

Electrophysiological evidence for a multisensory speech-specific mode of perception

Jeroen J. Stekelenburg*, Jean Vroomen

Department of Psychology, Tilburg University, The Netherlands

ARTICLE INFO

Article history:

Received 16 June 2011

Received in revised form 18 January 2012

Accepted 24 February 2012

Available online 4 March 2012

Keywords:

Multisensory integration

Audiovisual speech

Sine-wave speech

McGurk illusion

Mismatch negativity

ABSTRACT

We investigated whether the interpretation of auditory stimuli as speech or non-speech affects audiovisual (AV) speech integration at the neural level. Perceptually ambiguous sine-wave replicas (SWS) of natural speech were presented to listeners who were either in 'speech mode' or 'non-speech mode'. At the behavioral level, incongruent lipread information led to an illusory change of the sound only for listeners in speech mode. The neural correlates of this illusory change were examined in an audiovisual mismatch negativity (MMN) paradigm with SWS sounds. In an oddball sequence, 'standards' consisted of SWS/onso/coupled with lipread/onso/, and 'deviants' consisted of SWS/onso/coupled with lipread/omso/. The AV deviant induced a McGurk-MMN for listeners in speech mode, but not for listeners in non-speech mode. These results demonstrate that the illusory change in the sound by incongruent lipread information evoked an MMN which presumably takes place at a pre-attentive sensory processing stage.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

An important question about speech perception is whether speech is processed as all other sounds (Fowler, 1996; Kuhl, Williams, & Meltzoff, 1991; Massaro, 1998), or whether there are specialized mechanisms responsible for translating the acoustic signal into phonetic segments (Repp, 1982; Tuomainen, Andersen, Tiippana, & Sams, 2005). A relevant finding favoring the notion of 'speech-specificity' was provided by Remez, Rubin, Pisoni, and Carrell (1981). They created time-varying sine-wave speech (SWS) replicas of natural speech that were perceived by naïve listeners as non-speech whistles, bleeps, or 'computer sounds', but when another group of subjects was instructed about the speech-like nature of the stimuli, they could easily assign linguistic content to the sounds. This ambiguous nature of SWS has provided researchers a tool to study the neural and behavioral specificity of speech sound processing, because identical acoustic stimuli can be used that are perceived differently, depending on the mode of the listener. In this way, it has been shown with functional magnetic resonance imaging (fMRI) that SWS stimuli elicit stronger activity within the left posterior superior temporal sulcus for listeners in speech mode than for listeners in non-speech mode (Möttönen et al., 2006). The phonetic content of SWS is also more likely integrated with visual information of lipread speech if listeners are in speech mode rather

than non-speech mode, suggesting that audiovisual (AV) integration of phonetic information only occurs when listeners perceive the sound as speech (Tuomainen et al., 2005; Vroomen & Baart, 2009). Not all aspects of audiovisual integration, though, have been found to depend on the mode of the listener. One example is that perception of temporal synchrony between a heard SWS sound and lipread information is not different for listeners in speech or non-speech mode (Vroomen & Stekelenburg, 2011). Furthermore, lipread information can improve auditory detection of SWS targets in noise, but the size of the improvement does not depend on the mode of the listener (Eskelund, Tuomainen, & Andersen, 2011). This thus indicates that phonetic, but not temporal and loudness processing of the sound depend on the speech mode of the listener.

In the current study, we searched for a neural correlate of the distinction between a 'speech' and 'non-speech' mode of audiovisual speech processing. It is generally acknowledged that in audiovisual speech the auditory and visual signals are integrated at some processing stage into a coherent multisensory representation. Hemodynamic studies have shown that multisensory cortices (superior temporal sulcus/gyrus) (Callan et al., 2004; Calvert, Campbell, & Brammer, 2000; Skipper, Nusbaum, & Small, 2005), 'sensory-specific' cortices (Callan et al., 2003; Calvert et al., 1999; Kilian-Hütten, Valente, Vroomen, & Formisano, 2011; von Kriegstein & Giraud, 2006) and motor areas (Ojanen et al., 2005; Skipper et al., 2005) are involved in audiovisual speech integration. Electrophysiological studies have shown that AV speech interactions occur in the auditory cortex as early as 100 ms (Arnal, Morillon, Kell, & Giraud, 2009; Besle, Fort, Delpuech, & Giard, 2004; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, &

* Corresponding author at: Department of Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: +31 134668171.

E-mail address: j.j.stekelenburg@uvt.nl (J.J. Stekelenburg).

Poeppel, 2005). Here, we used the well-known mismatch negativity (MMN) component in the electroencephalogram (EEG) as a neural marker of audiovisual speech integration. The MMN is a fronto-centrally negative event-related potential (ERP) component that is elicited by sounds that violate the automatic predictions of the central auditory system (Näätänen, Gaillard, & Mäntysalo, 1978). The MMN is measured by subtracting the ERP of a frequent 'standard' sound from an infrequent 'deviant' sound, and it appears as a negative deflection with a fronto-central maximum peaking around 150–250 ms from the onset of the sound change. The MMN is most likely generated in the auditory cortex and presumably reflects pre-attentive auditory deviance detection (Näätänen, Paavilainen, Rinne, & Alho, 2007).

