily on feature information mainly from the lips, tongue and teeth as seeing only the mouth area is sufficient for speech reading and for eliciting the McGurk illusion (Hietanen et al., 2001; Jordan and Thomas, 2011; Rosenblum et al., 2000). Nevertheless, somewhat surprisingly, speech can also be read from faces even when the mouth area is entirely occluded and this can influence audiovisual speech perception (Jordan and Thomas, 2011). This effect is due to the fact that movements of extraoral face areas are correlated with movements of the mouth and articulators (Jordan and Thomas, 2011). Thus, the spatial relationship of these oral and extraoral features is a candidate for configuration information that may carry visual speech information.

Hietanen et al. (2001) examined the effect of configurational information in a very direct manner. They created visual stimuli consisting only of the eyes, nose and mouth by masking the rest of the face. The location of these facial features was either in their natural position or scrambled. While some effects of feature scrambling on the strength of the McGurk illusion were found, the effects were weak and dependent on speaker identity. Still, the study supports the notion that feature configuration can influence audiovisual speech perception.

Facial configuration has been shown to be difficult to perceive in inverted faces. Hence, face recognition (Farah et al., 1998; Valentine, 1988) and perception of facial expression (Prkachin, 2003) is impaired for inverted faces. Several studies have found face inversion effects for visual and audiovisual speech perception (Jordan and Bevan, 1997; Massaro and Cohen, 1996; Rosenblum et al., 2000). Some of these studies found strong effects and others none. The overall conclusion seems to be that the face inversion effect depends greatly on the visual stimulus as it can vary across speakers even when they articulate the same speech sounds. Thomas and Jordan (2002) extended this approach by examining the effect of different levels of visual blurring. They hypothesized that since feature information depends on higher resolution than configurational information (Goffaux and Rossion, 2007) observers must rely more on configuration information when the face is blurred. Thus, blurring should lead to a greater effect of inverting the orientation of the face. Their findings confirmed this hypothesis for speech reading, as well as for congruent and incongruent audiovisual speech.

Thompson (1980) devised a striking demonstration of our inability to perceive facial configuration in inverted faces, using a photograph of Margaret Thatcher. Misconfiguration, by vertical inversion of the mouth and eye segments (so-called Thatcherization), renders the face strikingly grotesque but this is only perceived when the face is upright and not when it is inverted (cf. Fig. 1). Thus the Thatcher illusion shows that configuration information is less effective when the face is presented upside down (Bartlett and Searcy, 1993; Bruce and Young, 2012; Carbon et al., 2005). Rosenblum et al. (2000) found that misconfiguration by Thatcherization could greatly reduce the strength of the McGurk illusion but only when the face was upright. However, this effect was not driven by inversion of the mouth segment, as it did not occur when the mouth segment was presented in isolation. These findings form strong support for configuration information being important for visual and audiovisual speech perception. Rosenblum and colleagues named this striking effect of face configuration on speech perception the McThatcher effect (Rosenblum, 2001).

In Rosenblum et al. (2000), the McThatcher effect was specific to certain phonemes just as the face inversion effect has been in most studies. For audiovisual stimuli, it was only for the visual dominance illusion of hearing acoustic /ba/+visual /va/ as /va/ that the full effect occurred. This indicates that facial configuration is more important for some phonemes than others. Thomas and Jordan (2002) came to the same conclusion noticing that the difference between visual /ga/ and /da/ is mostly visible in the oral cavity. Accordingly, this contrast seems less influenced by the face inversion effect and the McThatcher effect.

To summarize previous findings, we find that, on one side, many of them suggest an effect of facial configuration on speech perception but on the other, that the effects are highly variable and sensitive to details in the stimuli. Although deterred by this variability, we found the motivation for the current study in the power and usefulness of the McThatcher effect for investigating the relation between encoding of facial configuration and perception of audiovisual speech.

In the current study, we seek to find neural correlates of the McThatcher effect. If facial configuration truly influences audiovisual speech perception then it should be reflected in auditory evoked potentials such as the mismatch negativity (MMN, Näätänen et al., 1978). In its most basic form, the MMN is elicited by a deviant stimulus (e.g. a 1200 Hz tone) after a sequence of standard stimuli (e.g. 1000 Hz tones). Average ERPs due to deviant stimuli exhibit a negative deflection in the interval 100–250 ms covering a wide area of fronto-central electrodes. An MMN response can be produced by a noticeable deviance in a wide variety of acoustic features (pitch, intensity, duration, modulation or phoneme), and the magnitude of the negative deflection varies with the magnitude of the perceived difference (Näätänen and Alho, 1995; Näätänen et al., 2004). Although the MMN reflects early pre-attentive auditory perception, it is also evoked by visually induced auditory illusions, such as ventriloquism (Stekelenburg et al., 2004) and the McGurk illusion (Colin, 2002; Ponton et al., 2009; Saint-Amour et al., 2007; Sams et al., 1991; Stekelenburg and Vroomen, 2012). In typical McGurk-MMN paradigms, congruent audiovisual syllables (e.g. auditory /ba/+visual /ba/) are presented as standards, whereas incongruent (McGurk type) stimuli are deviants (e.g. auditory /ba/+visual /va/) (Colin, 2002; Stekelenburg and Vroomen, 2012; for a different method cf. Kislyuk et al., 2008). In such McGurk-MMN paradigms, stimulus deviance is only present in the visual signal. Thus, it is an auditory differential response evoked by the incongruent visual speech signal (i.e. the McGurk illusion), which produces the audiovisual McGurk-MMN response.

In the current study, we measured the McGurk-MMN for normal and Thatcherized faces with either upright or inverted orientation. We used the congruent audiovisual syllable /ba/ as the standard stimulus and the incongruent audiovisual combination of acoustic /ba/+visual /va/ as deviant stimulus as these were the phonemes for which Rosenblum et al. (2000) found the effect to be the strongest. To ensure that the McThatcher effect occurs for these specific stimuli, we also replicate Rosenblum et al.'s behavioral paradigm. Our hypothesis is that the McGurk-MMN will mirror behavioral findings and confirm the effect as being a truly perceptual effect. As the amplitude of the MMN is known to increase with perceived stimulus difference (Garrido et al., 2009; May and Tiitinen, 2010; Näätänen et al., 1978, 2004) we expect MMN amplitudes to be correlated with levels of behavioral McGurk responses.

2. Methods

2.1. Subjects

19 subjects (11 females) with a mean age of 24 years (range 18–38) participated in the experiment. MMN is known to show high inter-individual variability (Lang et al., 1995). Therefore, as the present study targets differences in McGurk-MMN with manipulated visual speech, we defined an exclusion criterion on basis of a recording of

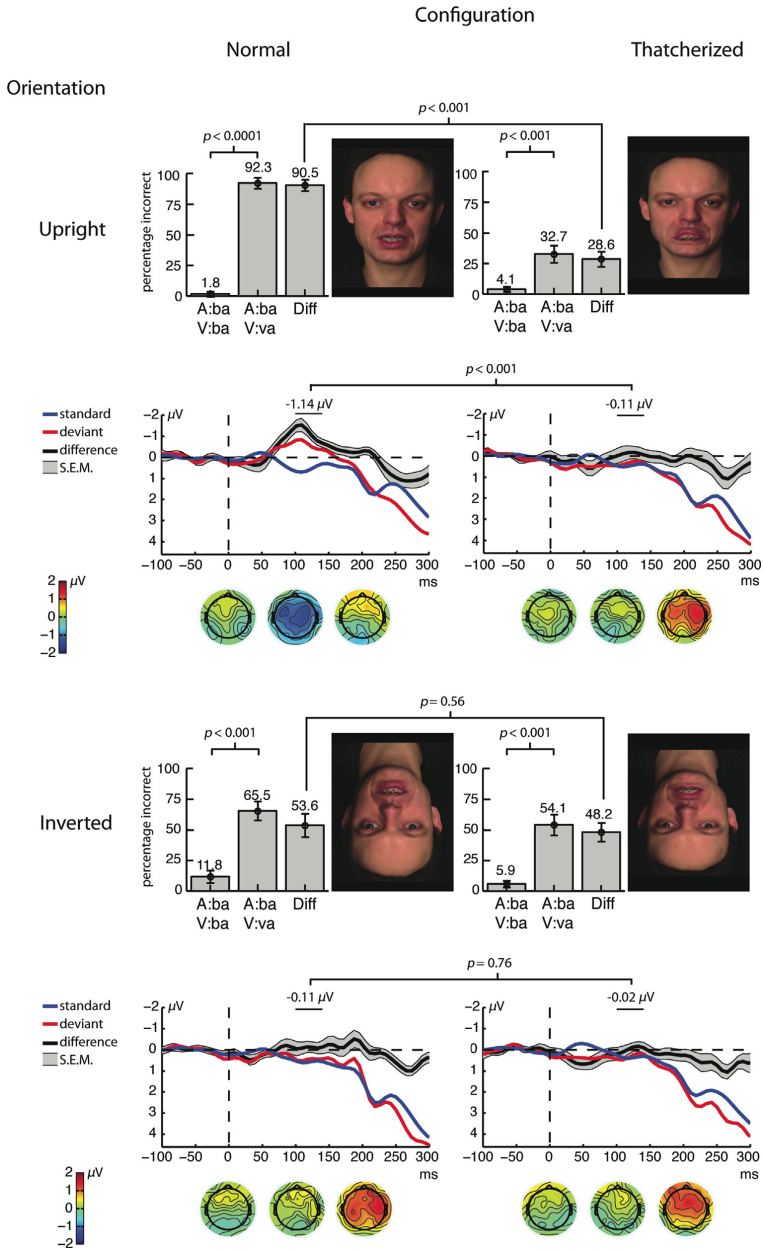


Fig. 1. Stimulus, response percentages, ERPs and scalp-maps for each of the four stimulus conditions. Stimulus: still frame from video showing maximal mouth opening. Response percentages: bar plot showing the percentage of incorrect responses as a measure of the strength of the McGurk illusion for congruent (A:ba/V:ba) and incongruent (A:ba/V:va) stimuli and the difference between them (Diff). Error-bars indicate standard error of mean. Shown are also *P*-values from statistical tests described in main text. ERPs: ERPs for standard and deviant stimuli and their difference, recorded at electrode Fz. Time zero is fixed at the voicing onset in the acoustic stimulus /ba/. Shaded area indicates standard error of the mean. The horizontal floating bar marks the interval 100–140 ms post-stimulus and the value above it indicate the average amplitude in that interval. Shown are also *P*-values from statistical tests described in main text. Scalp-maps show interpolated mean MMN amplitude at 0, 100 and 200 ms.

pure-tone MMN (Näätänen et al., 1978), as to reduce noise in our dataset by excluding subjects with a generally weak MMN response. For all subjects, acoustic MMN was recorded for 1000 Hz (standard) and 1200 Hz (deviant) tones of 100 ms duration with a SOA of 500 ms presented at 60 dB(A) SPL. The rate of deviant stimuli was 15% and a total of 1200 trials were presented. Subjects whose pure-tone MMN did not exceed −1 μV in the 100–200 ms interval were excluded. On basis of this, 8 subjects (5 female) were excluded.

2.2. Stimuli

Stimuli were generated from a video recording of syllables /ba/ and /va/. Each video was recorded at 30 fps and lasted 30 frames. Sound was recorded at 22.05 kHz sampling rate and 16-bit depth. The two visual speech tokens were edited in Adobe Premiere Elements 10 to produce the following visual manipulations of each: normal configuration, upright orientation; Thatcherized configuration, upright orientation; normal configuration, inverted orientation; Thatcherized configuration, inverted orientation (see Fig. 1). These eight visual speech tokens (/ba/ and /va/ × four visual manipulations) were combined with the acoustic /ba/ in Adobe Premiere Elements 10 to produce four congruent and four incongruent audiovisual speech stimuli.

McGurk-driven MMN responses may be confounded by purely visual responses to the visual speech signal. This is a problem in particular when studying perception of audiovisual speech compared to unimodal speech. In such studies, it is common practice to record ERPs due to the visual speech stimulus alone, and subsequently correct the audiovisual ERPs with these (cf. e.g. Colin, 2002; Möttönen et al., 2002; Sams et al., 1991). In contrast, the current experiment compares changes in the McGurk-driven MMN across four audiovisual conditions. Thus, in these audiovisual conditions, the visual response should be equal, eliminating the necessity of a correction for visual activation.

Subjects were seated in a comfortable armchair in a dimly lit, shielded EEG booth at a distance of 1.2 m from the visual display. Visual stimuli were presented on a 19 in. ViewSonic G90F CRT screen at a 60 Hz refresh rate. Sound was presented with a single Genelec 6010B monitor speaker positioned directly beneath the visual display, at an intensity of 60 dB (A) SPL measured at the head position of the subject. Stimulus presentation was controlled with Psychophysics Toolbox 3.0 (Kleiner et al., 2007).

