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

KASPER ESKELUND[1,2,*], LAURA FRØLICH[1], AND 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*

Visual speech plays a central role in general speech perception. Through audiovisual integration, visual speech may facilitate auditory detection and identification for people with normal hearing in noisy conditions. Further, a visual syllable may alter the auditory phonetic percept, as can be seen 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 audio-visual speech – relies on information from the talking face. However, visual speech encoding and face perception are known to be functionally separate. 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 functional dependency between 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 face perception and stimuli that do not, 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 (Grant and Seitz, 2000; Eskelund *et al.*, 2010). Visual speech can also facilitate speech comprehension (Sumby and Pollack, 1954). When visual and auditory speech are phonetically incongruent, an illusory alteration of the phoneme perceived in the voice can occur. This is known as the McGurk illusion (McGurk and MacDonald, 1976).

Perception of natural audiovisual speech evidently relies on visual perception of the talking face. Here, the visual 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.

*Corresponding author: kaes@dtu.dk

Proceedings of ISAAR 2013: Auditory Plasticity – Listening with the Brain. 4th symposium on Auditory and Audiological Research. August 2013, Nyborg, Denmark. Edited by T. Dau, S. Santurette, J. C. Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. ISBN: 978-87-990013-4-7. The Danavox Jubilee Foundation, 2014.

Interestingly, the functions of visual speech and face perception have been suggested to be separate cognitive modules (Bruce and Young, 1986). But does 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 different 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, holistic or configurational mismatch is only perceived in the upright facial context. Thus, facial configuration is only perceived for stimuli with upright facial context (UF).

To investigate the influence of facial configuration on audiovisual speech perception, Rosenblum and colleagues (2000) used video stimuli based on the Thatcher illusion. 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 *et al.* 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 in speech perception on holistic 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 testing the influence of configurational face information by electrophysiological means, using the mismatch negativity (MMN) paradigm developed by Näätänen *et al.* (1978). MMN is a component in the auditory event-related potential (ERP), generated by presenting a sequence of identical standard auditory stimuli at a constant inter-trial interval. At random places in the sequence, usually in 9-15% of stimulus presentations, the stimulus is altered (deviant trials). The alteration must be noticeable and can be in, 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. This has been hypothesized to be due to a memory process comparing each incoming stimulus with the established trace of standard stimuli (Näätänen, 2003). Whenever a deviant stimulus occurs, a differential neural response is evoked.

Sams and colleagues (1991) showed that the McGurk illusion can elicit MMN without any acoustic difference between standard and deviant stimuli. In this McGurk MMN

paradigm 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/. Thus, only the visual phoneme is altered in deviant trials.

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. The 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-second break where subjects were instructed to relax. In total, 1100 stimuli were presented in each condition, of which 165 were deviants. The duration of each EEG recording 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 were bandpass-filtered between 1 and 30 Hz 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. 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 MMN paradigm of the experiment relies on multiple perceptual and neuro-physiological effects. These are well-known effects, but do not occur in all members of a given population. The prevalence of acoustic MMN is high, but not universal. 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 displayed 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 vertex electrode (Cz) are shown in Fig. 1. For the UF-UM condition presented in Fig. 1, the standard and deviant ERPs follow the same pattern until approx. 200 ms post stimulus, where a negative deflection of the deviant ERP starts.

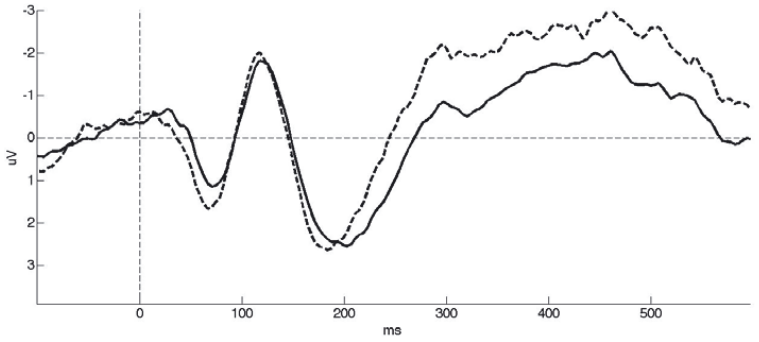

Fig. 1: Average ERPs recorded from UF-UM stimuli at electrode Cz. Auditory onset at 0 ms. Full line represents ERPs due to standard stimuli. Dashed line represents ERPs due to deviant stimuli.