Important from our perspective is that the MMN has also been successfully used to probe the neural mechanisms underlying the integration of information from different senses, as in the case of hearing and seeing speech (i.e., lipreading; Colin et al., 2002; Kislyuk, Möttönen, & Sams, 2008; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Sams et al., 1991). In a typical experiment, an intersensory conflict is created between the heard and lipread information, e.g., hearing/ba/, but seeing the face of a speaker articulate/ga/. The crucial aspect of this stimulus is that the incongruent lipread information leads to an illusory change in the perceived quality of the sound, and a perceiver typically reports to 'hear'/da/, when in fact auditory/ba/was presented combined with visual/ga/(McGurk & MacDonald, 1976). This change in the quality of the sound then evokes a 'McGurk-MMN', despite that the acoustic information remains unchanged. This finding has generally been taken as evidence that lipread information can penetrate auditory cortex and modulate its activity at a very basic level.

In the current study, we examined whether the McGurk-MMN indeed depends on the illusory change of the sound. So far, this has only been shown *indirectly* because the illusion itself has not been manipulated in any direct way. Another issue that has not yet been resolved is by what mechanism the distinction between a speech and non-speech mode of sound processing actually affects audiovisual integration. Tuomainen et al. (2005) speculated that attention may play a key role in the effect. It was argued that in speech mode, attention enhances processing and binding of those features in the stimuli that form a phonetic object, whereas in non-speech mode attention is focused on some other acoustic aspect (e.g., loudness, pitch, duration) that discriminates the stimuli. Whether attention is indeed the crucial factor can be tested using the McGurk-MMN because this brain potential does not require attention to be evoked. Here, we employed SWS while listeners were in speech or non-speech mode. Our 'standard' stimulus was an SWS sound derived from natural auditory/onso/that was combined with the video of a speaker articulating the same syllables/onso/(AnVn). The 'deviant' stimulus contained exactly the same SWS sound/onso/, but now combined with incongruent lipread information of/omso/(AnVm). The incongruent lipread information was expected to change the percept of the SWS sound from/onso/into/omso/if the SWS sound was heard as speech, but this illusory change should be almost completely abolished if the SWS sound is perceived as non-speech (for behavioral evidence, see Eskelund et al., 2011; Tuomainen et al., 2005; Vroomen & Stekelenburg, 2011). This dissociation then allowed us to test, with identical stimuli, whether the illusory change in the sound actually affects the McGurk-MMN. If so, one would expect a McGurk-MMN with SWS sounds for listeners in speech mode, but not for listeners in non-speech mode. For comparative purpose, we included a third group of listeners that heard the original natural recordings of/omso/and/onso/to confirm that the McGurk-MMN would be elicited by the original speech sounds. For all three groups, we also included a visual-only (V-only) and an auditory-only (A-only) condition. The V-only condition served as a control to rule out that the McGurk-MMN was based on the

visual difference between standard and deviant (a visual MMN), and to correct the McGurk-MMN accordingly by subtracting the V-only deviance wave from the AV-wave (Saint-Amour et al., 2007). With the A-only condition, we could test whether an actual auditory change from SWS/onso/into/omso/resulted in a (non-illusory) auditory MMN, and whether that differed for listeners in speech and non-speech mode.

2. Methods

2.1. Participants

Forty-five healthy students (13 males, 32 females, mean age of 21.0 years) with normal hearing and normal or corrected-to-normal vision participated after giving written informed consent (in accordance with the Declaration of Helsinki). They received course credits for their participation. They were equally divided into three between-subjects conditions (i.e., natural speech, SWS speech mode, and SWS non-speech mode). Note that a between-subject design was required because once participants perceive an SWS sound as speech, they cannot switch back to a non-speech mode again.

2.2. Stimuli

The experiment took place in a dimly lit, sound-attenuated, and electrically shielded room. Visual stimuli were presented on a 19-in. monitor positioned at eye-level, 70 cm from the participant's head. Sounds were presented from a central loudspeaker directly below the monitor. The stimuli were identical to the ones used previously by Tuomainen et al. (2005) because they have been shown to produce a strong McGurk-effect. Stimuli were the (Dutch) pseudowords/omso/and/onso/pronounced by a male speaker whose entire face was visible on the screen. The videos were presented at a rate of 25 frames/s, and at an auditory sampling rate of 22.05 kHz. The size of the video frames subtended 14° horizontal and 12° vertical visual angle. Peak intensity of the auditory stimuli was 63 dB(A). The duration of/omso/was 640 ms, and of/onso/it was 600 ms. Sine-wave replicas of both/onso/and/omso/were created in the Praat software (Boersma & Weenink, 1999–2005) with a script by Chris Darwin (<http://www.biols.susx.ac.uk/home/Chris.Darwin/Praatscripts/SWS>). The script creates a three-tone stimulus by positioning time-varying sine waves at the center frequencies of the three lowest formants of the natural speech.

2.3. Procedure

Participants were randomly assigned to either a natural speech group, an 'SWS in speech mode' group, or an 'SWS in non-speech mode' group. Note that the first two groups were intended to perceive the sounds as speech, and the latter as non-speech. Before the start of the actual experiment, participants in the SWS speech mode condition were trained to perceive the SWS stimuli as speech. This was done by alternating the original (natural) audio-recording and the SWS tokens for 15 times before the start of the experiment. Participants in the non-speech mode condition heard the SWS sound equally often, but they were told that it was an artificial sound generated by the computer. Hereafter, participants in the non-speech mode condition were asked to describe the auditory stimuli. One participant described the SWS stimuli as speech-like and was replaced by another participant. For each group, there were three different conditions comprising either A-only, V-only, or AV stimulus presentations. Each condition contained 1020 standards and 180 deviants, administered across 4 blocks per condition. For the A-only and V-only conditions, the standard was the unimodally presented stimulus/onso/(denoted as An and Vn, respectively). The deviant was the unimodally presented/omso/(Am and Vm, respectively). In the AV condition, the standard was auditory/onso/combined with visual/onso/(AnVn), while the deviant was auditory/onso/combined with visual/omso/(AnVm). Trial order was randomized with the restriction that at least two standards preceded each deviant. The inter-stimulus interval was 1200 ms. The A-only, V-only and AV blocks alternated with block-order counter-balanced across participants. To ensure that participants were watching the screen during stimulus presentation they had to detect, by key press, the occasional occurrence of catch trials (5% of total number of trials). Catch trials contained a superimposed small spot between the lips and nose (in the middle of the screen for the A-only condition).