2.3. Behavioral experiment

The behavioral task consisted of a random presentation of 20 repetitions of each of the eight audiovisual stimuli. After each trial, subjects were prompted to identify what they just heard in response categories “ba”, “da”, “fa”, or, “va”.

2.4. EEG experiment

A BioSemi ActiveTwo 64-channel EEG system referenced to the mean of two mastoid electrodes was used for recording EEG. Data were sampled at 512 Hz. EEG measurements were recorded in four conditions, each employing one of the four manipulations of the visual stimulus (cf. Fig. 1). Each condition was split into two blocks each containing an oddball sequence of 600 trials for a total of 1200 trials in each condition. The oddball sequence consisted of 85% standards, which were the audiovisual congruent syllable /ba/+/ba/, and 15% deviants, which were the incongruent audiovisual syllable /ba/+/va/. Stimuli were presented with a constant inter-trial interval of 100 ms, during which there was a crossfading between the last frame of the preceding stimulus and the first frame of the following. The sequence was randomized with the condition that at least two standards succeeded each deviant. Each block was preceded by 30 presentations of the standard stimulus. No data collected during those trials were used in the analysis. The sequence of blocks was randomized with the constraint that blocks presenting every condition were presented once, before any block was presented a second time.

2.5. EEG preprocessing

Analysis of EEG data was performed within the EEGLAB toolbox (Delorme and Makeig, 2004). First, continuous EEG data were bandpass filtered between 1 and 30 Hz (for similar filtering choices, cf. e.g. Möttönen et al., 2002; Näätänen et al., 2004; Sams et al., 1991; Stekelenburg and Vroomen, 2012), before downsampling to 128 Hz. After filtering, data were epoched in the interval −100 to 600 ms with auditory onset at 0. Epochs were baselined to the 100 ms preceding auditory onset. Electrodes dominated by unusual, non-biological waveforms were selected by a measure of kurtosis and data in these channels was interpolated from surrounding electrodes. An ICA algorithm (runica) was used (not including interpolated channels) to prepare data for rejection of independent components generated by eye artifacts, by means of the EyeCatch algorithm (Bigdely-Shamlo et al., 2013). Epochs were finally thresholded at ±100 μV to remove remaining artifacts. The proportion of epochs removed from any subject's dataset during preprocessing did not exceed 2%. ERPs were generated by averaging the preprocessed data epochs. Individual MMN waveforms were computed by subtracting average ERPs due to standard stimuli from mean ERPs due to deviant stimuli (cf. Fig. 1).

3. Results

3.1. Behavioral experiment

Responses from the behavioral task were re-categorized as correct (“ba”) and incorrect (all other responses). Among the incorrect responses to the acoustic /ba/, categories “fa” and “va”, which are visually indistinguishable, were clearly dominant, while the response category “da” only accounted for 1.2% of all incorrect auditory identifications. We use the percentage of incorrect responses as the independent variable (cf. Fig. 1).

Responses were arcsine-transformed to correct for heterogeneity of variances and subjected to a three-way, repeated-measures ANOVA with factors Orientation × Thatcherization × Congruence. Factor Orientation had two levels (Normal and Inverted), factor Configuration had two levels (Normal and Thatcherized), and factor Congruence had two levels (Congruent and Incongruent). The analysis revealed that the interaction between the three factors was significant ($F(1,10)=40.7$, $P<0.0001$).

We proceeded to perform two-way, repeated-measures ANOVAs with factors Configuration × Congruence on data from the two Orientation conditions. For upright stimuli, the interaction between Configuration and Congruence was significant ($F(1,10)=137.2$, $P<0.0001$), indicating that the conflicting direction of the mouth segment did reduce audiovisual integration. For inverted stimuli, however, the interaction between Configuration and Congruence was not significant ($F(1,10)=0.58$, $P=0.56$). This suggests that Thatcherization did not alter audiovisual integration when in the context of an inverted face.

Subsequently, we performed two-way, repeated-measures ANOVAs with factors Orientation × Congruence on data from the two Thatcherization conditions. Here, normally configured stimuli revealed a significant interaction between Orientation and Congruence ($F(1,10)=20.5$, $P<0.01$), as did Thatcherized stimuli ($F(1,10)=7.9$, $P<0.05$). This indicates that Orientation influenced audiovisual integration both when the face was Thatcherized and when it was not.

Interestingly, while inverting the orientation of the face reduced audiovisual integration for stimuli with normal configuration, it improved audiovisual integration for Thatcherized stimuli. This could indicate a role for the orientation of the mouth segment. To investigate this, we conducted a separate two-way, repeated-measures ANOVA on normal, upright and Thatcherized, inverted stimuli, which share direction of the mouth segment but within either a matching or conflicting facial orientation. The difference in the strength of the McGurk illusion was significant ($F(1,10)=31.36$, $P<0.001$) suggesting that even with a shared mouth segment direction, the two facial contexts still influenced audiovisual speech perception differently.

Individual repeated-measures ANOVAs on Congruence revealed that the difference between performance with congruent and incongruent stimuli was significant for all visual stimulus types (normal configuration, upright $F(1,10)=480.6$, $P<0.0001$; Thatcherized, upright $F(1,10)=25.0$, $P<0.001$; normal configuration, inverted $F(1,10)=35.5$, $P<0.001$; Thatcherized, inverted $F(1,10)=40.1$, $P<0.001$). Thus, we found a significant McGurk illusion in all four stimulus conditions.

3.2. Mismatch negativity experiment

We subjected the 0–200 ms interval of the difference wave to a repeated-measures, one-tailed clustered permutation test with 2500

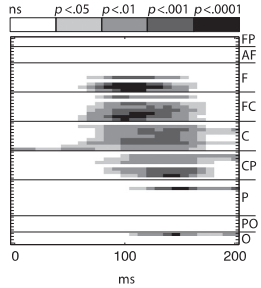


Fig. 2. Plot of significance of the McGurk-MMN for upright stimuli with normal facial configuration based on a repeated-measures, one-tailed clustered permutation test. Labels on the *y*-axis indicate frontal to anterior electrode clusters ranging from prefrontal (FP) to occipital (O). Electrodes are arranged from left (top) to right (bottom) within clusters.

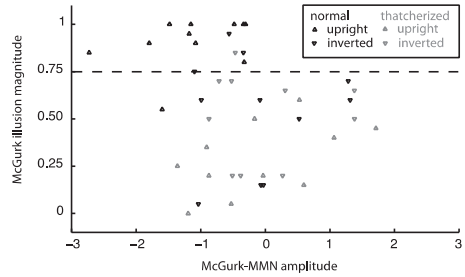


Fig. 3. Correlation of audiovisual MMN amplitude and behavioral audiovisual integration response for individual subjects. *X*-axis represents mean amplitude of the McGurk-MMN (difference wave) at electrode Fz. *Y*-axis represents the strength of the McGurk illusion measured as the percentage of incorrect responses in the incongruent condition minus the percentage of incorrect responses in the congruent condition. The dashed line indicates a threshold for the strength of the McGurk illusion. The McGurk-MMN is consistently negative above this threshold while very variable below the threshold.

permutations (for a detailed description cf. Groppe et al., 2011). We used a family-wise alpha level of 0.05 for determining statistical significance. Also, a one-tailed test was used, as any effect due to mismatch negativity would only be on the negative tail.

Only the upright face with normal configuration produced a difference wave that was significantly less than 0 at any time-point in any channel. In this condition, the difference wave recorded at a large ensemble of centro-parietal, central and fronto-central electrodes exceeded a *P*-value of 0.05 (cf. Fig. 2) during extended contiguous periods. Notably, the topographical distribution of the difference wave is centered over frontal and central electrodes. This is typical for auditory MMN (cf. e.g. Garrido et al., 2009; Näätänen and Alho, 1995), whereas differential potentials produced by visual-only MMN paradigms are centered at occipital and parietal sites (cf. e.g. Czigler, 2007; Stefanics et al., 2011).

The remaining conditions did not produce any MMN as their difference waves were not significantly below 0 in the target interval (0 to 200 ms). This suggests that the audiovisual MMN response is highly sensitive to both orientation inversion and Thatcherization.

We proceeded to compare difference waves from the four stimulus conditions. For this comparison, we again chose electrode Fz, which is a commonly used site for location of MMN in both auditory and audiovisual paradigms (Colin, 2002; Garrido et al., 2009; Näätänen and Alho, 1995; Stekelenburg and Vroomen, 2012). We extracted the mean amplitude in the 100–140 ms interval as a measure of mismatch negativity (Fig. 1). These values were subjected to a Wilcoxon signed-rank test, to test for differences across the difference waves in the four stimulus conditions. Here, the normal, upright condition proved significantly different from the other three conditions ($P < 0.02$ for each comparison). Comparisons not including the normal, upright condition did not yield any significant difference ($P > 0.70$ for each comparison). This again suggests a high sensitivity of the McGurk-MMN to Thatcherization and inversion.

In order to investigate the apparent discrepancy between the MMN and behavioral data we calculated the correlation between the mean MMN amplitude in the 100–140 ms interval and the difference in incorrect identifications between incongruent and congruent stimuli conditions across all subjects and conditions (cf. Fig. 3). The correlation between MMN amplitude and this behavioral McGurk measure was not significant ($P > 0.2$). As the estimated correlation may depend on the behavioral measure being constrained to values between zero and one, we also calculated the correlation between the MMN amplitude and the *Z*-score ($P > 0.2$; only for values greater than zero and less than one); as well as between the MMN amplitude and the arcsine transformed behavioral measure ($P > 0.1$) but this only confirmed the lack of correlation.

When looking at correlations between behavioral responses and McGurk-MMN, the results are inconclusive. The lack of correlation may seem surprising as we did find that both measures were higher in the normal, upright condition where we expected integration to be maximal. One explanation, which can never be excluded for a negative finding, is lack of statistical power due to an insufficient amount data. Another explanation is that this could indicate a non-linear effect where the MMN response only occurs when the McGurk illusion is very strong. This is unlike findings from several auditory MMN paradigms showing that MMN amplitude correlates well with perceived difference (May and Tiitinen, 2010; Näätänen, 2003; Näätänen et al., 1978). This has, to our knowledge, not been investigated for the McGurk-MMN. To investigate if the relationship between the perceived difference and the McGurk-MMN amplitude could be nonlinear we calculated the minimal behavioral response (difference in percentage incorrect between congruent and incongruent conditions) that elicited MMN negativity for all subjects in all conditions. We found that for behavioral measures of 75 percent points and above there was a consistent MMN. For behavioral measures below 75 percent points we found that the MMN was much more variable. This indicates that the McGurk illusion needs to be very strong to elicit the MMN consistently.

4. Discussion

4.1. Behavioral experiment

The present behavioral findings replicate Rosenblum et al.'s (2000) primary finding: The McThatcher effect. Thatcherization greatly reduces the influence of vision upon the auditory speech percept for an upright, but not for an inverted face. As a secondary finding we found a stronger face inversion effect than Rosenblum et al. (2000) in that inverting the normal face reduced the McGurk illusion more than in their study. While others have found smaller effects (Bertelson et al., 1994; Jordan and Bevan, 1997; Thomas and Jordan, 2002), our results are similar to those of Massaro and Cohen (1996). Given that the magnitude of the inversion effect has varied substantially across previous reports, this is not surprising. Overall, the McThatcher effect replicated in the present study supports the hypothesis that audiovisual speech perception is based not only on facial features but also on facial configuration (Rosenblum et al., 2000).

4.2. Mismatch negativity experiment

The non-Thatcherized upright face produced a strong McGurk-MMN response. The amplitude, latency and scalp distribution of this MGurk-MMN response is comparable with those reported in previous studies (Colin, 2002; Ponton et al., 2009; Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012). This signifies that the McGurk illusion we found in the behavioral experiment influenced activity in auditory cortex confirming that the effect is truly perceptual.

The two elements of the McThatcher effect were also reflected in the McGurk-MMN. First, we found no McGurk-MMN for an upright Thatcherized face reflecting that Thatcherization disrupts the McGurk illusion for upright faces. Second, we found no effect of Thatcherization on the McGurk-MMN for inverted stimuli. This mirrors the lack of difference found in the two matching behavioral conditions. However, as none of the inverted faces produced an MMN irrespective of facial configuration, only limited conclusions can be drawn from this regarding the effect of facial configuration for inverted faces. Thus, surprisingly, we found a much stronger face inversion effect in the McGurk-MMN than in the behavioral data.

4.3. General discussion

Our primary question was directed towards the role of facial configuration information in perception of visual and audiovisual speech. Our findings answer this and related questions in multiple ways.