Interestingly, ERPs from the UF-IM condition displayed in Fig. 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.

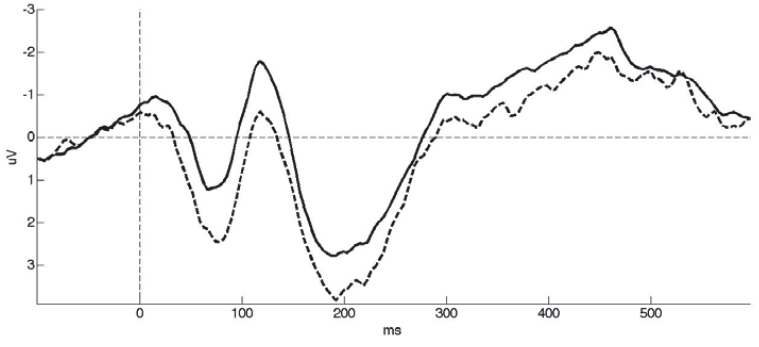

Fig. 2: Average ERPs recorded from UF-IM stimuli at electrode Cz. Auditory onset at 0 ms. Full line represents ERPs due to standard stimuli. Dashed line represents ERPs due to deviant stimuli.

In the UF-UM condition, a mismatch negativity pattern is easily seen in the difference between deviant and standard ERPs. As is evident in Fig. 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. The maximal *t*-score was −10.8 at 290 ms.

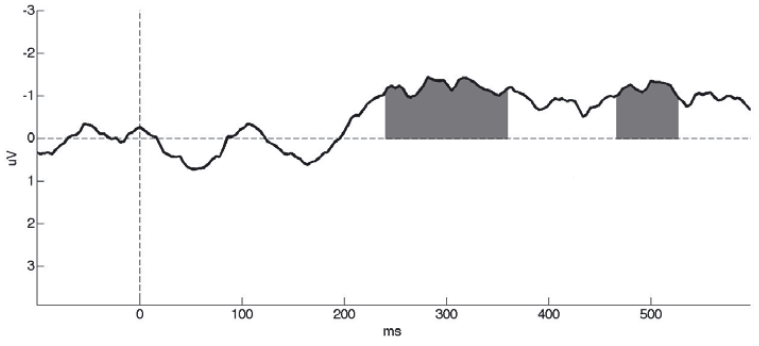

Fig. 3: Difference wave representing the difference between deviants and standards in the UF-UM condition at electrode Cz. Auditory onset at 0 ms. Shaded area marks statistically significant portions of the difference wave (exceeding the critical *t*-score of +/−4.31).

As can be seen in Fig. 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 Fig. 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).

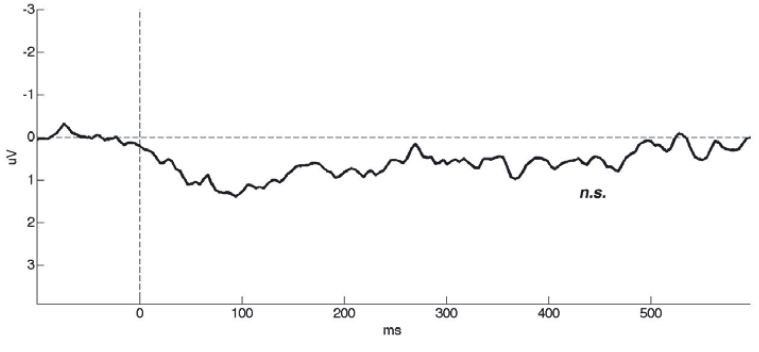

Fig. 4: Difference waves representing the difference between standards and deviants in the UF-IM condition at electrode Cz. Auditory onset at 0 ms.

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).

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.

	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)

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.

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 were 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 was not significant ($F(1,7) = 2.3$, $p > 0.1$). 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 *et al.* (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 large MMN 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,

we conclude that facial configuration had a significant impact on MMN generated by audiovisual integration.

It is worth noting, that subjects were pre-selected for analysis on the 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 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.

Bruce, V., and Young, A. (**1986**). "Understanding face recognition," Br. J. Psychol., **77**, 305-327.

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. Meth., **134**, 9-21.

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

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**, 1197-1208.

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

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

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

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.

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.

Sumby, W.H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am., **26**, 212-215.

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