After the EEG experiment, a behavioral control experiment was run to verify that the audiovisual stimuli were indeed perceived as intended. The auditory and visual information in the AV stimuli was either congruent (AmVm and AnVn trials) or incongruent (AmVn and AnVm; 20 trials for each stimulus). In the natural speech and the SWS speech-mode conditions, participants had to label the stimuli on the basis of whether they had heard/onso/or/omso/, while in the SWS non-speech condition the tokens were labeled as '1' or '2' (see Tuomainen et al., 2005). Before the start of this behavioral experiment, participants first learned to distinguish between the two SWS replicas. Training started by alternating the SWS replicas of/omso/and/onso/15 times each. During the presentation of/onso/, the number '1' was shown for the non-speech mode condition, and the word 'onso' for the speech mode condition. During/omso/the number '2' was shown for the non-speech mode condition, and

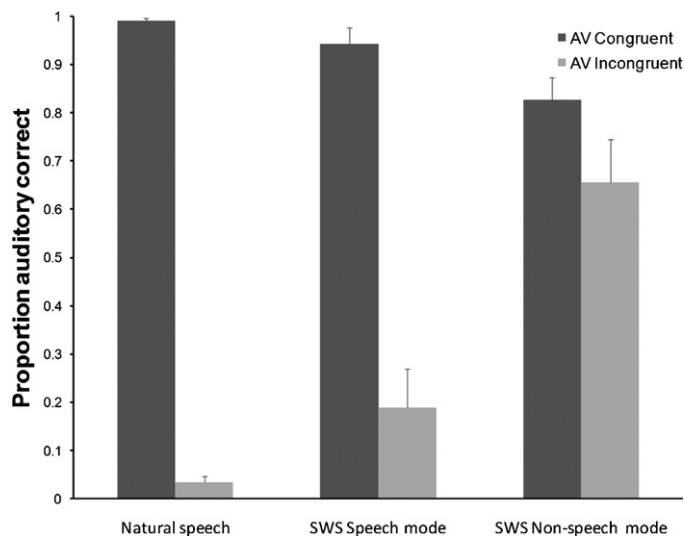


Fig. 1. Results of the behavioral experiment. The bars denote the proportion of correct auditory identification per group (natural speech, sine-wave speech in speech mode, sine-wave speech in non-speech mode) for congruent and incongruent audio-visual presentations. Error bars represent 1 Standard Error of the Mean (SEM).

the word 'omso' for the speech mode condition. Once participants were acquainted with the two sounds, they were further trained to discriminate the two auditory stimuli using two designated buttons. Feedback was given after each trial. If the accuracy in a block of 32 trials was below criterion (80% correct), a second block was run. A short practice session (containing A-only, AV congruent, and AV incongruent trials) preceded the behavioral experiment to familiarize the participants with the experimental task.

2.4. Electrophysiological recording and analysis

The EEG was recorded at a sampling rate of 512 Hz from 128 locations using active Ag–AgCl electrodes (BioSemi, Amsterdam, The Netherlands) mounted in an elastic cap and two mastoid electrodes. Electrodes were positioned radially equidistant from the vertex across the scalp (BioSemi ABC electrode positioning system). Two additional electrodes served as reference (Common Mode Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). Eye movements were monitored by bipolar horizontal and vertical EOG derivations. EEG was referenced offline to an average of left and right mastoids and band-pass filtered (0.1–30 Hz, 24 dB/octave). The raw data were segmented into epochs of 800 ms, including a 100-ms prestimulus baseline. ERPs were time-locked to sound onset. After EOG correction (Gratton, Coles, & Donchin, 1983), epochs with an amplitude change exceeding $\pm 80 \mu\text{V}$ at any EEG channel were rejected. ERPs of the non-catch trials were averaged for standards and deviants, separately for the A-only, V-only and AV blocks. Individual difference waves per modality were computed by subtracting the averaged standard ERP from the averaged deviant ERP. The difference wave in the AV condition may be composed of overlapping components pertaining to the illusory change in the sound as well as the change in mouth movements. To suppress ERP activity evoked by the visual change, the difference waveform of the V-only condition was subtracted from the difference waveform of the AV condition. This AV–V difference wave represents the EEG activity evoked by the illusory change in the sound in its purest form, thus without contribution of the visual component (Saint-Amour et al., 2007; Stekelenburg, Vroomen, & de Gelder, 2004). To track the time-course of the effect of perceptual mode on the McGurk-MMN, we conducted point-by-point *t*-tests between the AV–V difference waves of the speech and non-speech mode. The *t*-tests started at the point where the mouth of the actor of the deviant stimulus (V_m) began to differ from the standard (V_n), which was estimated at ~ 140 ms after stimulus onset. Using a procedure to minimize type I errors (Guthrie & Buchwald, 1991), the difference between the two conditions was considered significant when at least 12 consecutive points (i.e., 32 ms when the signal was resampled at 375 Hz) were significantly different from zero.