First, our main finding is that the McThatcher effect is reflected in both behavioral and MMN responses. We found a strong McGurk illusion and a corresponding MMN for a normal upright face. The McGurk illusion was strongly reduced and the MMN eliminated when the face was Thatcherized. This confirms that speech perception is affected by facial configuration when the facial orientation is upright. For inverted faces, we found no effect of Thatcherization on the McGurk illusion and the corresponding MMN. This is in agreement with the notion that facial configuration has little influence on visual and audiovisual speech perception for inverted faces.

As a secondary finding we found a discrepancy between our behavioral findings and the MMN for the face inversion effect. Although face inversion decreased the strength of the McGurk illusion, it did not eliminate it. While we found a moderate McGurk illusion for inverted faces, these stimuli did not elicit an MMN response. Unfortunately, this means that our MMN data tells us little about the effect of Thatcherization for inverted faces.

Investigating this discrepancy further we found no correlation between the magnitude of the McGurk illusion and the amplitude of the McGurk-MMN. We suggest three possible explanations of this. First, the statistical power of our MMN data is limited, and it may be insufficient for finding a McGurk-MMN of smaller effect size. Another possibility is that the McGurk illusion for inverted faces is not truly perceptual but based on changes in behavior at another stage of perceptual processing, e.g. in response selection. As we find the McGurk illusion perceptually convincing even for inverted faces we do not believe that this is the correct interpretation but admit that this remains to be tested formally. Here, McGurk responses to the specific incongruent syllable combination (acoustic /ba/+visual /va/) is dominated by the categories "va" and "fa". These responses could in principle be due to both audiovisual integration and to response bias towards the visual stimulus. In the latter case, no McGurk-driven MMN would be produced. Repeating the experiment using a discrimination task (Rosenblum et al., 2000) or sensitivity measures from signal detection theory (Kislyuk et al., 2008) could help elucidate this.

Finally, the McGurk-driven MMN may differ from auditory MMN in having a non-linear relation to the perceived difference. We find this a likely explanation as the McGurk-MMN was consistent only when the behavioral data indicated a strong McGurk illusion and highly variable for weaker McGurk illusions. Whereas the relation between MMN amplitude and stimulus deviation is well-described for acoustic stimuli, we are not aware of any study targeting the relation between McGurk illusion strength and amplitude of the McGurk-MMN response. However, studies using McGurk-MMN, which also report the level of behavioral McGurk response, report near 100% McGurk illusion with incongruent stimuli (Stekelenburg and Vroomen, 2012), or, "a strong McGurk illusion" (Saint-Amour et al., 2007). Kislyuk et al. (2008) exclude subjects with "a weak McGurk effect". From these reports of behavioral McGurk illusion strength it may be that a strong behavioral McGurk response is a prerequisite for evoking an audiovisual MMN response. If this is the case, McGurk-driven MMN differs from auditory MMN (Garrido et al., 2009; May and Tiitinen, 2010; Näätänen et al., 1978, 2004) in not being a graded response, proportional to the degree of stimulus deviance, but only being evoked by a strong audiovisual integration response. This warrants caution in basing conclusions about audiovisual speech perception on the McGurk-MMN.

References

Bartlett, J.C., Searcy, J., 1993. Inversion and configuration of faces. Cogn. Psychol. 25, 281–316.

Bertelson, P., Vroomen, J., Wiegeraad, G., de Gelder, B. Exploring the relation between McGurk interference and ventriloquism, Proceedings of ICSLP 1994, Yokohama, pp. 559–562.

Bigdely-Shamlo, Kreutz-Delgado, K., Kothe, C, Makeig, S., 2013. EyeCatch: data-mining over half a million EEG independent components to construct a fully-automated eye-component detector. In: Proceedings of the IEEE Engineering in Biology and Medicine Conference, Osaka.

Bruce, V., Young, A.W., 2012. Face Perception. Psychology Press, London; New York.

Carbon, C.-C., Schweinberger, S.R., Kaufmann, J.M., Leder, H., 2005. The Thatcher illusion seen by the brain: an event-related brain potentials study. Cogn. Brain Res. 24, 544–555.

Colin, C., 2002. Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. 113, 495–506.

Czigler, I., 2007. Visual mismatch negativity: violation of nonattended environmental regularities. J. Psychophysiol. 21, 224–230.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134, 9–21.

Farah, M.J., Wilson, K.D., Drain, M., Tanaka, J.N., 1998. What is "special" about face perception? Psychol. Rev. 105, 482–498.

Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009. The mismatch negativity: a review of underlying mechanisms. Clin. Neurophysiol. 120, 453–463.

Goffaux, V., Rossion, B., 2007. Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. J. Exp. Psychol. Hum. Percept. Perform. 33, 995–1002.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48, 1711–1725.

Hietanen, J.K., Manninen, P., Sams, M., Surakka, V., 2001. Does audiovisual speech perception use information about facial configuration? Eur. J. Cogn. Psychol. 13, 395–407.

Jordan, T.R., Bevan, K., 1997. Seeing and hearing rotated faces: influences of facial orientation on visual and audiovisual speech recognition. J. Exp. Psychol. Hum. Percept. Perform. 23, 388–403.

Jordan, T.R., Thomas, S.M., 2011. When half a face is as good as a whole: effects of simple substantial occlusion on visual and audiovisual speech perception. Atten. Percept. Psychophys. 73, 2270–2285.

Kislyuk, D.S., Möttönen, R., Sams, M., 2008. Visual processing affects the neural basis of auditory discrimination. J. Cogn. Neurosci. 20, 2175–2184.

Kleiner M., Brainard D.H., Pelli D.G., What's New in Psychtoolbox-3?, *Perception*, 36, ECVP 2007 abstract supplement.

Lang, A.H., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S., Aaltonen, O., 1995. Practical issues in the clinical application of mismatch negativity. Ear Hear. 16, 118–130.

Massaro, D.W., Cohen, M.M., 1996. Perceiving speech from inverted faces. Percept. Psychophys. 58, 1047–1065.

May, P.J.C., Tiitinen, H., 2010. Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. Psychophysiology 47, 66–122.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.
Möttönen, R., Krause, C.M., Tiippana, K., Sams, M., 2002. Processing of changes in visual speech in the human auditory cortex. Brain Res. Cogn. Brain Res. 13, 417–425.
Näätänen, R., 2003. Mismatch negativity: clinical research and possible applications. Int. J. Psychophysiol. 48, 179–188.
Näätänen, R., Alho, K., 1995. Mismatch negativity—a unique measure of sensory processing in audition. Int. J. Neurosci. 80, 317–337.
Näätänen, R., Gaillard, A.W.K., Mäntysalo, S., 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. 42, 313–329.
Näätänen, R., Pakarinen, S., Rinne, T., Takegata, R., 2004. The mismatch negativity (MMN): towards the optimal paradigm. Clin. Neurophysiol. 115, 140–144.
Ponton, C.W., Bernstein, L.E., Auer, E.T., 2009. Mismatch negativity with visual-only and audiovisual speech. Brain Topogr. 21, 207–215.
Prkachin, G.C., 2003. The effects of orientation on detection and identification of facial expressions of emotion. Br. J. Psychol. 94, 45–62.
Rosenblum, L.D., 2001. Reading upside-down lips, ⟨http://www.faculty.ucr.edu/~rosenblu/VSinvertedspeech.html⟩.
Rosenblum, L.D., Yakel, D.A., Green, K.P., 2000. Face and mouth inversion effects on visual and audiovisual speech perception. J. Exp. Psychol. Hum. Percept. Perform. 26, 806–819.
Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., Foxe, J.J., 2007. Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia 45, 587–597.
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. Neurosci. Lett. 127, 141–145.
Stefanics, G., Kimura, M., Czigler, I., 2011. Visual mismatch negativity reveals automatic detection of sequential regularity violation. Front. Hum. Neurosci. 5.
Stekelenburg, J.J., Vroomen, J., 2012. Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia, 1425–1431.
Stekelenburg, J.J., Vroomen, J., de Gelder, B., 2004. Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. Neurosci. Lett. 357, 163–166.
Sumby, W., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust Soc Am. 26, 212–215.
Thomas, S.M., Jordan, T.R., 2002. Determining the influence of Gaussian blurring on inversion effects with talking faces. Percept. Psychophys. 64, 932–944.
Thompson, P., 1980. Margaret Thatcher: a new illusion. Perception 9, 483–484.
Valentine, T., 1988. Upside-down faces: a review of the effect of inversion upon face recognition. Br. J. Psychol. 79, 471–491.
Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J., 1977. Effects of training on the visual recognition of consonants. J. Speech Hear. Res. 20, 130–145.

C Electrophysiological correlates of the temporal window of audiovisual integration in speech perception (unsubmitted manuscript)

Electrophysiological correlates of the temporal window of audiovisual integration in speech perception

Kasper Eskelund[1,2], Jeroen J. Stekelenburg[3], Tobias S. Andersen[1,2]

[1]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark

[2]CHeSS, Oticon Centre for Hearing and Speech Sciences, Technical University of Denmark.

[3]Department of Cognitive Neuropsychology, University of Tilburg, Netherlands

Abstract

Integration of audiovisual speech is known to be tolerant to high levels of intersensory asynchrony (Conrey and Pisoni, 2006; Munhall et al., 1996; van Wassenhove et al., 2007). A series of behavioral studies have targeted this temporal window of audiovisual integration in speech perception, by observing the strength of integration phenomena, such as the McGurk illusion or by using direct measures of perceived synchrony such as simultaneity judgment or temporal order judgment, while varying audiovisual asynchrony.

Here, we ask if behavioral estimates of this temporal window of audiovisual integration are mirrored in neural responses. Specifically, we first measure the level of behavioral McGurk-responses to incongruent audiovisual syllables at a wide range of audiovisual asynchronies. On basis of individual data, we then estimate audiovisual integration response curves as a function of asynchrony level. Using these individual functions to predict which asynchrony levels will yield specific levels of audiovisual integration responses (maximal, 70%, and 20% integration responses), we measure electrophysiological differential responses in a mismatch negativity paradigm to McGurk stimuli at these lags. Our hypothesis is that neurophysiological findings mirror behavioral McGurk response levels. However, our mismatch negativity findings only partly reflect the behavioral response.

Introduction

Speech perception integrates hearing and vision across temporal shifts

Speech perception integrates signals from ear and eye. This is evident e.g. in the detection advantage associated with audiovisual speech (Sumby and Pollack, 1954), and in audiovisual illusion effects, such as ventriloquism (Bertelson et al., 2000) and the McGurk illusion (McGurk and MacDonald, 1976). Interestingly, these influences of vision upon hearing do not require strict synchrony between acoustic and visual signals. Rather, vision may interact with hearing within a remarkably wide window of audiovisual asynchronies (Conrey and Pisoni, 2006; Munhall et al., 1996; Navarra et al., 2010; van Wassenhove et al., 2007).

The window is wider for speech than non-speech

Audiovisual asynchrony in abstract, non-speech stimuli, such as flashes and beeps, are detected at intersensory asynchronies of just 25-50 ms (cf. e.g. Zampini et al., 2003). In comparison, audiovisual speech stimuli are perceived as simultaneous with considerably larger intersensory lags. For such stimuli, audiovisual integration is reported to be effective with acoustic lags of 200 ms or more (Massaro and Cohen, 2000; Munhall et al., 1996; van Wassenhove et al., 2007). Tolerance to asynchrony is often reported in terms of a temporal window, encompassing the interval of intersensory shifts from maximal tolerable lag in the visual stimulus to the corresponding acoustic lag. Most studies report that larger delays are tolerated in the acoustic stimulus than in the visual. Thus, the temporal window is asymmetrically arranged around the point of synchronous presentation. Estimates of this window, however, are dependent on the aspect of audiovisual perception addressed, on the experimental paradigm, and on the stimulus material. Thus, indicators of simultaneous perception of asynchronous audiovisual speech may vary considerably between methods.

Simultaneity judgment

One straightforward measure of asynchrony tolerance is simply to administer a simultaneity judgment (SJ) task while presenting visual and acoustic speech with different temporal onset. In typical SJ paradigms, subjects are asked to indicate whether an audiovisual speech stimulus was presented in synchrony, or if any

modality was lagging, disregarding all other aspects of the stimulus (Conrey and Pisoni, 2006; van Wassenhove et al., 2007). In a study comparing speech and nonspeech audiovisual stimuli, Dixon and Spitz (1980) estimated that asynchronuous speech stimuli within 131 ms visual lag and 258 ms acoustic lag were perceived as simultaneous. Grant and colleagues (2004) found a narrower window, within 35 ms visual lag and 160 ms acoustic lag, but their study differed in using full sentences as stimulus material. Conrey and Pisoni (2006) found that stimuli with visual lags up to 144 ms and acoustic lags up to 254 ms were deemed simultaneous. On basis of a similar experiment, van Wassenhove and colleauges (2007) suggested a window between visual lag of 74-80 ms and acoustic lag of 125-131 ms. However, Maier and his colleagues (2011) found a narrower integration window of approximately 164 ms based on simultaneity judgment, while still observing asymmetry around synchronous presentation. In conclusion, most SJ studies report a window of perceived simultaneity of 200 ms or wider, arranged asymmetrically around the point of stimulus synchrony.