3. Results

3.1. Behavioral data

The proportion of correctly identified auditory stimuli was calculated for the congruent and incongruent stimuli. As clearly visible in Fig. 1, and as expected, lipread information changed the

percept of the sound if that sound was perceived as speech, but not if perceived as non-speech. This was confirmed in a MANOVA for repeated measures with as within-subjects variable Congruency (congruent vs. incongruent), and as a between-subject variable Group (SWS non-speech mode, SWS speech mode, natural speech). There were main effects of Congruency, $F(1, 42) = 187.58$, $p < 0.001$ and Group, $F(1, 42) = 10.62$, $p < 0.001$. The proportion of correct responses was higher for congruent than for incongruent stimulus pairings, and it was higher in the SWS non-speech mode condition than in the SWS speech mode and natural speech conditions (p -values < 0.01). Most importantly, there was an interaction between Group and Congruency, $F(2, 42) = 26.55$, $p < 0.001$. Simple-effect tests showed that, compared with congruent lipread information, incongruent lipread information hampered sound identification in the natural speech and the SWS speech mode conditions (because of the McGurk effect, $p < 0.001$), but incongruent lipread information did not affect sound identification in the SWS non-speech condition ($p = 0.1$).

3.2. ERP data

Participants in the AV condition detected 98.9% of the catch trials which did not differ between natural speech, SWS speech mode and SWS non-speech mode groups ($F < 1$). Participants thus apparently complied with instructions and were watching the screen.

3.2.1. A-only MMN

The overall A-only MMN ($-2.35 \mu\text{V}$) was larger than the pre-stimulus baseline, $t(44) = 12.29$, $p < 0.001$. As clearly visible in Fig. 2a, there was no difference between the two SWS conditions (the speech vs. the non-speech group), but the peak of the (acoustically different) natural speech sound was later and somewhat smaller than the two SWS conditions. To test these observations, we ran an ANOVA on the amplitude and latency of the MMN at electrode Cz (where the MMN was maximal) with Group as between-subjects factor (SWS non-speech mode, SWS speech mode, natural speech). For MMN latency, there was main effect of Group, $F(2, 42) = 7.23$, $p < 0.01$. Tukey's post hoc tests (all reported post hoc tests are two-tailed) revealed no differences between the two SWS conditions ($p = 0.26$), but the MMN in the natural speech condition was about 60 ms later compared to the SWS speech mode ($p < 0.01$), and 33 ms later if compared to the SWS non-speech mode condition ($p = 0.08$). For the MMN amplitude, there was no effect of Group, $F(2, 42) = 1.61$, $p = 0.21$.

We also tested for possible differences in the scalp distribution of the MMN by collapsing the 128 electrodes into 18 electrode clusters: prefrontal, frontal, frontal–central, central–parietal, parietal–occipital and occipital for left, central and right side. A MANOVA for repeated measures was conducted with the variables Group (SWS non-speech mode, SWS speech mode, natural speech), Hemisphere (left, middle, right) and Region (6 levels). Condition did not interact with Hemisphere, Region or Hemisphere $\times$ Region ($p > 0.24$), suggesting similar neural generators of the A-only MMN.

3.2.2. V-only MMN

Fig. 2b and c shows the difference waves between standard and deviants (i.e., V_n vs. V_m) for the V-only conditions for occipital and frontal electrodes. The V-only difference wave (a visual MMN) peaked at the midline occipital electrodes after ~ 165 ms if measured from sound onset in the original recording (Fig. 2b), and its amplitude and latency was therefore tested at the electrode Oz. The mean amplitude of the V-only difference wave was $-1.48 \mu\text{V}$, which was significantly different from the pre-stimulus baseline level, $t(44) = 10.56$, $p < 0.001$. For amplitude and latency, though, there was no effect of Group ($F < 1$). The scalp distributions of the V-only difference waves were also similar for the

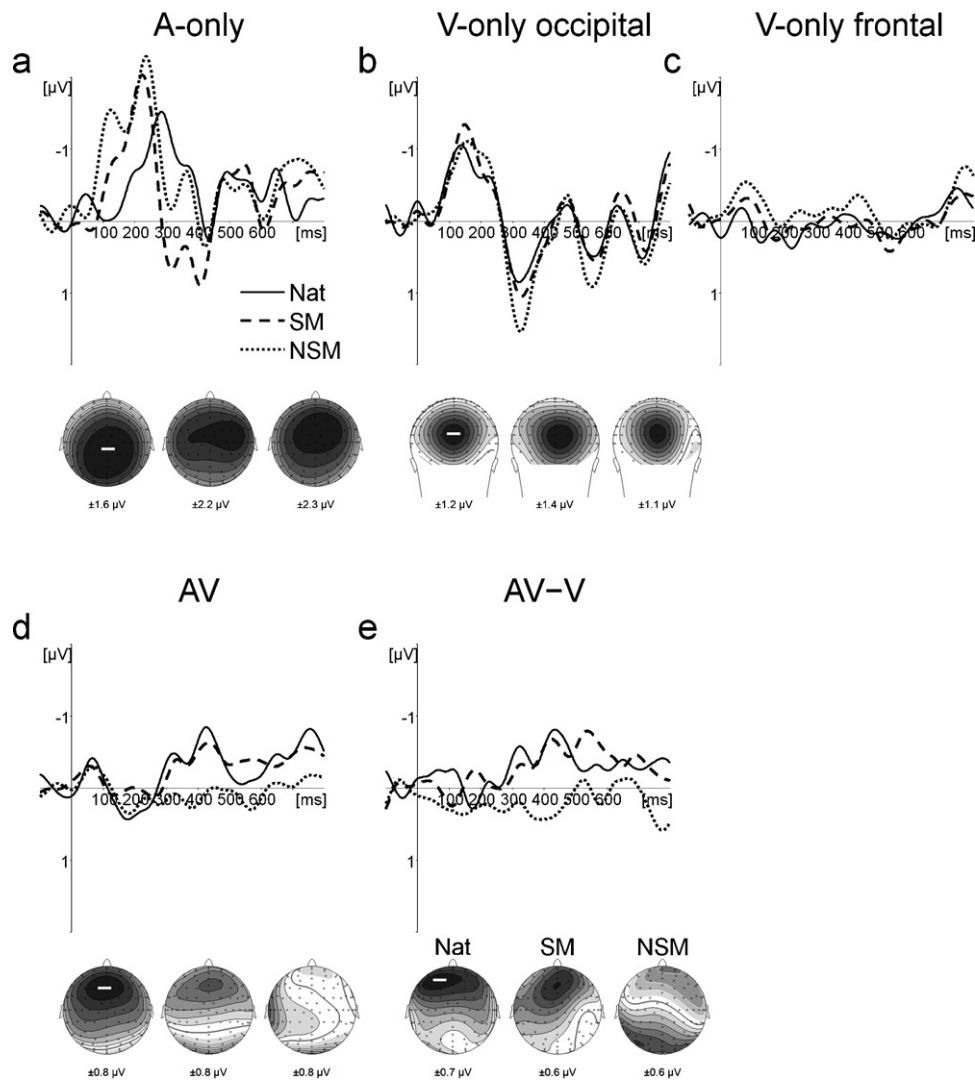


Fig. 2. Grand average ERPs, time-locked to auditory onset, for natural speech (Nat), sine-wave speech in speech mode (SM), and sine-wave speech in non-speech mode (NSM). In panels a, b, c and d the difference waves between deviant and standard are depicted for auditory-only (at a central electrode), visual-only (at an occipital electrode and the same frontal electrode as for the audiovisual blocks) and audiovisual (at a frontal electrode) blocks. Panel e shows the difference waves between the audiovisual and visual-only difference waves at a frontal electrode. Below each panel the scalp topography of the MMN for each mode is displayed. The range of the voltage maps in microvolts is displayed below each map.

three conditions, as the Group $\times$ Hemisphere, Group $\times$ Region, and Group $\times$ Hemisphere $\times$ Region interactions were all non-significant ($F < 1$).

3.2.3. AV MMN

As is clearly visible in Fig. 2d, a reliable AV-MMN was obtained in listeners that were intended to perceive the sounds as speech (i.e., the natural speech and the SWS speech mode conditions), but not in the SWS non-speech mode condition. The ERPs of the natural speech and the SWS speech mode conditions showed a negative deflection between 250 and 500 ms, whereas the difference wave of the SWS non-speech condition hardly deviated from baseline level. Fig. 2d also shows that the ERPs in the critical window actually had two peaks. More careful examination of the individual waves revealed that some participants had two peaks, whereas other participants only had an early peak or a late peak. Peak picking would therefore lead to unreliable and inconsistent results. On basis of visual inspection of the difference wave we therefore computed for all three groups the mean activity in a window of 300–450 ms at an electrode located between Fz and FCz. In this analysis, there was a Group effect, $F(2, 42) = 7.41$, $p < 0.01$. Tukey post hoc

comparisons showed that there was no difference between the natural speech and SWS speech mode condition ($p = 0.65$). Mean activity of the SWS non-speech mode condition, though, was lower than for SWS speech mode ($p < 0.05$) and the natural speech conditions ($p < 0.01$). Testing the difference wave against zero showed that both the SWS speech mode condition, $t(14) = 5.26$, $p < 0.001$, and the natural speech condition, $t(14) = 4.91$, $p < 0.001$, were significantly different from zero, whereas this was not the case for SWS non-speech mode condition, $t(14) = 0.26$, $p = 0.80$.

3.2.4. McGurk-MMN (AV-V)

The AV-V difference wave between standard and deviant (the McGurk-MMN) was tested by comparing the mean activity in the time window of 300–450 ms at Fz between groups. This analysis on the McGurk-MMN led to comparable results as on the raw AV-MMN. There was again a main effect of Group, $F(2, 42) = 5.96$, $p < 0.01$, and Tukey post hoc comparisons showed that there was no significant difference between the natural speech and the SWS speech mode condition ($p = 0.79$), whereas mean activity in the SWS non-speech mode condition was again lower than in the SWS speech mode ($p < 0.05$) and natural speech conditions ($p < 0.01$).