Sensitivity of the McGurk illusion to asynchrony

Another way of targeting simultaneous perception of asynchronous speech is to observe the asynchrony levels within which binding of acoustic and visual speech is effective. The interaction of hearing and vision in speech perception is observed in various audiovisual illusions, such as ventriloquism (Bertelson et al., 2000) or the McGurk illusion (McGurk and MacDonald, 1976). In the latter, a co-occurring, incongruent visual phoneme (e.g. /gi/) alters perception of an acoustic phoneme (e.g. /bi/), resulting in hearing a third phoneme (either /di/ or /bgi/) or in the auditory percept being dominated by the visual stimulus (hearing a /gi/). Basic paradigms (e.g. Munhall et al., 1996; van Wassenhove et al., 2007) employ an identification task which records McGurk responses, while the incongruent speech tokens are presented at different intermodal lags. On basis of the strength of the McGurk illusion over the tange of lags, the temporal window within which vision influences the auditory percept is estimated. Such findings represent the temporal correlation sufficient for supporting binding of phonetic information in acoustic and visual speech signals.

In this way, Munhall and colleagues (1996) investigated sensitivity of the McGurk illusion to asynchrony. In their study, audiovisual incongruent bisyllabes (e.g. acoustic /aba/ + visual /aga/) were presented at asynchrony levels ranging from 360 ms visual lag to 360 ms acoustic lag, in 60 ms steps. Here, audiovisual integration as measured in incorrect identifications of the acoustic consonant followed a v-shaped function, with maximal McGurk response at 60 ms acoustic lag. The width of the integration window was estimated to range from 60 ms visual lag to 240 ms acoustic lag.

In a related study, van Wassenhove and coworkers (2007) targeted the relation between the temporal window of the SJ task and asynchrony sensitivity in the McGurk illusion. For both tasks, congruent and incongruent monosyllables were present at lags ranging from 467 visual lag to 467 ms acoustic lag in steps of 33 ms. SJ results suggested a temporal window if 205 ms centered at 23-29 ms acoustic lag of for congruent speech. With McGurk stimuli, integration was effective in a window of 208 ms, centered around 70 ms acoustic lag. Interestingly, the SJ task also included the incongruent syllables, used for the McGurk identification task. For these stimuli, the window of simultaneous perception was narrower at 159-161 ms, centered at 37-43 ms acoustic lag. This signifies that the coherence of acoustic and visual speech supports temporal integration, whereas incongruence reduces tolerance.

Tolerance of the audiovisual detection advantage to asynchrony

The auditory detection advantage associated with a co-occurring visual speech signal (Sumby and Pollack, 1954) is another effect of audiovisual binding. Grant and colleagues (2004) investigated the sensitivity of this effect to audiovisual asynchrony, using sentence-length stimuli. Interestingly, the audiovisual detection advantage exhibited a similar window of asynchronies, being effective within visual lags of 45 ms to acoustic lags of 200 ms.

Different paradigms measure different capacities

The studies reviewed above all target the tolerance of audiovisual perception to asynchrony. However, the capacities measured by SJ tasks, identification of incongruent syllables and audiovisual detection tasks are quite different. SJ tasks directly target the temporal properties of the audiovisual stimulus and may thus

be more sensitive to asynchrony than identification or detection tasks, which do not direct attention towards temporal properties of the stimulus. Processes involved in binding of acoustic and visual speech have been thought either to be aspects of a single, unified binding process (Vatakis and Spence, 2007; Vatakis et al., 2008), or elements of a multi-stage process (Eskelund et al., 2010; Schwartz et al., 2004). Finally, based on the observation that phonetic integration, as observed in the McGurk illusion, can be fully separated from simultaneity perception, Soto-Faraco and Alsius (Soto-Faraco and Alsius, 2009) have suggested that at least these two processes be independent. In their findings, the McGurk illusion could be induced with asynchronies wider than in any previous report. Here, visual lags of 320 ms and acoustic lags of 480 ms still were dominated by illusory phoneme identifications.

Differences in stimulus material produce different temporal window estimates

Another important variable is the dynamics of the audiovisual stimulus material. Here, mono-syllables, bi-syllables and full-sentence stimuli may provide different resolutions in the audiovisual temporal dynamics, producing different results. Furthermore, audiovisual phonemes differ in their audiovisual dynamics, and onset asynchrony may be easier to detect in specific phonemes. Lastly, articulation characteristics of the individual speaker may produce different audiovisual dynamics with phonetically identical stimuli (see e.g. Hietanen et al., 2001 for the influence of individual articulation characteristics). In identification paradigms using incongruent speech, the difference in dynamics between acoustic and visual signals also may alter audiovisual binding. Thus, the binding observed in the McGurk illusion may very well be different from binding of natural, congruent speech stimuli, and the asynchrony tolerance with incongruent speech as measured in SJ is lower (van Wassenhove et al., 2007).

Current experiment

The reviewed studies broadly agree on a rather wide window of asynchronies, width estimates ranging from 161 to 245 ms. Furthermore, they all report that this window is shifted towards acoustic lags, centered at lags in the range between 23 ms to 70 ms. However, a difference in window estimate of more than 80 ms demands further explanation. The extreme findings of Soto-Faraco and

Alsius (2009) highlight the variability in two regards. Firstly, the difference in audiovisual integration as measured in McGurk illusion strength and SJ is striking. Secondly, their findings display the variability between behavioral estimates of the sensitivity of the McGurk illusion to audiovisual asynchrony.

To target the temporal window of audiovisual integration in speech from a different angle, we here look for a neural correlate to the behavioral window estimates in prior research in the hope of producing a supplementary estimate. Starting with a behavioral identification task of incongruent phonemes, we first estimate individual audiovisual integration functions across different asynchrony levels. Based on this function, we choose asynchrony levels on these individual functions, representing the maximal, 70% and 20% levels of audiovisual integration. At these asynchronies, we then record event-related potentials (ERPs) in a mismatch negativity (MMN) paradigm. The MMN stimulus sequence presents congruent and incongruent audiovisual speech stimuli. On basis of a differential auditory neural response to incongruent phonemes (evoking the McGurk illusion), we aim at producing a neural estimate of the temporal window of audiovisual integration, which then can be compared to behaviorally estimated integration windows in the same subjects.

Mismatch negativity

MMN is a component in the auditory ERP, which first was produced in passive, auditory oddball paradigms (Näätänen et al., 1978). In its most basic form, MMN is evoked by presenting an oddball sequence consisting of standard stimuli (e.g. a 1000 Hz sine tone) and deviants (e.g. a 1200 Hz sine tone) with a constant inter-stimulus interval (Näätänen et al., 1978). Deviant trials usually represent between 9 and 15% (Garrido et al., 2009; Lang et al., 1995), and the stimulus sequence is randomized, with the condition that no deviant trial follows upon another deviant trial. In standard acoustic MMN paradigms, average ERPs evoked by deviant stimuli exhibit a negative deflection within a latency interval of 100-250 ms after auditory onset, over a wide area of fronto-central electrodes. An MMN response can be produced by noticeable differences in any basic acoustic stimulus feature (pitch, modulation, intensity, duration, onset lag, phoneme, etc.), and the depth of the negative deflection varies with the depth of

the perceived stimulus difference. (Näätänen and Alho, 1995; Näätänen et al., 2004).

Although originally understood as an effect of auditory sensory memory, MMN is also evoked by visually induced auditory illusions, such as ventriloquism (Stekelenburg et al., 2004) and the McGurk illusion (Colin, 2002; Ponton et al., 2009; Saint-Amour et al., 2007; Sams et al., 1991; Stekelenburg and Vroomen, 2012). In typical McGurk-MMN paradigms, congruent audiovisual syllables (e.g. auditory /ba/ + visual /ba/) are presented as standards, whereas incongruent (McGurk type) stimuli are deviants (e.g. auditory /ba/ + visual /va/) (Colin, 2002; Stekelenburg and Vroomen, 2012). In such McGurk-MMN paradigms, stimulus deviance is only present in the visual signal (for a different approach cf. Kislyuk et al., 2008). It is thus an auditory illusion provoked by the incongruent visual speech signal (i.e. the McGurk illusion), which produces the MMN response.

A visual oddball stimulus may however produce a differential potential in itself, within the latency range of the MMN response (for review of visual MMN see Czigler, 2007; Pazo-Alvarez et al., 2003). A common method for correcting this, is to record ERPs due to the visual standard and deviant stimuli alone in a similar oddball sequence. The resulting visual-only (VO) standard and deviant ERPs are subsequently subtracted from audiovisual standard and deviant ERPs (Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012). This produces an audiovisual MMN response, corrected for any purely visual response (AV-VO). One caveat of this approach is that the relation between ERPs due to audiovisual and visual-only speech is assumed to be additive, where it is actually unknown. For instance, visual speech is known to stimulate activity in auditory cortex (Möttönen et al., 2002). This introduces the risk, that subtraction of visually-driven potentials in auditory cortex effectively blurs the representation of the audiovisual response. However, we recognize the need to control visual evoked potentials and thus here employ the method for subtracting VO ERPs.

In the following, we first estimate the temporal window of audiovisual integration behaviorally by means of an identification task with incongruent speech stimuli. Based on the level of McGurk-responses to incongruent stimuli,

we estimate individual audiovisual integration functions. On basis of these, we select three asynchronies, representing maximal, 70% and 20% behavioral McGurk-response levels at which we record audiovisual MMN. In this manner, we directly target the correlation between the behaviorally estimated integration window and a neural differential response to McGurk-type stimuli at points in this window.

Methods

Subjects

24 native Dutch-speaking subjects (21 female) with a mean age of 20 years (SD=2.38, range 18-27) participated in the experiment.

Exclusion criterion 1: An individual function for dependency of audiovisual integration upon asynchrony level must be attainable

The goal of the study is to measure the neural response to asynchronous audiovisual speech stimuli, along individual behaviorally estimated temporal windows of audiovisual integration. Thus, to be able to proceed to the MMN measurements, an individual audiovisual integration function must be estimated for each subject. Participants whose behavioral audiovisual integration response over the range of asynchronies did not lend itself to estimation of a double sigmoidal function (see below) are thus excluded from the MMN measurement. 12 subjects were excluded on this criterion.

Exclusion criterion 2: Screening for individual pure-tone MMN

MMN is known to show high inter-individual variability (Lang et al., 1995). The current study targets McGurk-MMN, which is known to be less consistent, reaching lower amplitudes than auditory MMN (Colin, 2002; Sams et al., 1991; Stekelenburg and Vroomen, 2012). We thus defined an inclusion criterion on basis of pure-tone MMN, as not to introduce noise in our dataset from including subjects with a generally poor MMN response. For all subjects, pure-tone MMN was recorded due to 1000 Hz (standard) and 1200 Hz (deviant) tones of 100 ms duration with a SOA of 650 ms presented at a peak intensity of 65 dB(A) SPL. The rate of deviant stimuli was 15% and a total of 1200 trials were presented. The inclusion criterion was defined as a pure-tone MMN larger than a -1 μV in

the 100-200 ms interval post stimulus (Colin, 2002). All subjects met this criterion.

Stimuli

Stimulus material was generated from a video recording of Dutch nonsense bisyllables /tabi/ and /tagi/. The two visual speech tokens were edited in Adobe Premiere Elements. Each video was recorded at 25 fps and lasted 40 frames. Before the onset of the first frame, four frames (duration 160 ms) presented a fade from black onto the first frame of the stimulus.

Sound was recorded at 44.1 kHz sampling rate. The acoustic /tabi/ was presented either together with the visual /tabi/ (congruent stimulus) or with the visual /tagi/ (incongruent stimulus).

The acoustic phoneme of interest is the second syllable /bi/, which would be prone to influence by the visual syllable /gi/ in incongruent audiovisual combinations. The onset of the /bi/ syllable follows 280 ms after onset of the auditory stimulus.

Subjects were seated in a comfortable chair in a dimly lit, shielded EEG booth at a distance of 1.2 meters from the visual display. Visual stimuli were presented on a 19" CRT screen at a 25 Hz refresh rate. The visual stimulus extended to 14 degrees horizontal and 12 degrees vertical of the visual field. Sound was presented with two monitor speakers located on each side of the visual display, at an intensity of 65 dB(A) SPL at the head position of the subject. Stimulus presentation was controlled with E-Prime 1.1 software.

Congruent and incongruent syllables were presented at varied audiovisual asynchronies, ranging from -600 ms (auditory lead) to +600 (auditory lag) in 40 ms steps, for a total of 30 different audiovisual asynchronies.

Catch trial stimuli

To ensure that gaze was always directed towards the visual stimulus in the otherwise passive MMN experiment, visual catch trials were used. In these trials, a white dot was overlaid the nose area of the face stimulus in a 120 ms interval, covering the onset of the target visual syllable /bi/. To avoid contamination from

irrelevant potentials evoked by the catch trial task, data from these trials were not included in the ERP analysis.

Procedure

Behavioral experiment

The behavioral experiment consisted an identification task, in which each of the 30 asynchrony level was presented 12 times in randomized order, totaling 360 trials. Subjects had to identify the acoustic syllable /bi/ as either /bi/, /gi/, or, /di/.

Selection of audiovisual asynchronies for the MMN experiment

After the experiment, identification data from each subject was used to model an audiovisual integration response function, fitting an asymmetrical double sigmoidal function to individual response proportions along the asynchrony levels.

Due to the number of trials necessary for a good MMN measurement, duration of the EEG recordings was a primary concern. As to keep measurement time within reasonable limits, MMN was measured at only three asynchrony levels per subject. However, due to the inter-individual variability in audiovisual integration response and simultaneity judgment scores, audiovisual asynchronies were chosen individually for each subject, along the behaviorally estimated individual audiovisual integration functions. Thus, for each subject, the onset asynchronies that evoked maximal, 70% and 20% McGurk responses respectively were selected for the MMN experiment.

MMN experiment

EEG was recorded with a BioSemi ActiveTwo 64-channel EEG system referenced to the average of two mastoid electrodes. Data was sampled at 512 Hz.

Each subject underwent four MMN conditions. Each condition consisted of 450 trials. The rate of deviants was 20%. This is slightly higher than most MMN studies (Garrido et al., 2009). A higher rate was chosen as to collect a sufficient number of deviant ERPs with a low total number of trials. The trial sequence was presented in randomized order, with the constraint that each deviant was

followed by at least two standard trials. Each condition was divided into three blocks. The sequence of blocks was randomized, with the constraint that every condition would be represented by one block in the sequence before any condition was presented in another block. The conditions were as follows:

1. Visual-only MMN, presenting only visual stimul (/tabi/ as standard, /tagi/ as deviant).
2. Audiovisual MMN with stimulus asynchrony corresponding to maximal McGurk response on individual audiovisual integration functions (acoustic /tabi/ + visual /tabi/ as standard, acoustic /tabi/ + visual /tagi/ as deviant) .
3. Audiovisual MMN with stimulus asynchrony corresponding to 70% McGurk responses on individual audiovisual integration functions (acoustic /tabi/ + visual /tabi/ as standard, acoustic /tabi/ + visual /tagi/ as deviant).
4. Audiovisual MMN with stimulus asynchrony corresponding to 20% McGurk responses on individual audiovisual integration functions (acoustic /tabi/ + visual /tabi/ as standard, acoustic /tabi/ + visual /tagi/ as deviant).

MMN paradigms are passive and require no effort from the participant (Lang et al., 1995). Thus, when presenting audiovisual stimuli, control of the subject's gaze is necessary. To achieve this, 5% visual catch stimuli were added to the randomized stimulus sequences, always presenting a congruent audiovisual syllable /tabi/. Subjects were instructed to detect these trials by a key press, while remaining passive during all other trials.

EEG preprocessing

EEG data were analyzed using the EEGLAB toolbox (Delorme and Makeig, 2004). First, continuous EEG data was bandpass filtered between 1 and 30 Hz (Näätänen et al., 2004), before downsampling to 128 Hz. After filtering, data was epoched in the interval -200 to 2600 ms with visual onset at 0. Epochs were then baselined to the 200 ms preceding visual onset. Electrodes dominated by unusual, non-biological waveforms were selected by a measure of kurtosis and data in these channels were interpolated from surrounding electrodes. An ICA

algorithm (runica) was used (not including interpolated channels) to prepare data for rejection of independent components generated by eye artifacts, by means of the EyeCatch algorithm (Bigdely-Shamlo et al., 2013). Epochs were finally thresholded at +/- 100 μV to remove remaining artifacts. The proportion epochs removed from any dataset during preprocessing did not exceed 15%. Data recorded from catch trials were excluded from the analysis.

Correction for visual responses

At this first stage, each subject contributed with data from four conditions, one visual-only (VO), and three audiovisual (AV) with different asynchrony levels (corresponding to maximal, 70% and 20% McGurk response levels), each consisting of ERPs due to standard and deviant stimuli. To correct the audiovisual recordings for visual response activity, mean VO ERPs due to deviants and standards were subtracted from corresponding ERPs in every audiovisual condition. This yielded audiovisual datasets corrected for visual-only ERPs (AV-VO), each consisting of ERPs due to standard and deviant stimuli. These datasets were subsequently epoched to the interval of interest, the onset of the auditory stimulus (-500 to 800 ms around auditory onset). AV-VO epochs were baselined to 200 ms preceding auditory onset. Finally, average AV-VO ERPs due to standard stimuli were subtracted from AV-VO ERPs due to deviant stimuli, producing an AV-VO difference wave, representing an audiovisual MMN corrected for visual MMN.

Results

Behavioral results

Responses from the identification task were recategorized as correct (identification of the auditory (/bi/), or incorrect (identification of the auditory syllable as /gi/ or /di/). Here, we consider the proportion incorrect responses as a measure of the strength of audiovisual integration, as observed in the McGurk illusion. Figure 1 shows the distribution of incorrect auditory identifications as a function of SOA.

Response proportions were arcsine-transformed as to correct for the heterogeneity of variances. We performed a repeated-measures ANOVA on the

response proportions, which revealed a significant effect of SOA upon audiovisual integration as seen in incorrect auditory identifications ($F(13,29)$ = 7.5, $P < 0.0001$).

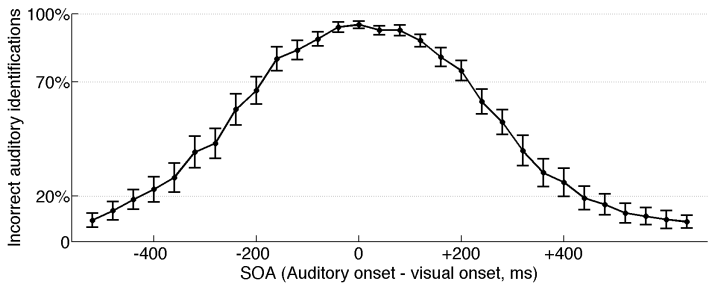


Fig.1. Average incorrect identifications as a function of SOA (ms) in incongruent audiovisual syllable combining auditory /tabi/ with incongruent visual /tagi/. Data points represent mean proportion incorrect identifications. Errorbars represent standard error of mean.

A bonferroni-corrected multiple comparisons test revealed that the rate of incorrect identifications was not significantly different between SOAs of -120 and +200 ms.

Estimation of audiovisual integration function

We further analyzed individual identification results. An asymmetrical double sigmoidal curve was fitted to individual identification data (see Figure 2).

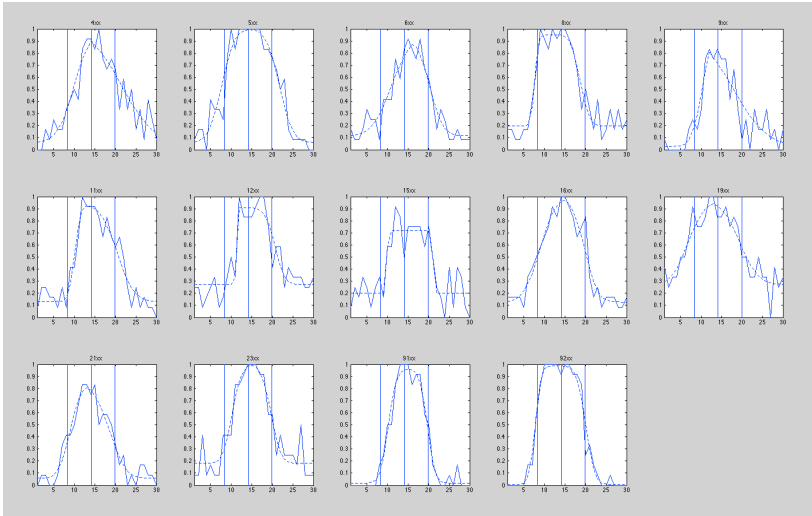


Fig. 2. Individual incorrect response proportions relative to SOAs. An asymmetrical double sigmoidal curve is fitted to each dataset individually.

On basis of these curves, we estimated the SOA at which maximal audiovisual integration was produced. We further estimated the auditory lags, which resulted in 70% integration responses and 20% integration responses, respectively. These individual SOAs were the basis of the MMN measurements.

MMN experiment

Catch trial detection

All participants in the MMN experiment exhibited uniformly high levels in catch trial detection task (mean 97.4% correct detections, SD = 1.7%) This indicates that all participants consistently directed their gaze towards the central area of the visual speech stimulus.

MMN results

Average ERPs and MMN difference waves are represented in Figures 3 and 4. Figure represents the singular ERPs due to visual-only and audiovisual stimuli, before subtraction of visual-only ERPs, however, baselined to acoustic onset. Plots of the scalp distribution of the differential potential at baseline and at negative peaks within the typical MMN range are displayed in Figure 6.

Our marker for audiovisual integration is the McGurk response to the auditory syllable /bi/ in combination with the visual syllable /gi/. As the auditory /bi/ occurs approximately 280 ms after auditory onset, we expect the MMN response to be delayed by the same interval. Where an auditory MMN response is commonly observed in the 100-250 ms interval after onset of the target auditory stimulus (Alho et al., 1990; Garrido et al., 2009; May and Tiitinen, 2010), we would thus expect the MMN for our stimuli to occur in the 380-530 ms range. Previous studies of the McGurk-driven MMN response has shown a latency shifted towards a later interval. Saint-Amour and colleagues (2007) reported an audiovisual MMN response in the 175-400 ms latency range after acoustic onset. Colin (2004) described an McGurk-evoked MMN response in the range 100-500 ms. Sams and colleagues (1991) also reported a wide latency range (200-500 ms). To encompass the latency ranges described in previous findings, our statistical analysis targets the area of 0 to 600 ms after acoustic onset.

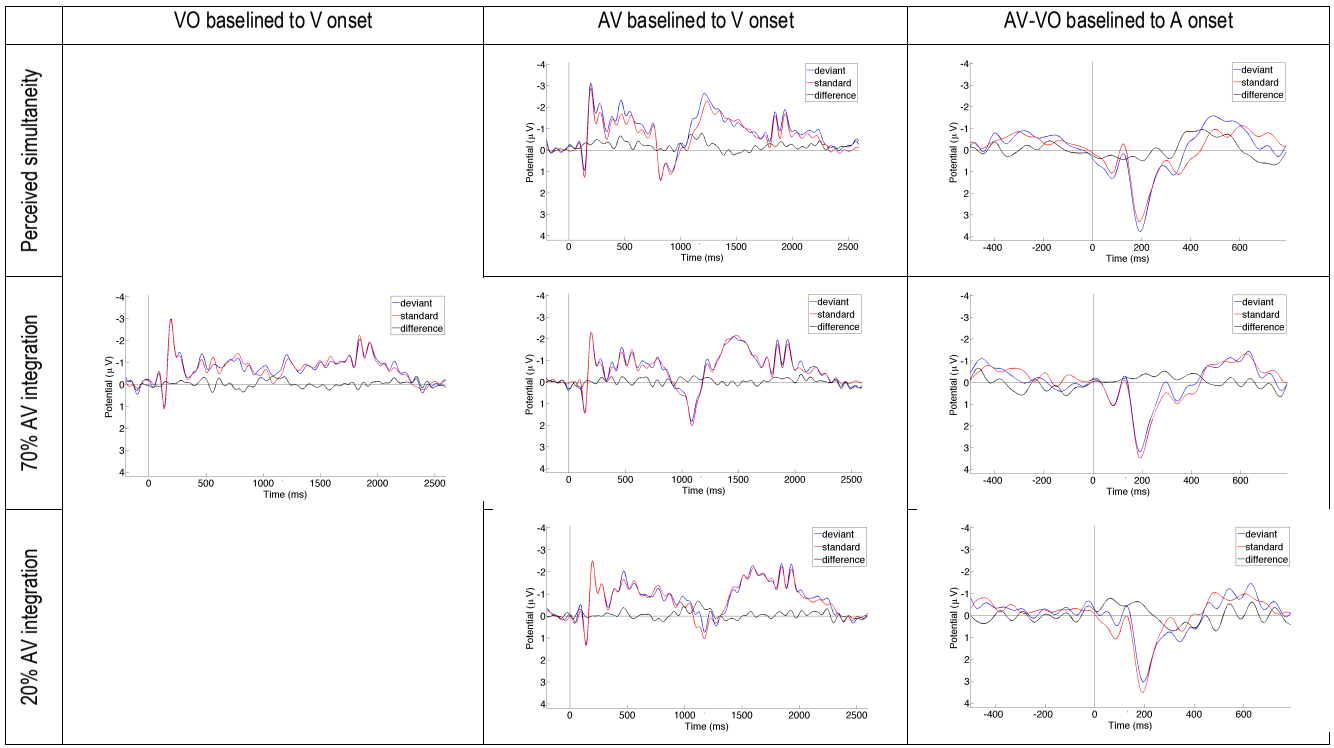


Fig.3. ERPs and difference waves at a centro-parietal electrode, from the VO and AV conditions, and the VO-corrected AV condition (AV-VO). The blue line represents average ERP due to deviants, the red line represents average ERP due to standards, and the black line represents the MMN differential potential. VO and AV potentials are epoched and baselined to the visual onset. The AV-VO potentials are epoched and baselined to the auditory onset. Note that the interval between visual and auditory onset is subject to interindividual variability.