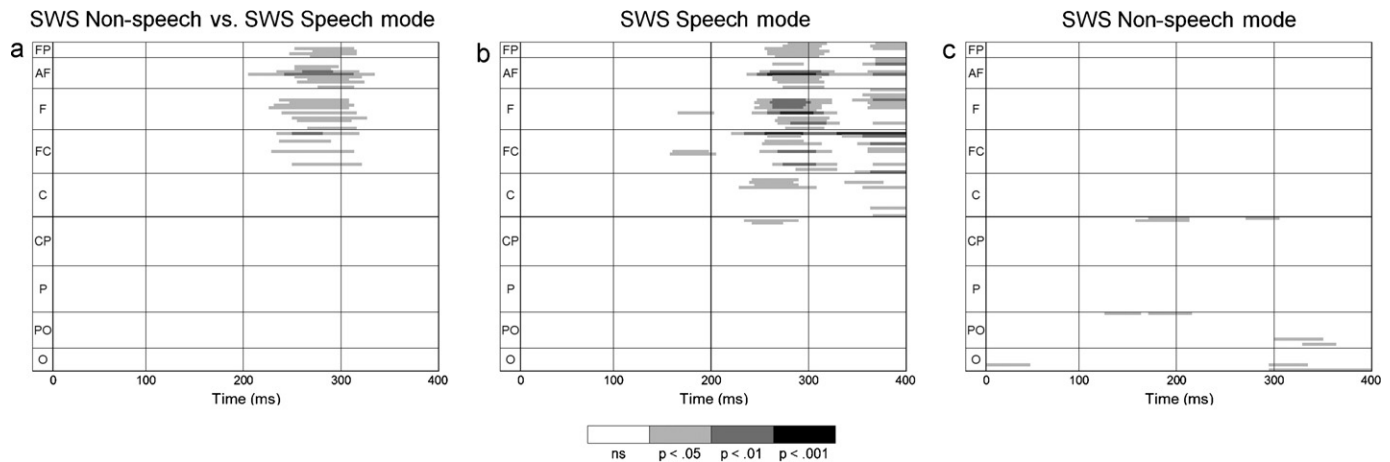


Fig. 3. (a) Time-course of the effect of perceptual mode on the McGurk-MMN using pointwise t -tests at every electrode between the AV–V difference waves of the speech and non-speech mode. (b and c) Pointwise t -test testing the AV–V difference waves against prestimulus baseline level for the speech and non-speech mode conditions, respectively. On the x-axis time starts at 140 ms after sound onset. This is the point where the mouth of the actor of the deviant stimulus (omso) began to differ from the standard (onso). On the y-axis electrode position are clustered into nine scalp regions ranging from prefrontal (FP) to occipital (O). Within each region, electrode laterality (from right to left) is arranged from top to bottom.

Testing the difference wave against zero showed that both the SWS speech mode condition, $t(14) = 2.72, p < 0.05$, and the natural speech condition, $t(14) = 2.83, p < 0.05$, were significantly different from prestimulus baseline level, whereas this was not the case for the SWS non-speech mode condition, $t(14) = 1.41, p = 0.18$. Incongruent lipread information thus evoked a McGurk-MMN in the natural speech condition and in the SWS speech mode condition, but not in the non-speech condition.

Point-wise running t -tests between the AV–V difference waves of the speech and non-speech mode show that the McGurk-MMN differed between both speech modes at the fronto-central electrodes in a 230–320 ms window (Fig. 3a). Subsequent testing of the individual AV–V difference waves against prestimulus baseline level revealed that the deviant stimulus elicited a McGurk-MMN in SWS speech mode, but not in the SWS non-speech mode condition (Fig. 3b and c). In the SWS non-speech mode the AV–V difference wave at the fronto-central electrodes did not differ from prestimulus baseline level.

4. Discussion

Using SWS, we demonstrated that the speech mode of a listener affects the McGurk-MMN. A single auditory token of SWS was presented repeatedly, together with either congruent or incongruent lipread information while listeners were either in speech or non-speech mode. The incongruent lipread information in the AV condition evoked a McGurk-MMN for listeners in speech mode, but not for listeners in non-speech mode. A behavioral experiment further demonstrated that the incongruent lipread information led to an illusory change of the sound only when listeners were in speech mode, but not when listeners were in non-speech mode. Thus only when the SWS stimuli were interpreted as speech, a coherent audiovisual phonetic percept was formed evoking a McGurk effect to incongruent AV stimuli. The McGurk effect on the deviant trials thus modified the auditory percept, and this illusory change triggered the MMN. This also suggests that the auditory representation was modified by the visual input prior to the generation of the MMN. When participants considered the auditory stimuli as non-speech, no McGurk effect was induced, indicating that the visual and acoustic tokens were processed independently. Therefore no MMN was elicited by AV incongruent deviant trials. Taken together, this is rather compelling evidence that it is indeed the illusory change of the sound – induced by audiovisual integration

– that drives the McGurk-MMN. The beauty of this demonstration is that this dissociation in the McGurk-MMN could be evoked with stimuli that were acoustically and visually *identical* for listeners in speech and non-speech mode. This therefore effectively rules out any kind of (low level) stimulus confound that may arise when comparing natural speech to non-speech stimuli.

Still, one must consider alternative accounts of the effect of speech mode on the McGurk-MMN. One might argue, for example, that the McGurk-MMN in non-speech mode did not emerge because participants were simply not watching the visual stimuli. This would then reduce the visual effect on the auditory percept and would render the deviant to be perceptually identical to the standard stimulus. There is, however, no reason to assume that participants did not watch the visual stimuli in the non-speech mode condition because performance on the secondary visual task (detecting a short-duration small spot) was virtually flawless and did not deviate between the groups. A second concern may be that, despite the fact that participants in the non-speech mode condition watched the video equally well as those in the speech mode condition, the difference in McGurk-MMN between speech mode and non-speech mode was caused solely by differences in visual processing and not by differences in audiovisual integration per se. That is, lip movements in speech mode may have been more ‘meaningful’ or attract more attention than in non-speech mode with the result that the visual deviance triggered a stronger visual mismatch process. The modulation of the McGurk-MMN by perceptual mode would then mainly reflect an effect of meaning or attention devoted to the visual stimuli. Alternatively, because of the manipulation of perceptual mode participants may have been lipreading in speech mode but not in the non-speech mode. This would imply that the effect is entirely visually driven. To check this, we analyzed the mean activity in the time window of 300–450 ms for the V-only conditions at the same electrode as for the AV condition (Fig. 2c). In this analysis, the three groups did not differ from each other ($F < 1$), and the difference waves at these electrode positions did not exceed the prestimulus baseline level (all $p > 0.12$). There is thus no evidence for these alternative accounts because the relevant neural markers of visual processing as indexed by the difference wave in the V-only blocks did not differ between speech and non-speech conditions.