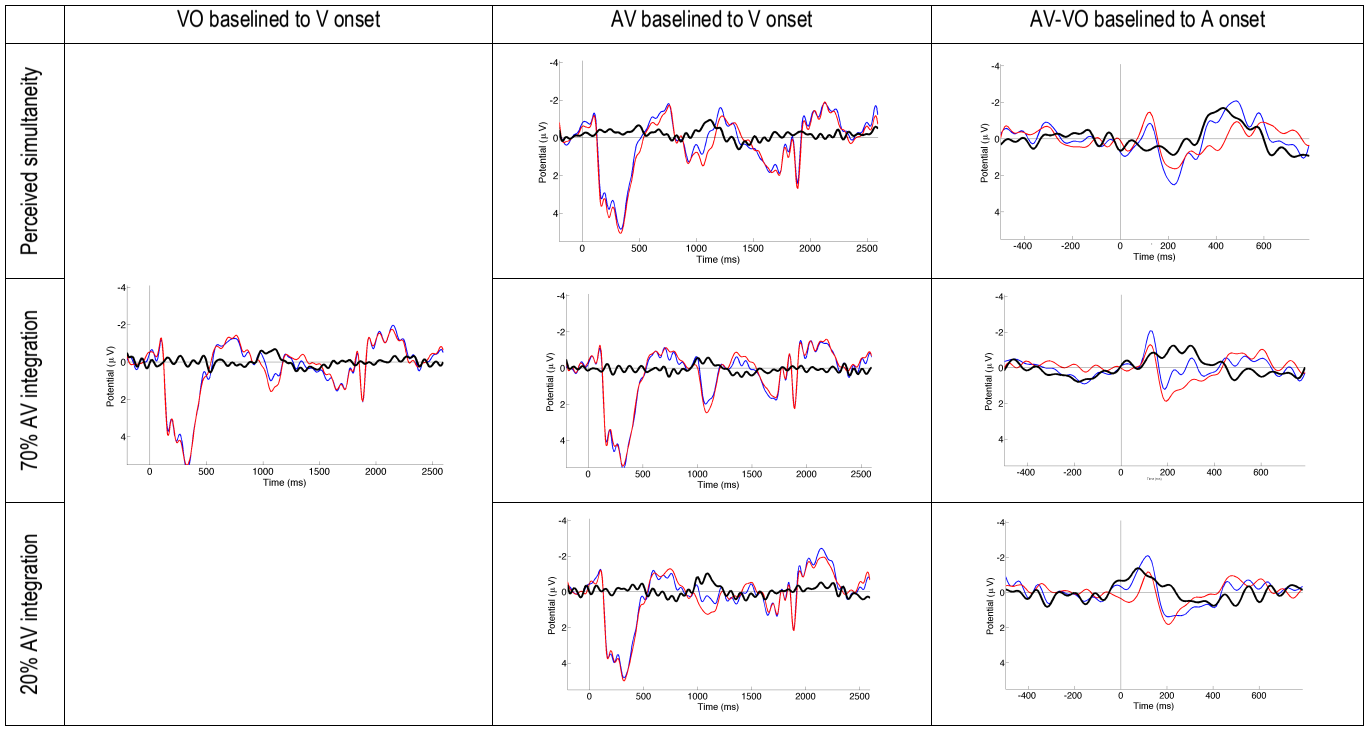


Fig.4. ERPs and difference waves at a parieto-occipital electrode, from the VO and AV conditions, and the VO-corrected AV condition (AV-VO). The blue line represents average ERP due to deviants, the red line represents average ERP due to standards, and the black line represents the MMN differential potential. VO and AV potentials are epoched and baselined to the visual onset. The AV-VO potentials are epoched baselined to the auditory onset. Note that the interval between visual and auditory onset is subject to interindividual variability.

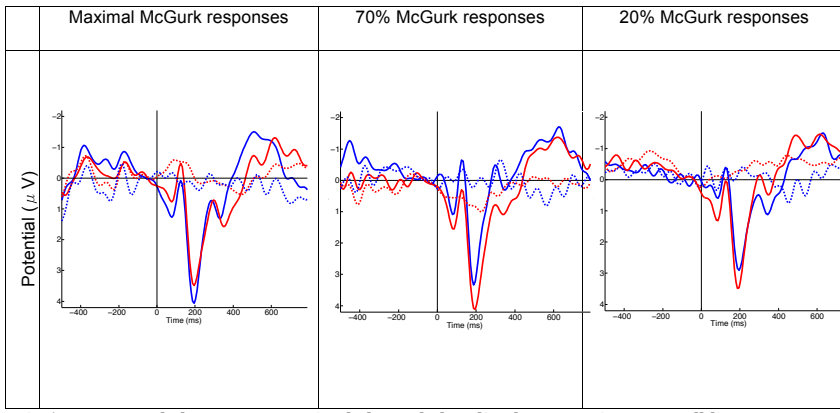


Fig.5. ERPs recorded at a centro-parietal electrode baselined to acoustic onset. Full lines represent ERPs due to audiovisual stimuli (not corrected by subtraction of visual-only ERPs), dashed lines represent ERPs due to visual-only stimuli. Blue lines represent deviant stimuli, red lines represent standard stimuli.

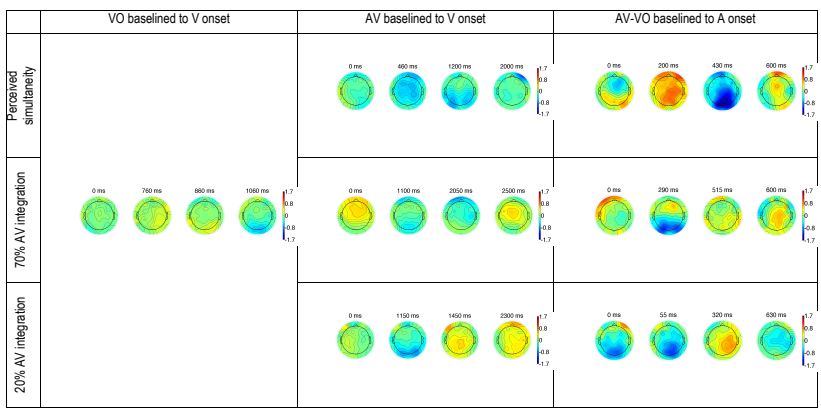


Fig.6. Scalp plots of the differential potential generated by subtracting the average ERP due to standard trials from the average ERP due to deviant trials. VO and AV scalp potentials are baselined and epoched to visual onset. AV-VO scalp potentials are epoched and baselined to auditory onset.

Individual difference waves for each subject were generated by subtracting the mean ERP due to standard stimuli from the mean ERP due to deviant stimuli in each condition. Targeting the negative deviation in the 0-600 ms interval as a signature of an MMN response, this period of the difference waves from each condition was subjected to a repeated-measures, one-tailed clustered permutation test (Bullmore et al., 1999; for a detailed description, cf. Groppe et al., 2011) utilizing a family-wise alpha level of 0.05 and 2500 random permutations. A one-tailed test was used, as any effect due to mismatch

negativity would only be on the negative tail. All time points in the target interval at all 64 scalp electrodes were included.

Difference waves produced by stimuli with SOAs corresponding to maximal integration and 70% integration responses proved significant mainly at parieto-occipital and occipital electrodes (see fig. 7 and fig. 8, respectively). For the maximal integration stimuli, the MMN wave reached significant levels in the interval 350-584 ms, exceeding a critical t-score of -1.89, corresponding to a *P*-value of 0.05. In the 70% integration-condition, the distribution of electrodes yielding significant differential potentials (exceeding a critical t-score of -1.81 corresponding to a *P*-value of 0.05) was slightly wider, and the latency was shifted forward to the 100-350 ms interval. In the case of SOAs corresponding to 20% integration responses however, the MMN did not differ significantly from zero (lowest *P*-value 0.4880).

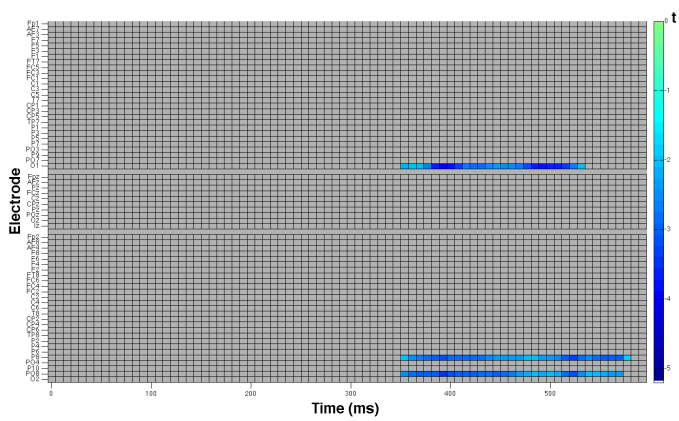


Fig.7. Plot of t-scores for the difference wave produced by AV-VO stimuli in the SOA corresponding to perceived simultaneity, resulting from a repeated-measures, one-tailed clustered permutation test. Colored areas represent t-scores below the critical t-score of –1.89 corresponding to a *P*-value of 0.05, thus being deemed to be statistically significant.

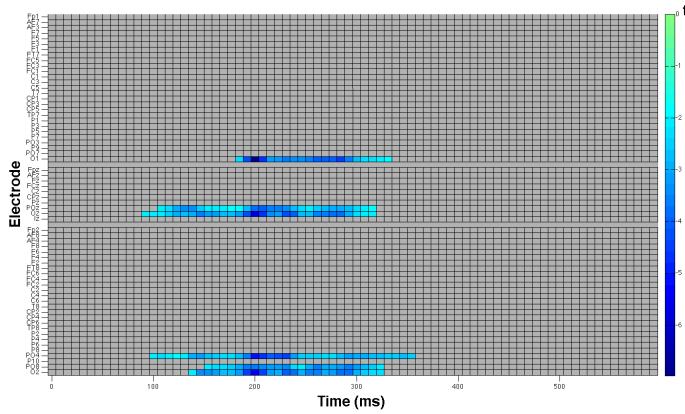


Fig.8. Plot of t-scores for the difference wave produced by AV-VO stimuli in the individual SOAs producing 70% McGurk responses in the behavioral task, resulting from a repeated-measures, one-tailed clustered permutation test. Colored areas represent t-scores below the critical t-score of –1.81 corresponding to a *P*-value of 0.05, thus being deemed to be statistically significant.

Discussion

Behavioral estimate of the temporal window

Identification performance with incongruent stimuli revealed an audiovisual integration function of SOAs with an asymmetrical distribution around stimulus synchrony, similar to previous studies (van Wassenhove et al., 2007). The width of the integration window is estimated to -120 to +200 ms. This is wider than findings of van Wassenhove and colleagues (2007). The width is on level with findings of Munhall and coworkers (1996), although it is shifted slightly towards visual lags.

MMN correlates to the behavioral temporal window

When presenting stimuli at SOAs determined by individual maximal behavioral McGurk response, deviant stimuli evoke a negative differential potential, which is clearly seen in the MMN. The latency of this potential is approximately 360-580 after auditory onset, which correspond to 80-300 ms after onset of the target consonant /bi/. The distribution of the difference wave over the scalp is limited to parietal, parieto-occipital and occipital electrodes. This clearly differs from previous studies of both auditory MMN (Näätänen and Alho, 1995; Näätänen et al., 1978) and audiovisual MMN (Ponton et al., 2009; Saint-Amour et

al., 2007; Stekelenburg and Vroomen, 2012), which is observed widely over central, fronto-central and frontal sites.