The results of the current experiment provide insight into the mechanism that underlies the mediating effect of perceptual mode on audiovisual integration. It was suggested that depending on

the perceptual mode attention guides perceivers either to phonetic or non-phonetic features in both auditory and visual stimuli (Tuomainen et al., 2005). Accordingly, the McGurk effect to SWS is evoked because speech mode directs attention to features of the stimulus relevant for extracting phonetic content (Eskelund et al., 2011; Tuomainen et al., 2005). This attentional account, though, is difficult to reconcile with the current MMN data. In the visual detection task we used here, the auditory aspects of the stimuli were completely task-irrelevant. In addition, the generation of the MMN does not require attention to the auditory stimuli, and for the McGurk effect, visual fixation is not that important, provided that the viewer can see the face of the speaker (Pare, Richler, Ten Hove, & Munhall, 2003). Our data therefore rather suggest that the 'top-down effect' of the auditory interpretation of the sound takes place automatically, affecting intersensory, and thereby also acoustic stages of stimulus processing.

Another theoretically interesting question is whether this top-down effect is specific for speech, or whether it occurs with other learned associations as well. To the best of our knowledge this has not been tested. One could create ambiguous sounds, though, that remain ambiguous in a 'naive' mode, but receive meaning (and are labeled as belonging to either of the two sounds) in an 'informed' mode. As an example, Saldaña and Rosenblum (1993) created 'pluck' and 'bow' sounds of a cello, and showed that the visual information of a pluck (or bow) on a cello biased the interpretation of that sound toward the visual information. The critical question would be whether this same bias would occur if listeners were unaware that the sounds were derived from a musical instrument and thus belonged to the cello. One caveat for future research is that it will be difficult to find non-speech sounds that can sufficiently be biased by visual information. With the stimuli used by Saldaña and Rosenblum (1993), we expect the effect on an AV-MMN to be weak or non-existent because even in the original study, where listeners thus were "informed", there was only a very small behavioral effect (Saldaña & Rosenblum, 1993), while others simply failed to obtain a visual bias effect on non-speech sounds (Kroos & Hogan, 2009).

Another relevant question is to which extent the stimuli of the present study would yield similar results as the one of the original McGurk effect. In a strict sense, one can argue that the only 'real' McGurk effect is a fusion in which a new percept emerges that was not present in the auditory and visual channels in isolation (e.g., auditory/b/ + visual/g/fuses into/d/). In the present study, there is no fusion into a new percept, but a bias (visual/m/biases auditory/n/towards/m/). Is there a critical difference between these two examples? We would argue that there is not. Crucially, the McGurk effect is not solely an exotic laboratory phenomenon that can be observed with unnatural dubbings of audio and video channels, but rather an elegant example of a much more widespread phenomenon, namely that auditory and lipread speech are, within limits, integrated at a profound level. From that perspective, there is no reason to believe that integration of (/b/ + /g/) is fundamentally different from (/m/ + /n/): both lead to an optimal solution in which the phonetic percept is different from the acoustic information.

Because the McGurk illusion allows the elicitation of an MMN in the absence of any acoustic change, it has been used as a paradigm to dissociate acoustic from phonetic processing. Previous MMN studies have suggested the existence of a phonetic-specific neural trace that is distinct from an acoustic change-detection processes (e.g., Dehaene-Lambertz, 1997; Näätänen et al., 1997; Winkler, 1999). While the acoustic MMN is thought to rely on both hemispheres, the phonetic MMN appears to occur predominantly on the left side (Alho et al., 1998; Näätänen et al., 1997). This is in line with fMRI data showing that SWS induces stronger activity in the left superior temporal sulcus for listeners in speech mode than for listeners in non-speech mode (Möttönen et al., 2006). Somewhat to

our surprise, though, we found no support for a distinction between a phonetic and acoustic MMN. That is, the comparison of the A-only difference wave between speech and non-speech mode revealed no difference in MMN amplitude and topography, which suggests that the perceptual mode did not affect auditory mismatch detection.¹ How then can expectation about the origin of auditory stimuli modulate the McGurk-MMN? The existence of an MMN to the McGurk illusion indicates that auditory sensory memory is modified by visual input, presumably by audiovisual interactions that occur prior to the generation of the auditory mismatch signal (Colin et al., 2002; Kislyuk et al., 2008; Saint-Amour et al., 2007; Sams et al., 1991). Several studies have indeed shown early integration of auditory and visual speech signals (usually starting around 80–100 ms), followed by integration that depended on the AV congruency between the sound and the lipread information (Arnal et al., 2009; Besle et al., 2004; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). Here, we argue that the effect of expectation about the origin of auditory stimuli on the McGurk-MMN is the consequence of altered audiovisual interactions that give rise to either change or no change in the auditory percept depending on perceptual mode.