In the case of MMN recorded with stimuli at SOAs corresponding to individual 70% McGurk responses, a significant differential potential was found across a wider range of parietal, parieto-occipital and occipital sites. However, the latency is shifted to a range *preceding* onset of the auditory stimulus, in the range from 100 to 300 ms after onset of the auditory bisyllable. This excludes the possibility of the effect being a modulation of auditory activity. As ERPs due to visual-only stimuli were subtracted before analysis, an interpretation as an oddball response to the visual stimulus alone is also left out. A possible third explanation is that the MMN reflects genuine multisensory activity evoked by the visual stimulus, or, by a modulation of auditory processing by means of the visual stimulus, even before the targeted auditory syllable is presented. Visual speech is known to modulate auditory neural activity (Möttönen et al., 2002; Pekkola et al., 2005, 2006), but ERPs due to visual-only stimuli were subtracted in calculating the AV-VO potential. This only leaves room for a modulation of auditory processing *before* onset of the target auditory syllable as a possible explanation.

For stimuli presented at SOAs individually corresponding to 20% McGurk responses, the differential potential was indiscriminable from background noise, and it did not reach significant levels at any point in the latency range 0 to 600 ms.

General discussion

While our behavioral findings are similar to previous findings, our MMN results are less clear. The topographical distribution of the MMN potential does not match prior audiovisual MMN studies. Also, the latency shift of the MMN difference wave in the 70% McGurk response condition is hard to explain, without reference to confounding factors. Below we address a few specific concerns.

The correlation between behavioral response strength and AV MMN strength is unknown

In auditory MMN studies, MMN amplitude varies with the depth of the perceived stimulus deviance, or, its distance from the just noticeable difference (Garrido et al., 2009; Näätänen et al., 1978). Following this, a reduction of the McGurk MMN could be expected due to the reduction of McGurk responses with increasing SOA. However, whereas the correlation between acoustic stimulus deviance and MMN depth is known, a corresponding relation between audiovisual stimulus deviance, let alone McGurk illusion strength and audiovisual MMN depth remains unknown. If the relation between behavioral McGurk illusion strength and MMN is nonlinear, i.e. if an MMN response requires a uniform high level of McGurk illusion to be elicited at all, it would severely hamper our efforts at looking for correlations between grades of behavioral responses and neural responses. While the correlation between behavior and MMN strength for audiovisual stimuli remains to be investigated, previous studies of McGurk-driven MMN report high levels of behavioral McGurk responses (e.g. Saint-Amour et al., 2007; Stekelenburg and Vroomen, 2012) or even exclusion of subjects with weak behavioral McGurk response (Kislyuk et al., 2008).

Using individually estimated SOAs

In our approach, we aimed at producing a neural correlate to individually estimated audiovisual integration functions. Due to individual differences, each subject was presented with different audiovisual asynchronies in the three MMN conditions. The advantage of this is that we were able to target the relevant points on the function for each subject. This, however, rests on the assumption that the behavioral data on basis of which the function is estimated accurately represents individual levels of audiovisual integration. Furthermore, it also requires that the individual differences in behavior are reflected in a difference in neural processing. If so, audiovisual ERPs recorded with individually estimated audiovisual asynchronies should yield similar responses. However, these assumptions may not hold, e.g. if the audiovisual neural response is more uniform than represented in behavior. In that case, ERPs due to stimuli with

individual audiovisual lags would basically produce potentials with different latencies. If so, the resulting ERPs and difference waves would be blurred.

References

Alho, K., Sainio, K., Sajaniemi, N., Reinikainen, K., and Näätänen, R. (1990). Event-related brain potential of human newborns to pitch change of an acoustic stimulus. Electroencephalogr. Clin. Neurophysiol. Potentials Sect. *77*, 151–155.

Bertelson, P., Vroomen, J., Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. *62*, 321–332.

Bigdely-Shamlo, Kreutz-Delgado, K., Kothe, C, and Makeig, S. (2013). EyeCatch: Data-mining over Half a Million EEG Independent Components to Construct a Fully-Automated Eye-Component Detector. (IEEE Engineering in Biology and Medicine Conference, Osaka),.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M.J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imaging *18*, 32–42.

Colin, C. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. *113*, 495–506.

Colin, C., Radeau, M., Soquet, A., and Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. Clin. Neurophysiol. *115*, 1989–2000.

Conrey, B., and Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J. Acoust. Soc. Am. *119*, 4065.

Czigler, I. (2007). Visual Mismatch Negativity: Violation of Nonattended Environmental Regularities. J. Psychophysiol. *21*, 224–230.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods *134*, 9–21.

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

Eskelund, K., Tuomainen, J., and Andersen, T.S. (2010). Multistage audiovisual integration of speech: dissociating identification and detection. Exp. Brain Res. *208*, 447–457.

Garrido, M.I., Kilner, J.M., Stephan, K.E., and Friston, K.J. (2009). The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol. *120*, 453–463.

Grant, K.W., Wassenhove, V. van, and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. Speech Commun. *44*, 43–53.

Groppe, D.M., Urbach, T.P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. Psychophysiology *48*, 1711–1725.

Hietanen, J.K., Manninen, P., Sams, M., and Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? Eur. J. Cogn. Psychol. *13*, 395–407.

Kislyuk, D.S., Möttönen, R., and Sams, M. (2008). Visual Processing Affects the Neural Basis of Auditory Discrimination. J. Cogn. Neurosci. *20*, 2175–2184.

Lang, A.H., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S., and Aaltonen, O. (1995). Practical issues in the clinical application of mismatch negativity. Ear Hear. *16*, 118–130.

Maier, J.X., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. J. Exp. Psychol. Hum. Percept. Perform. *37*, 245–256.

Massaro, D.W., and Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. J. Acoust. Soc. Am. *108*, 784–789.

May, P.J.C., and Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. Psychophysiology *47*, 66–122.

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

Möttönen, R., Krause, C.M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. Brain Res. Cogn. Brain Res. *13*, 417–425.

Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. Percept. Psychophys. *58*, 351–362.

Näätänen, R., and Alho, K. (1995). Mismatch negativity-a unique measure of sensory processing in audition. Int. J. Neurosci. *80*, 317–337.

Näätänen, R., Gaillard, A.W.K., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst.) *42*, 313–329.

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. Clin. Neurophysiol. *115*, 140–144.

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., and Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. Brain Res. *1323*, 84–93.

Pazo-Alvarez, P., Cadaveira, F., and Amenedo, E. (2003). MMN in the visual modality: a review. Biol. Psychol. *63*, 199–236.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A., and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport *16*, 125–128.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., and Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. Hum. Brain Mapp. *27*, 471–477.

Ponton, C.W., Bernstein, L.E., and Auer, E.T. (2009). Mismatch Negativity with Visual-only and Audiovisual Speech. Brain Topogr. *21*, 207–215.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., and Foxe, J.J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia *45*, 587–597.

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

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

Soto-Faraco, S., and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. J. Exp. Psychol. Hum. Percept. Perform. *35*, 580–587.

Stekelenburg, J.J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia.

Stekelenburg, J.J., Vroomen, J., and de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. Neurosci. Lett. *357*, 163–166.

Sumby, W.H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Am. *26*, 212.

Vatakis, A., and Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. Percept. Psychophys. *69*, 744–756.

Vatakis, A., Ghazanfar, A.A., and Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. J. Vis. *8*, 14–14.

Van Wassenhove, V., Grant, K.W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. Neuropsychologia *45*, 598–607.

Zampini, M., Shore, D.I., and Spence, C. (2003). Audiovisual temporal order judgments. Exp. Brain Res. *152*, 198–210.

8 Other work published during the PhD study

D Audiovisual integration in speech perception: a multi-stage process

Audiovisual integration in speech perception: a multi-stage process

KASPER ESKELUND[1] , JYRKI TUOMAINEN[2] AND TOBIAS ANDERSEN[1]

[1] *Cognitive Systems, Department of Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Speech Hearing and Language Sciences, University College London, UK*

Integration of speech signals from ear and eye is a well-known feature of speech perception. This is evidenced by the McGurk illusion in which visual speech alters auditory speech perception and by the advantage observed in auditory speech detection when a visual signal is present. Here we investigate whether the integration of auditory and visual speech observed in these two audiovisual integration effects are specific traits of speech perception. We further ask whether audiovisual integration is undertaken in a single processing stage or multiple processing stages.

INTRODUCTION

Integration effects such as the McGurk effect (McGurk and MacDonald, 1976) and the detection advantage associated with audiovisual speech (Grant and Seitz, 2000) show that vision and hearing integrate when perceiving speech. It is, however, unknown whether the processes underlying such audiovisual integration are specific for perception of speech, or if they pertain to audiovisual perception in general. Moreover, audiovisual integration is often tacitly assumed to be undertaken in a single step (Massaro, 1998; Vatakis and Spence, 2007). In the experiment reported here, we test whether audiovisual integration as seen in the McGurk effect and the audiovisual detection advantage occurs for both non-speech and speech perception. We further test these integration effects as to investigate whether they show different properties in non-speech and speech conditions. If the latter is the case, it may indicate that the effects are related to dissociated processes supporting the claim that audiovisual integration of speech is multi-faceted.

Grant and Seitz (2000) showed that seeing a synchronous visual speech signal is advantageous when detecting an acoustic speech signal masked by noise. Presenting three sentences in audiovisual and auditory-only formats masked by acoustic noise, they found that the advantage associated with the presence of the

visual speech signal in the audiovisual stimulus was equivalent to a 1.6 dB gain of the auditory-only stimulus. Investigating the dynamics of the acoustic and visual stimuli, they showed that the magnitude of the advantage depends on the degree of correlation between changes in lip opening area and sound intensity. On this basis, they proposed the *peak listening hypothesis*, stating that cues in the visual signal guides the listener to the spectral and temporal parts of the acoustic signal with the most favourable signal-to-noise ratio.

Experimenting with single-syllable audiovisual speech stimuli, Bernstein and colleagues (2004) found that preparatory lip gestures preceding acoustic onset may be responsible for the effect. Thus, even if the visual stimulus was exchanged with a non-speech geometric figure, it still evoked a detection advantage as long as the onset of the preparatory articulatory movements was retained. The authors concluded that the effect was not specific for speech stimuli and could be produced by any visual pre-cueing of auditory onset.

In a similar experiment, however, Schwartz and colleagues (Schwartz et al., 2004) observed that the audiovisual detection advantage was eliminated for non-speech visual stimuli, even if the dynamics of speech was represented. Further, in the results of Bernstein and co-workers (2004), the detection advantage was lower for non-speech visual stimuli. However, it is difficult to determine whether these findings are due to the geometrical visual stimuli lacking relevant cues present in natural visual speech or whether they are due to observers not using available cues because the non-speech figures seem irrelevant to the observer and visual cues are thus to a lesser degree bound together with the auditory signal. These studies have thus targeted a contrast between visual stimuli in addition to the difference between non-speech and speech perception, or, the perceptual set. Thus, the findings on the speech-specificity of the audiovisual detection advantage are inconclusive. In contrast, the purpose of the current experiment is to ask directly if the stimulus needs to represent speech.

As any stimulus containing a minimum of phonetic cues will be perceived as speech, it is difficult to devise a meaningful comparison of speech perception with non-speech perception using the same stimulus. Tuomainen et al. (2005)

provided an elegant solution to this, using sine wave speech (SWS) stimuli (Remez et al., 1981). In SWS, centre frequencies of the three lowest formants of a natural speech token are extracted. A novel stimulus is generated by letting three sine tones track these frequencies and their amplitudes. This synthetic stimulus thus contains only faint phonetic cues. When listening to SWS, naïve subjects tend not to perceive any phonetic content, but rather report hearing synthetic, meaningless sounds. However, when informed on the phonetic content, the weak phonetic cues are perceived and SWS heard as speech. Remez and colleagues (1981) interpreted this as evidence for a speech-specific mode of perception. Since SWS can be perceived as speech or as non-speech it is an ideal stimulus for investigating effects that supposedly occur specifically in speech perception.

With this approach, Tuomainen et al. (2005) found that the McGurk illusion only occurred for SWS when it is perceived as speech. This result indicates that the audiovisual integration process underlying the McGurk effect is speech-specific. To test the speech-specificity of the audiovisual detection advantage, we investigated if visual speech may assist auditory detection of SWS when perceived as non-speech and when perceived as speech (Eskelund et al., 2010).

METHODS

18 participants (6 female), mean age 25 (range 21 to 30) all reported normal hearing and normal or corrected-to-normal vision. 4 were excluded; 3 due to recognizing SWS as speech before entering the speech condition of the experiment and 1 due to not being able to discriminate among the SWS stimuli.