The current data together with other data on AV speech perception indicate that AV speech integration is a multistage process. AV speech may be integrated by speech-specific and more general multisensory integrative mechanisms. One piece of evidence that distinct processes underlie integration of AV speech comes from a study of Eskelund et al. (2011). As in Tuomainen et al. (2005), these authors reported a McGurk effect for SWS stimuli only when participants were in speech mode. By contrast, for the same participants, the improvement of auditory detection by the talkers' face was equally large for speech and non-speech mode. This is in line with other behavioral studies showing that the audiovisual detection advantage is not speech-specific (Bernstein, Auer, & Takayanagi, 2004; Schwartz, Berthommier, & Savariaux, 2004). According to Schwartz et al. (2004) this could be explained by the notion that increased auditory detection and speech intelligibility may result from enhanced auditory signal-to-noise ratio by audiovisual covariation, which is identical for speech and non-speech modes. The identification of phonetic content and auditory detection in noise can thus be affected by speech-specific and non-specific audiovisual integration effects.

The notion of multiple stages of AV speech processing in which different attributes of the AV speech signal are integrated by different integration processes is also corroborated by electrophysiological studies (Arnal et al., 2009; Klucharev, Möttönen, & Sams, 2003; Stekelenburg & Vroomen, 2007). Arnal et al. (2009) conjectured that predictive visual information (lip movements) that naturally precedes the actual utterance, affects auditory perception via a fast direct visual to auditory pathway which conveys physical visual but no phonological characteristics. After the visual-to-auditory predictive mechanism a secondary feedback signal is followed via superior temporal sulcus (STS), which signals the error (if present) between visual prediction and auditory input. As the McGurk effect acts on the phonetic level, the current data show that knowledge about the origin of the auditory data affects the AV integration at a phonetic level. This suggests that the slow indirect route via STS is implicated in the auditory expectation effect on AV integration. A recent fMRI study (Lee & Noppeney, 2011) on

¹ Fig. 2a shows that the A-only MMN for the natural speech condition is somewhat smaller and later than the MMN for both SWS conditions. It is unlikely that this reflects different neural mechanisms underlying the generation of the MMN because the scalp distribution did not differ between conditions. The most plausible explanation is that the perceptual difference between standard and deviant was larger for SWS than for natural speech, which typically reduces MMN latency and increase MMN amplitude (Sams, Paavilainen, Alho, & Näätänen, 1985).

audiovisual SWS speech perception specified that the left anterior mid-STS depended on higher-order linguistic information whereas bilateral posterior and left mid STS integrated audiovisual inputs on the basis of physical factors.

To conclude, expectation about the nature of auditory stimuli strongly affects audiovisual speech integration on both the behavioral and the neural level. Only when SWS tokens were interpreted as speech, lipread information influenced the perception of auditory stimuli as shown in the McGurk effect and the McGurk-MMN. These findings support the existence of an audiovisual speech-specific mode of perception. Given the pre-attentive and perceptual nature of the MMN, we conclude that the effect of perceptual mode on audiovisual integration has consequences at early processing stages.

References

- Alho, K., Connolly, J. F., Cheour, M., Lehtokoski, A., Huottilainen, M., Virtanen, J., et al. (1998). Hemispheric lateralization in preattentive processing of speech sounds. *Neuroscience Letters*, 258(1), 9–12.
- Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, 29(43), 13445–13453.
- Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44, 5–18.
- Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20(8), 2225–2234.
- Boersma, P., & Weenink, K. (1999–2005). Praat: Doing phonetics by computer. Retrieved from: <http://www.fon.hum.uva.nl/praat/>
- Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, 14(17), 2213–2218.
- Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, 16(5), 805–816.
- Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, 10(12), 2619–2623.
- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10(11), 649–657.
- Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(4), 495–506.
- Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, 8(4), 919–924.
- Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research*, 208(3), 447–457.
- Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3), 1730–1741.
- Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography & Clinical Neurophysiology*, 55(4), 468–484.
- Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, 28(2), 240–244.
- Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience*, 31(5), 1715–1720.
- Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience*, 20, 1–10.
- Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research, Cognitive Brain Research*, 18(1), 65–75.
- Kroos, C., & Hogan, J. (2009). Visual influence on auditory perception: Is speech special? In B. J. Theobald, & R. Harvey (Eds.), *Proceedings of the international conference on auditory-visual speech processing 2009* Norwich, UK.
- Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech-perception in adults and infants using nonspeech auditory-stimuli. *Journal of Experimental Psychology-Human Perception and Performance*, 17(3), 829–840.
- Lee, H., & Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *Journal of Neuroscience*, 31(31), 11338–11350.
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
- Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage*, 30(2), 563–569.
- Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect in evoked potential reinterpreted. *Acta Psychologica*, 42, 313–329.
- Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huottilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615), 432–434.
- Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12), 2544–2590.
- Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage*, 25(2), 333–338.
- Pare, M., Richler, R. C., Ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audio-visual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, 65(4), 553–567.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947–949.
- Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1), 81–110.
- 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(3), 587–597.
- Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics*, 54(3), 406–416.
- Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Louasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, 127(1), 141–145.
- Sams, M., Paavilainen, P., Alho, K., & Näätänen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalography & Clinical Neurophysiology*, 62(6), 437–448.
- Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2), B69–B78.
- Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage*, 25(1), 76–89.
- Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19(12), 1964–1973.
- Stekelenburg, J. J., Vroomen, J., & de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neuroscience Letters*, 357(3), 163–166.
- Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, 96(1), B13–B22.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Science of the USA*, 102, 1181–1186.
- von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, 4(10), 1809–1820.
- Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition*, 110, 254–259.
- Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, 118(1), 75–83.
- Winkler, I. (1999). Auditory and phonetic representations for isolated vowels: Cross-language studies. *Psychophysiology*, 36, S2.