All stimuli were based on the speech recordings and SWS replicas produced by Tuomainen et al. (2005). Four auditory stimuli were used, SWS /omso/ and /onso/, and natural /omso/ and /onso/. A total of eight audiovisual stimuli were produced by combining SWS and natural speech tokens with video of the talking face, resulting in congruent and incongruent audiovisual combinations of /omso/ and /onso/.

In identification tasks, sound intensity of both SWS replicas were 77 dB SPL, while natural speech stimuli had intensities of 68 dB SPL and 70 dB SPL for natural /omso/ and /onso/ respectively. In detection tasks, a noise masker with a constant intensity of 65 dB SPL was added, while the intensity of the acoustic

stimulus was varied, using a 2AFC paradigm and an adaptive staircase procedure. Duration of the masker was that of the stimulus plus two random intervals of 100-300 ms added before and after stimulus onset to eliminate any cues from onset of masker and target.

In the non-speech condition, subjects perceive SWS as non-speech sounds. Thus, when identifying and detecting audiovisual SWS tokens, they have little reason to look at the talking face, precluding any integration of sight and hearing. This might be a trivial confound for any reduction in the McGurk illusion in the non-speech condition. To control that subjects were actually looking at the screen, we included a secondary visual detection task. A white dot was overlaid the nose of the talking face for the same duration as each stimulus plus surrounding random intervals. In 20% of trials, the white dot disappeared for 200 ms at the onset of consonants /m/ and /n/. Subjects had to detect if the dot blinked.

Before the experiment, subjects were trained in discriminating the auditory SWS tokens /omso/ and /onso/ in arbitrary non-speech categories ('sound 1' and 'sound 2'). The experiment began with a non-speech condition during which subjects were naïve about the speech origin of SWS. First, subjects identified SWS auditory-only, audiovisual congruent and audiovisual incongruent stimuli in arbitrary categories, then they performed the detection task with auditory-only and audiovisual congruent tokens of /omso/. After a short break, subjects were then informed about the speech-like nature of SWS and then followed the speech condition (Eskelund et al., 2010), repeating the identification and detection tasks, with the change in the identification task, that stimuli were now categorised as 'omso' and 'onso'. Additionally, in the speech condition, a separate task of identifying natural auditory-only, audiovisual congruent and audiovisual incongruent speech tokens was performed.

As the experiment hinges on a shift in perceptual set between non-speech and speech perception, the hearing experience of participants was checked before and after each condition. Included subjects did not associate SWS with speech before being informed about its phonetic origin and they all reported hearing SWS as speech during all tasks in the speech condition.

RESULTS

Identification tasks

Proportions of correct responses in the identification of auditory stimuli are displayed in Figure 1 for each stimulus type. The results were subjected to an arcsine transformation and analyzed with a two-way (Stimulus x Conditions) repeated-measures ANOVA (Eskelund et al., 2010). The interaction between Stimulus and Conditions was significant. This indicated that the effect of Condition differed for the three stimulus types.

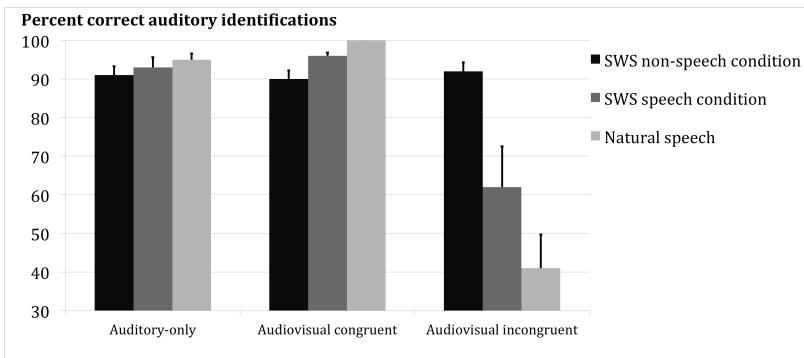


Fig. 1: Results from the auditory identification tasks. Bars represent percent correct auditory identifications by stimulus type and condition. Error bars represent standard error of mean. With audiovisual incongruent stimuli, the difference in identification performance between SWS in non-speech and speech conditions indicates that audiovisual integration of phonetic content only occurs in speech perception.

For auditory-only stimuli, there was no significant effect of Condition. In the case of congruent audiovisual speech, there was a significant effect of Condition. This indicated that performance was highest for natural speech, lower for SWS in the speech condition and lowest for SWS in non-speech condition. For congruent audiovisual stimuli, integrating the talking face with the voice should improve performance. Therefore this effect can be interpreted as a stronger influence from vision on audition when the stimulus is perceived as speech.

For incongruent audiovisual stimuli, which would tend to induce a McGurk illusion and hence is the pivotal stimulus class of the identification task, the effect of Condition was significant, reflecting that performance was lowest for natural speech, somewhat higher for SWS in the speech condition and highest for

SWS in the non-speech condition. For audiovisual speech, seeing an incongruent talking face should obstruct identification due to the McGurk effect. This result could be interpreted as a stronger influence of vision on audition when SWS is perceived as speech.

A comparison of performance with auditory-only SWS, audiovisual congruent SWS and audiovisual incongruent SWS in the non-speech condition revealed no significant difference, indicating that the visual signal was not integrated into the auditory signal while perceived as non-speech.

Detection tasks

Detection thresholds were calculated as the mean of the last 10 responses of the adaptive staircase. Average thresholds are shown in Figure 2. Mean detection advantage of audiovisual stimuli over auditory-only presentation was 2.66 dB SPL. The detection difference between non-speech and speech conditions was negligible.

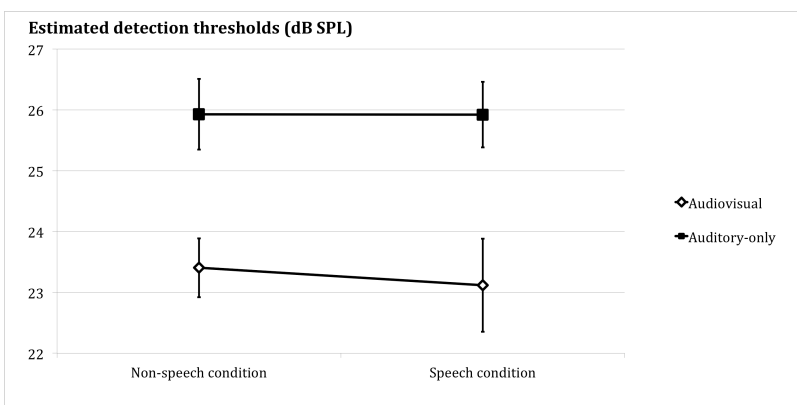


Fig. 2: Results from the auditory detection tasks. Points represent auditory detection threshold per stimulus type and condition. Error bars represent standard error of mean.

Results were subjected to a two-way (Stimulus x Conditions) repeated-measures ANOVA (Eskelund et al., 2010). In contrast to the interaction seen in identification tasks, no significant interaction between factors Stimulus and Condition was found, indicating that the audiovisual detection advantage is not influenced by the shift from non-speech to speech perception. A significant main effect of Stimulus was found, however, expressing that a detection advantage for

audiovisual SWS over auditory-only SWS occurred. No main effect of Condition was found.

Secondary task

Detection of the occurrence of the white dot remained consistently high across all tasks. No significant difference in secondary task performance was found between tasks (Eskelund et al., 2010). This indicates that participants were following the instructions to look at the screen in all tasks, even in the non-speech condition where the talking face was irrelevant to tasks.

DISCUSSION

Identification results confirmed the observation of Tuomainen and colleagues (2005) that the McGurk illusion does occur for SWS but only when perceived as speech. This finding suggests that audiovisual integration of phonetic content is a speech-specific effect.

The audiovisual detection advantage observed for SWS in the present study is in concordance with Grant and Seitz' findings (2000) for natural speech. Interestingly, this effect was not influenced by whether SWS was perceived as non-speech or speech. The finding is in agreement with the interpretation of Bernstein and colleagues (2004), that the detection advantage is not specific for speech perception. In contrast, the McGurk effect only occurred in the speech condition. This suggests that the audiovisual detection advantage is not speech-specific whereas the McGurk effect is.

Our results thus further suggest that the detection advantage and the McGurk illusion are caused by two dissociated mechanisms, integrating different features of the audiovisual signal according to the perceptual set of the observer. This shows that audiovisual integration in speech perception is not, as Soto-Faraco and Alsius (2009) put it, a "monolithic", but rather a "multi-faceted" process.

Extending upon the concept of Auditory Scene Analysis (Bregman, 1990), Schwartz and colleagues (2004) proposed a two-stage model of audiovisual integration. In their concept of Audiovisual Scene Analysis, the early stage forms a correspondence between auditory and visual signals in a "primitive grouping"

(Barker et al., 1998). This bimodal correspondence facilitates auditory detection by aiding segregation of auditory sources. Phonetic content is identified at a later stage, which receives the grouped bimodal signal.

In a recent series of experiments, Nahorna and colleagues (2011, 2010) showed that the illusory phonetic percept in the McGurk illusion could be disintegrated when the expectation of phonetic audiovisual congruence was changed. In one condition, subjects were presented with a series of congruent audiovisual syllables followed by an incongruent audiovisual syllable, which had to be identified. This produced a McGurk illusion. In a second condition, subjects were presented with a series of incongruent audiovisual speech syllables, again followed by an incongruent audiovisual syllable, which had to be identified. Now the McGurk illusion disappeared. According to the two-stage model, the early stage operates under the assumption of congruence, thus integrating unexpected incongruent auditory and visual signals as observed in the first condition. However, when evidence for audiovisual incongruence accumulates as in the second condition, the weight of the coherence evaluation in the early stage changes. As the grouping thus is reduced, the weight of the non-matching visual signal is decreased in the phonetic decision in the second stage. This results in the unbinding of the incongruent signals, eliminating the McGurk illusion. Our results agree with this approach. The early stage groups auditory and visual signals and facilitates auditory detection regardless of whether the listener is perceptually set for speech. In contrast, the integration of phonetic cues occurs in the later stage, which our results suggest is speech-specific.

Our current findings thus fit well with a multi-stage model as suggested by Schwartz (2004) and Nahorna (2011, 2010). An early stage would assess audiovisual coherence and exploit bimodal covariation to enhance the effective auditory signal-to-noise ratio. This is the stage involved in the audiovisual detection advantage. A later stage would identify phonetic content on basis of the percept generated in the first stage. This stage underlies the McGurk effect and is speech-specific.

REFERENCES

Barker, J. P., Berthommier, F., and Schwartz, J.-L. (1998). Is Primitive AV Coherence an Aid to Segment the Scene? In Proceedings from the International Conference on Auditory-Visual Speech Processing 1998 (Terrigal, Australia).

Bernstein, L., Auer, E. T. J., and Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading.

Bregman, A. S. (1990). Auditory scene analysis: the perceptual organization of sound (MIT Press).

Eskelund, K., Tuomainen, J., and Andersen, T. S. (2010). Multistage audiovisual integration of speech: dissociating identification and detection. Experimental Brain Research *208*, 447–457.

Grant, K. W., and Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. The Journal of the Acoustical Society of America *108*, 1197.

Massaro, D. W. (1998). Perceiving talking faces: from speech perception to a behavioral principle (MIT Press).

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

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (2011). Binding and unbinding in audiovisual speech fusion: Follow-up experiments on a new paradigm. In Proceedings from the International Conference on Auditory-Visual Speech Processing 2011 (Volterra, Italy: Kungliga Tekniska Högskolan, Sweden).

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (2010). Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context. In In Proceedings from the International Conference on Auditory-Visual Speech Processing 2010 (Hakone, Kanagawa, Japan: Kumamoto University, Japan).

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (1981). Speech perception without traditional speech cues. Science *212*, 947–949.

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

Soto-Faraco, S., and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. Journal of Experimental Psychology: Human Perception and Performance *35*, 580–587.

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (2005). Audio-visual speech perception is special. Cognition *96*, B13–B22.

Vatakis, A., and Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. Perception & Psychophysics *69*, 744–756.

Speech perception integrates signals from ear and eye. This is witnessed by the hearing benefits provided by seeing the speaker while listening, as well as by perceptual effects produced by altering the relation between acoustic and visual speech signals.

What requirements must speech stimuli meet to support audiovisual integration? This thesis addresses two key aspects of this question:

First, the visual speech signal emanates from a talking face. How do face perception processes influence the audiovisual speech percept? Does the configuration of facial features influence audiovisual integration?

Second, acoustic and visual signals emanating from the same source coincide temporally. How tolerant is audiovisual integration of speech to asynchrony between signals in the two modalities?

These aspects are targeted here by a series of combined behavioral and electrophysiological experiments. In parallel, methodological findings on the use of electrophysiology for investigation of audiovisual speech perception are reported and discussed.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk