doi:10.1093/scan/nsu105 SCAN (2015) 10, 713–720

The role of emotion in dynamic audiovisual integration of faces and voices

Jenny Kokinous,[1] Sonja A. Kotz,[2,3] Alessandro Tavano,[1] and Erich Schröger[1]
[1]Institute of Psychology, University of Leipzig, 04109 Leipzig, Germany, [2]Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neuropsychology, 04103 Leipzig, Germany, and [3]School of Psychological Sciences, University of Manchester, Manchester, UK

We used human electroencephalogram to study early audiovisual integration of dynamic angry and neutral expressions. An auditory-only condition served as a baseline for the interpretation of integration effects. In the audiovisual conditions, the validity of visual information was manipulated using facial expressions that were either emotionally congruent or incongruent with the vocal expressions. First, we report an N1 suppression effect for angry compared with neutral vocalizations in the auditory-only condition. Second, we confirm early integration of congruent visual and auditory information as indexed by a suppression of the auditory N1 and P2 components in the audiovisual compared with the auditory-only condition. Third, audiovisual N1 suppression was modulated by audiovisual congruency in interaction with emotion: for neutral vocalizations, there was N1 suppression in both the congruent and the incongruent audiovisual conditions. For angry vocalizations, there was N1 suppression only in the congruent but not in the incongruent condition. Extending previous findings of dynamic audiovisual integration, the current results suggest that audiovisual N1 suppression is congruency- and emotion-specific and indicate that dynamic emotional expressions compared with non-emotional expressions are preferentially processed in early audiovisual integration.

Keywords: emotion; audiovisual; incongruity; cross-modal prediction; EEG

INTRODUCTION

Human social communication is multimodal by nature and involves combining emotional cues such as a speaker's facial and vocal expression. Audiovisual integration of emotion expressions has been subject to several studies using a wide range of experimental paradigms but there is only a small number of event-related potential (ERP) studies investigating audiovisual interaction effects in the auditory domain (e.g. de Gelder *et al.*, 1999; Pourtois *et al.*, 2000, 2002; Paulmann *et al.*, 2009; Jessen and Kotz, 2011; Jessen *et al.*, 2012). The majority of work has focused on static visual expressions (face images) (e.g. de Gelder *et al.*, 1999; Pourtois *et al.*, 2000, 2002; Paulmann *et al.*, 2009) with relatively low ecological validity. In real-life situations, however, dynamic visual information such as facial movements contributes to facial affect recognition (Bassili, 1979; Ambadar *et al.*, 2005). Nevertheless, the role of facial dynamics in emotion perception has largely been ignored (LaBar *et al.*, 2003). Dynamic information may also modulate audiovisual interactions (Bertelson and de Gelder, 2004), for example, multisensory integration effects are much stronger for dynamic faces (Ghazanfar *et al.*, 2005; Campanella and Belin, 2007). The present study used the electroencephalogram (EEG) to investigate how the processing of vocal emotion expressions is influenced by ecologically valid, dynamic facial expressions at the early stages of audiovisual integration.

The sensory input of one modality can alter and facilitate perception in another modality (Calvert *et al.*, 1997; de Gelder and Vroomen, 2000; Ethofer *et al.*, 2006), particularly when stimulus dynamics cause one modality to lag the other. Facial movements naturally precede vocal expressions in time (Chandrasekaran *et al.*, 2009) and as a consequence, audiovisual amplitude suppression of the brain responses occurring ~100–200 ms after auditory stimulus onset (N1 and P2 component of the ERP) compared with unisensory processing has been reported for audiovisual speech, an indicator for cross-modal facilitation and audiovisual integration (Klucharev *et al.*, 2003; van Wassenhove *et al.*, 2005; Knowland *et al.*, 2014). This effect has also been shown for audiovisual emotion expressions (Jessen and Kotz, 2011, 2013; Jessen *et al.*, 2012) and other dynamic human actions with sensory consequences such as clapping hands (Stekelenburg and Vroomen, 2007, 2012). Some studies have reported suppression of only the N1 (e.g. Besle *et al.*, 2004; Jessen and Kotz, 2011) or the P2 (Baart *et al.*, 2014), suggesting that audiovisual interactions at N1 and P2 can be dissociated.

Predictive processes have been proposed as the underlying mechanism for cross-modal auditory response suppression (Besle *et al.*, 2004; van Wassenhove *et al.*, 2005; Stekelenburg and Vroomen, 2007, 2012; Arnal *et al.*, 2009; Vroomen and Stekelenburg, 2010). During the perception of dynamic audiovisual stimuli, a visual signal can predict several aspects of a subsequent sound, such as the time of its onset (temporal prediction), its specific features and informational content (formal prediction) or its spatial location (spatial prediction) (see also e.g. Stekelenburg and Vroomen, 2007 for the terminology). Studies manipulating the validity of formal predictions by implementing audiovisual informational incongruity, using for example phonetically congruent and incongruent visual and auditory syllables, have found that N1-P2 suppression (Klucharev *et al.*, 2003; van Wassenhove *et al.*, 2005) or particularly N1 suppression (Stekelenburg and Vroomen, 2007) is insensitive to audiovisual semantic incongruity. However, the auditory N1 has been shown to be modulated by temporal and spatial predictability of the auditory input (Vroomen and Stekelenburg, 2010; Stekelenburg and Vroomen, 2012). The described effects relate to non-emotional processing, and the cross-modal 'predictive coding hypothesis' has rarely been addressed in studies on audiovisual emotion perception. Emotion signals may provide a salient context to modulate visual-to-auditory predictions and audiovisual integration. Jessen and Kotz (2013) propose that it is essential to consider cross-modal predictions to fully understand multisensory emotion perception and that emotional visual information may even allow more reliable predicting of auditory information. Thus, the aim of this

Received 1 November 2013; Revised 9 July 2014; Accepted 13 August 2014
Advance Access publication 20 August 2014

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the research training group 1182 'Function of Attention in Cognition' (scholarship to J.K.). The authors wish to thank Alexandra Bendixen for providing the scripts for the jackknife procedure.

Correspondence should be addressed to Jenny Kokinous, Cognitive including Biological Psychology, Institute of Psychology, University of Leipzig, Neumarkt 9–19, 04109 Leipzig, Germany. E-mail: kokinous@uni-leipzig.de.

study was to investigate whether audiovisual suppression effects are influenced by the emotional validity of preceding visual information and whether emotions are preferentially processed in early audiovisual integration.

We studied early audiovisual integration of neutral and angry dynamic face–voice pairs. Pursuant to previous studies (Ho *et al.*, 2014), we tested for the presence of early auditory emotion effects in an auditory-only condition, which was also crucial to measure audiovisual integration as indicated by audiovisual suppression effects. Further, we investigated how the processing of vocal expressions is differentially influenced by a predictive or a non-predictive visual context. To manipulate the validity of formal predictions, we used dynamic visual stimuli that were either emotionally congruent or incongruent with the auditory stimuli. Based on previous findings, we expected to find facilitation of auditory processing by the presence of congruent visual information reflected in an audiovisual N1 and P2 amplitude decrease. Owing to the evolutionary significance of emotion signals, we hypothesized that audiovisual N1 suppression would be modulated by audiovisual emotional (in)congruity. This would be in contrast to previous studies showing global congruency-unspecific audiovisual amplitude reductions in the N1-P2 complex and to studies showing congruency-specific audiovisual suppression for the P2, but not for the N1. Thus, the findings would indicate an advantage of emotion signals in audiovisual integration.

METHODS

Participants

Twenty healthy volunteers participated in the experiment of which three had to be excluded from further analyses due to excessive artifacts. The remaining sample consisted of 17 participants (10 female) with a mean age of 25.6 years (s.d. = 4.6 years). All participants had normal or corrected-to-normal vision and did not report hearing impairments. Participants gave written informed consent after the experimental procedure had been explained to them. They received course credit or monetary reimbursement for participating in the study. Exclusion criteria describe any history of brain injury, neurological disorder (e.g. stroke, epilepsy), any current treatment for mental illness or the intake of medication affecting the central nervous system. The experimental protocol adhered to the Declaration of Helsinki and the ethics guidelines of the German Association of Psychology (ethics board of the Deutsche Gesellschaft für Psychologie).

Stimulus material and design

The stimulus material had previously been developed and validated at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany, for research on multimodal affective processing (Ho *et al.*, 2014). Stimuli consisted of a series of non-linguistic interjections (/ah/,/oh/) uttered by a 24-year-old lay actress and expressing anger or no emotion (neutral). In the auditory-only condition, interjections were presented while participants viewed a black fixation cross on a gray computer screen. In the audiovisual conditions, the utterances were accompanied by congruent or incongruent face videos of the speaker. In all conditions, the delay between the onset of the visual and the onset of the auditory stimulus was variable owing to a natural jitter in the individual recordings (mean = 765 ms). Fifteen separately recorded videos per condition were selected for the EEG experiment to preserve the natural variability of the emotion expressions. Videos were in MPEG-1 format. The actress was instructed to begin each emotion expression with a neutral face to ensure that the emotion evolved naturally in time. Incongruent audiovisual stimuli were created artificially by overlaying the videos with a separately recorded sound of a mismatching emotion using the original sound onset for alignment. Thus, synchrony differences between emotionally congruent and incongruent stimuli were minimized. The sound in all videos was root mean square normalized, matching the mean intensity of neutral and angry interjections. No other modifications of the auditory material were performed to not distort the natural characteristics of the stimuli.

Several valence and arousal ratings (Bradley and Lang, 1994) and an emotion categorization study had been performed on the stimulus material before the present experiment (Ho *et al.*, 2014). In the rating study, 32 participants (16 female) were asked to rate the congruent and incongruent videos in terms of valence and arousal using a two-dimensional valence and arousal rating space (Schubert, 1999) with manikins taken from Bradley and Lang (1994), which represented the extreme ends of the valence and arousal scales. The ratings were subsequently converted to the 9-point SAM scale and confirmed that angry congruent face–voice combinations were rated as more negative and more arousing than neutral congruent face–voice pairs. Incongruent audiovisual combinations of an angry face paired with a neutral voice were rated as more negative and higher in arousal than a congruent neutral combination. On the other hand, when a neutral face was combined with an angry voice, overall valence was less negative and arousal lower than for congruent angry combinations. Finally, incongruent angry faces paired with neutral voices were rated as more negative but less arousing than an incongruent neutral face combined with an angry voice. To ensure that an expressed emotion is accurately recognized in the congruent stimuli, 40 additional participants (20 female) had been asked to classify the emotion expressed only in the face, only in the voice or in a multimodal condition. Six basic emotions had been tested (Ekman and Friesen, 1976: anger, happiness, sadness, fear, disgust and a neutral). Performance in the emotion categorization task, measured as unbiased hit rates (H_u; Wagner, 1993), showed reliable identification of anger and neutral expressions in all conditions but most accurate performance in the audiovisual condition ($H_u > 0.95$).

Procedure

Participants sat comfortably in a sound-attenuated, electrically shielded and dimly lit chamber looking at a computer screen placed ~120 cm in front of them and holding a response device (Microsoft SideWinder Plug & Play Game Pad). Duration of the EEG session was ~60 min including breaks. Each of the six experimental conditions (auditory-only neutral, auditory-only angry, audiovisual congruent neutral, audiovisual congruent angry, audiovisual incongruent neutral, audiovisual incongruent angry) comprised 128 trials, adding up to 768 trials. Trials were segmented into 12 blocks (four auditory, eight audiovisual) of 3.7 min length and were presented pseudo-randomized in a mixed design with a constant proportion of trials of each condition in each block. Based on a previous study showing that the effects of audiovisual incongruity on N1 and P2 amplitude are most robust in attend-voice situations (Ho *et al.*, 2014), participants engaged in a two-alternative forced choice auditory task, that is, they judged the voice conveyed emotion ('Was the voice angry or not?'). Sounds were presented binaurally via headphones (Sennheiser HD 25-1) at the same loudness level for all participants. A trial consisted of the presentation of a black fixation cross on a gray screen for 750 ms, followed by stimulus presentation for 1000–2250 ms depending on the individual stimulus length (voice only or audiovisual face–voice pair) and the subject's response. For the latter, after each stimulus, a response screen appeared showing the button assignment ('wütend' ['angry'], 'nicht wütend' ['not angry'] displayed on the left and the right side of the screen, corresponding to a left and a right button on the response device) with a question mark in the center. Participants were requested to respond as quickly and as accurate as

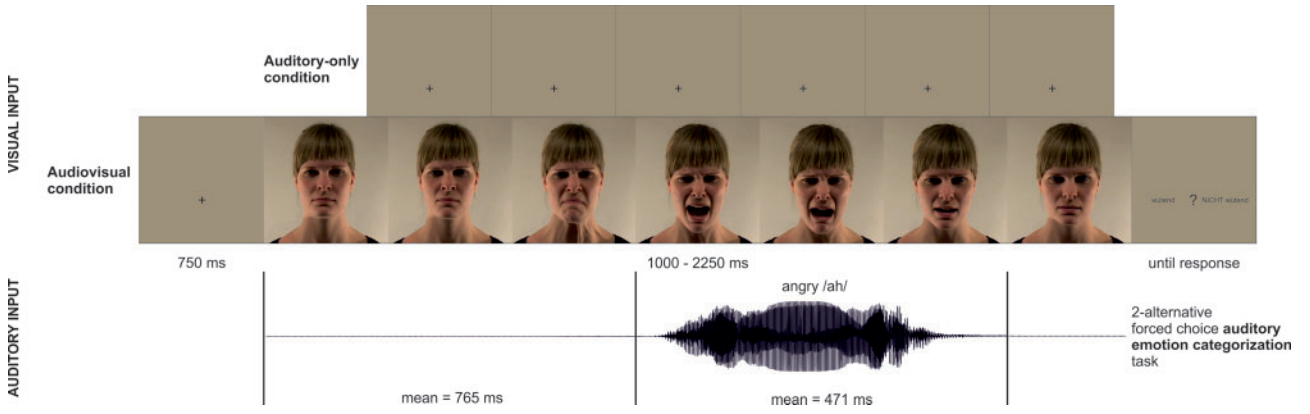

Fig. 1 Illustration of a trial in the auditory *vs* the audiovisual conditions, including a schematic depiction of an example video with congruent sound. The voice was either emotionally congruent or incongruent with the video.

possible. The button presses were done with the thumbs of both hands. The order of experimental parts (auditory, audiovisual) and the response button assignment (angry left, angry right) was counterbalanced across participants. Before the start of the actual experiment, subjects performed short training blocks (~ 2 min, 32 trials). The experiment was implemented using the Presentation software Version 15.0 (Neurobehavioral Systems, Inc.). Figure 1 shows an example trial including a schematic depiction of a video with accompanying sound.

The EEG was recorded from 58 active Ag/AgCl electrodes mounted in an elastic cap (actiCAP; Brain Products GmbH, Munich, Germany) according to the international extended 10–20 system using BrainVision Recorder software (Brain Products). Additional electrodes were placed at the left and the right mastoid as well as on the outer canthi of both eyes and above and below the right eye for the electrooculogram (HEOG and VEOG, respectively). The ground electrode was placed on a participant's forehead. During recording, the signal was commonly referenced to a nose electrode, amplified by BrainVision Professional BrainAmp DC amplifiers and digitized at a sample rate of 500 Hz.

In the EEG experiment, behavioral responses had been delayed to ensure that participants perceived the full length of the stimulus before making a decision and to avoid motor-response–related activity in the ongoing EEG signal. To collect accurate reaction times (RTs), an additional behavioral experiment was conducted subsequently to the EEG using a different participant sample (N = 10, 6 female, mean age = 24.6 years, s.d. = 2.5 years).

DATA ANALYSIS

Behavioral data

RTs (in milliseconds) as well as hit rates (in %) in the two-alternative auditory emotion classification task were computed for each participant of the follow-up behavioral experiment. To test how behavioral performance is influenced by preceding emotionally congruent or incongruent visual information, we computed repeated-measures analyses of variance (ANOVAs) with the factors voice conveyed emotion (neutral, angry) and visual context (auditory-only, audiovisual emotionally congruent, audiovisual emotionally incongruent).

ERP data

EEG data processing was performed using the EEGLAB 9.0.3.4 b toolbox (Delorme and Makeig, 2004) implemented in Matlab (Mathworks, Natick, MA). The data were filtered with a 0.5–30 Hz bandpass sinc FIR filter (Kaiser window, Kaiser beta 5.653, filter order 1812). Epochs started −100 ms before and ended 700 ms after audio onset. Baseline differences in the pre-auditory stimulus interval may be due to ongoing visual processing as a consequence of the visual stimulus dynamics. Therefore, following the procedure of Jessen and Kotz (2011), no baseline correction was performed to not distort the influence of visual preprocessing on auditory processing. Epochs containing artifacts exceeding amplitude changes of 75 μV were rejected. The first three subjects to take part in the EEG study were tested with only 96 trials per condition. To ensure that the EEG results are not confounded by a varying signal-to-noise ratio resulting from a smaller number of trials in the first participants, all statistical analyses described below were additionally conducted with only 14 subjects (excluding the first 3). Statistical effects did not change depending on the number of included participants (14 *vs* 17); thus, all following analyses are reported using the full number of participants.

Statistical analyses of the effects of visual context on the auditory N1 and P2 component were conducted on epochs time-locked to voice onset. Based on the typical topographical distribution of those components, a previous study (Ho *et al.*, 2014) suggesting that emotion and audiovisual congruence interact mostly at anterior sites and the present component topographies, we confined our analysis to electrode Fz. Latency and amplitude analyses were treated independently; thus, analysis parameters were chosen to optimize the informative value of each procedure. Peak latencies of the N1 and P2 component were determined using a jackknife-based technique (Kiesel *et al.*, 2008) to achieve more accurate and robust estimates of latency (Miller *et al.*, 2009). The jackknife procedure tests latencies not on N single-subject averages but on N grand averages of N-1 subjects (leave-one-out method). Resulting F-test statistics were adjusted using formulas provided in Kiesel *et al.* (2008). Subsequently, N1 peak amplitude was analyzed in a 70–150 ms time window across all conditions. Peak amplitude was chosen because prominent peaks were identifiable for N1 in most subjects. As the jackknife analysis revealed significant latency differences between neutral and angry voice conditions for the P2, two separate time-windows were chosen for its mean amplitude analysis, each window centered on the mean latency of neutral and angry conditions, respectively (±20 ms). Thus, the P2 amplitude was analyzed in a 177–217 ms time window for neutral and in a 205–245 ms time window for angry conditions. To control for the possibility that a baseline correction may diminish the component effects, P2-N1 peak amplitude differences were supplementary analyzed. Following the analyses of the behavioral data, a repeated-measures ANOVA with the factors voice conveyed emotion (neutral, angry) and visual context

(auditory-only, audiovisual emotionally congruent, audiovisual emotionally incongruent) was computed for latency and amplitude measures. Appropriate follow-up ANOVAs and pairwise comparisons were calculated. Statistical analyses were conducted with the IBM SPSS Statistics software for Windows, Version 17 (IBM; Armonk, NY, USA).

Complementary to the auditory responses (N1, P2), the visual evoked potentials to face onset are depicted in Figure A of the supplementary material. Additionally, we conducted an analysis of the non-auditory contributions to processing at the time of voice onset to show how audiovisual and visual analysis regions reflect the influence of emotional incongruity. The description and results of these analyses, including the Figures B and C, can also be found in the supplementary material.

RESULTS

In the following, only significant results are reported. For a complete list of ANOVA and follow-up statistics, see Table 1 for the behavioral data and Table 2 for the ERP analyses.

Behavioral results

Figure 2 shows bar graphs depicting the two different measures of behavioral performance derived from the follow-up behavioral experiment. The statistical analyses yielded no significant interactions or main effects of emotion and/or visual context for RTs or hit rates. All participants performed the auditory task with an accuracy of >95%.

Table 1 Behavioral data-results of the repeated-measures ANOVA (Emotion × Visual context)

ANOVA	Reaction times				Accuracy			
	df	*F*	*P*	η^2	*df*	*F*	*P*	η^2
Emotion	1, 9	0.30	0.595	0.033	1, 9	0.36	0.566	0.038
Visual context	1.1, 9.5	2.22	0.169	0.198	2, 18	2.04	0.160	0.184
Emotion × Visual context	1.1, 10.1	1.63	0.233	0.154	2, 18	1.76	0.200	0.164

ERP results

Figure 3 shows the auditory N1 and P2 response (ERPs) to the voice onset for the different visual context conditions at electrode Fz and the corresponding voltage distributions.

N1 latency

The jackknife analysis revealed significant effects of emotion with shorter latencies for neutral compared with angry voices (mean neutral = 99.9 ms, mean angry = 114.0 ms: $F(1, 16) = 21.2$, $P < 0.001$). The significant main effect of visual context ($F(1,5, 23,9) = 35.31$, $P < 0.001$) suggested overall latency reductions in the audiovisual congruent (mean = 100.2 ms) and the audiovisual incongruent (mean = 99.5 ms) compared with the auditory-only (mean = 121.3 ms) condition ($t(16) = -7.81$, $P < 0.001$ and $t(16) = -5.91$, $P < 0.001$, respectively). Such audiovisual latency shortenings may be modulated by emotion as indicated by a marginally significant emotion × visual context interaction ($F(2, 32) = 3.2$, $P = 0.054$).

P2 latency

The analysis of P2 peak latencies revealed a significant main effect of emotion with again shorter latencies for the neutral compared with the angry voices (mean neutral = 197.3 ms, mean angry = 225.3 ms: $F(1, 16) = 10.85$, $P = 0.005$). Based on this, separate time windows were selected for neutral and angry conditions for the P2 mean amplitude analysis.

N1 amplitude

The emotion × visual context ANOVA yielded a significant effect of emotion indicating smaller N1 peak amplitudes for angry compared with neutral voices ($F(1, 16) = 30.5$, $P < 0.001$). We found a significant main effect of visual context suggesting that N1 amplitudes vary as a function of the visual stimulus ($F(2, 32) = 14.19$, $P < 0.001$). Importantly, the significant emotion × visual context interaction showed that the influence of visual information on N1 amplitude is modulated by emotion ($F(2, 32) = 6.28$, $P = 0.005$). The interaction was unpacked by emotion in follow-up ANOVAs with the factor visual context. Both analyses yielded significant main effects of visual

Table 2 Results of all statistical analyses for the auditory N1 and P2 component (at electrode Fz) comprising the repeated-measures ANOVA (Emotion × Visual context) and the corresponding follow-up analyses

Effect	N1 peak latency jackknife			N1 peak amplitude (70–150 ms)				P2 peak latency jackknife			P2 mean amplitude (177–217 / 205–245 ms)				P2-N1 peak-to-peak amplitude (N1:70–150 ms; P2:177–217 / 205–245 ms)			
ANOVA	*df*	F^a	*P*	*df*	*F*	*P*	η^2	*df*	F^a	*P*	*df*	*F*	*P*	η^2	*df*	*F*	*P*	η^2
Emotion	1, 16	21.2	<0.001***	1, 16	30.5	<0.001***	0.656	1, 16	10.85	0.005**	1, 16	3.23	0.091	0.168	1, 16	22.55	<0.001***	0.585
Visual context	1.5, 23.9	35.31	<0.001***	2, 32	14.19	<0.001***	0.470	1.2, 19.8	1.7	0.564	2, 32	33.0	<0.001***	0.673	2, 32	67.23	<0.001***	0.808
Emotion × Visual context	2, 32	3.2	0.054	2, 32	6.28	0.005**	0.282	1.2, 19	0.23	0.964	1.4, 22.4	0.92	0.380	0.055	2, 32	4.22	0.024*	0.209
Follow-up ANOVAs																		
Visual context for neutral				2, 32	13.17	<0.001***	0.452								2, 32	39.78	<0.001***	0.713
Visual context for angry				2, 32	8.63	0.001***	0.350								2, 32	31.61	<0.001***	0.664
Pairwise comparisons	*df*	t^a	*P*	*df*	*t*	*P*					*df*	*t*	*P*		*df*	*t*	*P*	
A *vs* AVc	16	−7.81	<0.001***								16	−3.94	0.003**					
A *vs* AVic	16	−5.91	<0.001***								16	−7.81	<0.001***					
AVc *vs* AVic	16	−0.31	2.283								16	−4.56	0.001***					
A *vs* AVc neutral				16	−3.84	0.004**									16	6.43	<0.001***	
A *vs* AVic neutral				16	−4.24	0.002**									16	8.19	<0.001***	
AVc *vs* AVic neutral				16	1.55	0.424									16	−3.17	0.018*	
A *vs* AVc angry				16	−3.61	0.007**									16	6.36	<0.001***	
A *vs* AVic angry				16	−1.60	0.386									16	6.40	<0.001***	
AVc *vs* AVic angry				16	2.79	0.039*									16	−0.74	1.410	

A, auditory-only condition; AVc, audiovisual congruent condition; AVic, audiovisual incongruent condition.
[a]Adjusted according to Kiesel *et al.*, 2008.
***$P \leq 0.001$; **$P \leq 0.01$; *$P \leq 0.05$.

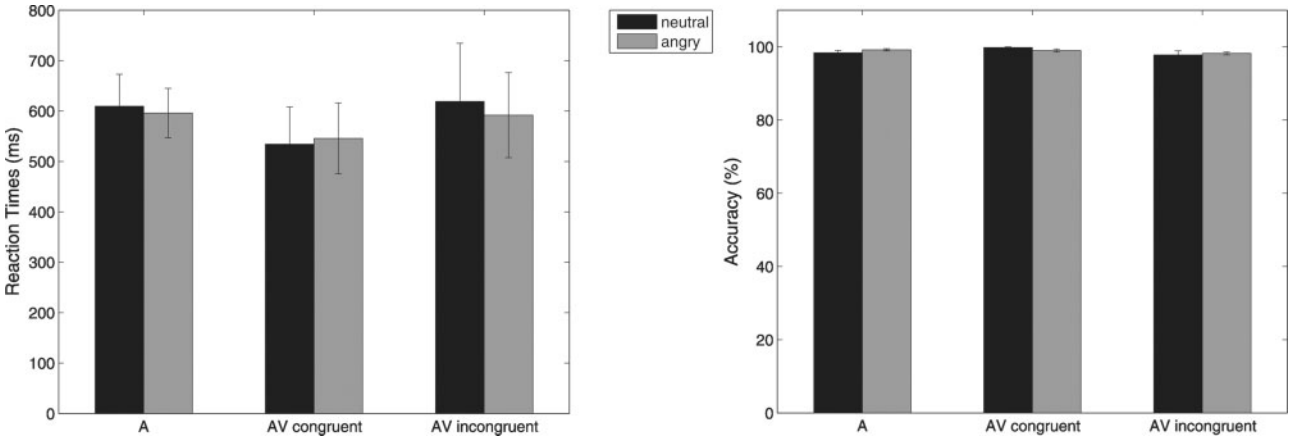

Fig. 2 Mean RTs (left)/mean accuracy (right) and standard errors in the 2-alternative forced choice auditory task ('Was the voice angry or not?'). Abbreviations: A, auditory-only; AV, audiovisual.

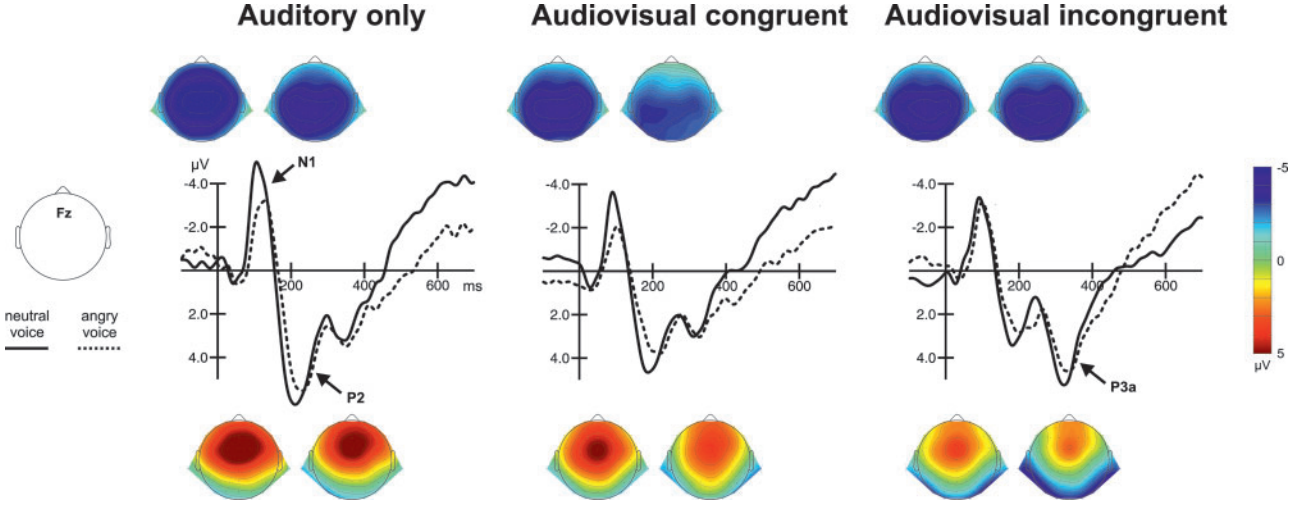

Fig. 3 ERPs (N1, P2) and corresponding topographies (left: neutral, right: angry) to voice onset for the two emotions showing the N1-P2 complex in the auditory-only condition (left) and the modulation of audiovisual N1 and P2 suppression by audiovisual congruency (middle and right). N1 topographies are plotted in time windows centered on the individual condition's peak ± 20 ms, P2 topographies are plotted from 177–217 ms for neutral and from 205–245 ms for angry conditions.

context (neutral: $F(2, 32) = 13.17$, $P < 0.001$; angry: $F(2, 32) = 8.63$, $P = 0.001$) and subsequent pairwise comparisons showed that for neutral voices audiovisual conditions did not differ significantly from each other ($t(16) = 1.55$, $P = 0.424$) but both differed significantly from the auditory-only condition (audiovisual congruent: $t(16) = -3.84$, $P = 0.004$; audiovisual incongruent: $t(16) = -4.24$, $P = 0.002$). This suggests the presence of audiovisual N1 suppression regardless of audiovisual congruency. For angry voices, the audiovisual congruent condition differed significantly from the auditory-only ($t(16) = -3.61$, $P = 0.007$) and the audiovisual incongruent condition ($t(16) = 2.79$, $P = 0.039$) but the latter two were not significantly different ($t(16) = -1.60$, $P = 0.386$), suggesting that audiovisual N1 suppression was present only in the audiovisual congruent but not in the incongruent condition.

P2 amplitude

For P2, the main effect of visual context was highly significant ($F(2, 32) = 32.95$, $P < 0.001$) and subsequent paired-samples t-tests showed that all visual context conditions differed significantly from each other (auditory-only–audiovisual congruent: $t(16) = -3.94$, $P = 0.003$; auditory-only–audiovisual incongruent: $t(16) = -7.81$, $P < 0.001$; audiovisual congruent–incongruent: $t(16) = -4.56$, $P = 0.001$). Thus, across emotions, P2 showed audiovisual amplitude suppression in the audiovisual congruent condition and was additionally affected by audiovisual congruency with smaller amplitudes for incongruent compared with congruent audiovisual stimulation.

DISCUSSION

We investigated how the processing of vocal expressions is influenced by preceding emotionally congruent or incongruent dynamic facial expressions. In contrast to previous emotion studies (e.g. Ho *et al.*, 2014; Pourtois *et al.*, 2002), we included a unisensory auditory condition, enabling us to determine audiovisual suppression of the auditory N1 and P2 component in the congruent and incongruent condition as an indicator for audiovisual integration.

The collected behavioral measures did not yield any significant results. However, by visual inspection the effect patterns seemed consistent with previous findings on behavioral performance in comparable multisensory situations using dynamic faces (e.g. Collignon *et al.*, 2008; Föcker *et al.*, 2011; Klasen *et al.*, 2012), showing a trend for an audiovisual behavioral benefit in the audiovisual congruent compared with the auditory-only condition and a reduction of such benefit for audiovisual incongruity (RTs: for neutral and angry voices, accuracy: for neutral voices). The incongruity effect may be caused by distraction or successful inhibition of task-irrelevant information (see e.g. Talsma *et al.*, 2007). Overall, the absence of any significant behavioral effects in

the current paradigm may be due to the nature of the task, which was easy (decision between two alternatives), as underlined by the generally high hit rates (>95%).

The electrophysiological data showed an auditory emotion suppression effect in the N1 for angry compared with neutral vocalizations. Following previous work (Dietrich *et al.*, 2006, 2008), the usage of non-linguistic interjections allowed eliminating potential semantic confounds on auditory affective processing, suggesting a pure effect of emotional salience. Such emotion effects have already been demonstrated at early sensory processing stages (Schirmer and Kotz, 2006; Paulmann *et al.*, 2009; Jessen and Kotz, 2011; Kotz and Paulmann, 2011; Jessen *et al.*, 2012), indicating that emotional significance can be derived from vocal expressions within a very short time. They have been interpreted as facilitated processing of emotional auditory stimuli (e.g. Paulmann *et al.*, 2009), but there are only few electrophysiological studies investigating the processing of vocal anger expressions (Ho *et al.*, 2014; Jessen and Kotz, 2011). Anger expressions fulfill an essential function in human social behavior, as they are threat signals (Schupp *et al.*, 2004) and demand behavioral adaptation from the observer (Frijda, 1986). Example frequency spectra of a neutral and an angry interjection used in the present study are shown in Figure 5.

Concerning the effects of visual context on auditory processing, the N1-P2 amplitude suppression for neutral and angry vocalizations in the audiovisual congruent compared with the auditory condition confirms our first hypothesis and earlier findings (e.g. Jessen *et al.*, 2012). For the N1, the amplitude suppression effect was accompanied by an unspecific audiovisual latency reduction (see also e.g. Jessen *et al.*, 2012), implying speeded-up processing in the case of bimodal stimulation. This latency modulation was also present in the incongruent audiovisual situation, which contrasts previous findings for audiovisual speech (van Wassenhove *et al.*, 2005; Knowland *et al.*, 2014) showing congruency-specific N1 latency reductions. However, this result supports Stekelenburg and Vroomen (2007), who found that temporal N1 facilitation is independent of audiovisual congruency. Together, these modality effects (amplitude, latency) suggest emotion-unspecific audiovisual facilitation.

Our second hypothesis predicted that owing to the saliency of emotion signals audiovisual N1 suppression would differ for neutral and emotional expressions depending on audiovisual congruency. As expected, we found an interaction of emotion and visual context for N1 amplitude, indicating that for neutral vocalizations, audiovisual response suppression did not differ between the congruent audiovisual condition where the faces were neutral and the incongruent audiovisual condition where the faces were angry. For angry vocalizations, however, audiovisual N1 suppression was evident only in the audiovisual congruent condition where the faces were also angry but not in the audiovisual incongruent condition where the faces were neutral. Thus, in the neutral voice condition, angry faces induced audiovisual response suppression despite audiovisual incongruity, whereas in the angry voice condition, incongruent neutral faces did not. The interaction is also depicted in Figure 4, which illustrates that there is pronounced N1 suppression in all audiovisual conditions except when a neutral face precedes an angry vocalization. Angry faces induced N1 suppression regardless of audiovisual (in)congruity, which implies that they lead to stronger predictions than neutral ones, in line with the predictive coding hypothesis for audiovisual perception. On the other hand, the absence of audiovisual N1 suppression in the incongruent angry voice condition could also be driven by the voice with intensified processing of the auditory emotion in situations of audiovisual mismatch. Either way, the interaction of audiovisual congruency and voice conveyed emotion, which was also confirmed in the N1-P2 peak-to-peak amplitude analysis (see Table 2 for the results of the statistical analysis) at such an early time point is in contrast to previous findings using dynamic non-emotional events and proposing non-specificity of audiovisual N1 suppression to the informational congruency of visual and auditory inputs (e.g. van Wassenhove *et al.*, 2005; Stekelenburg and Vroomen, 2007). Therefore, our findings emphasize the role of emotion and thus saliency in audiovisual integration of ecologically valid events. We suggest that dynamic emotion expressions are preferentially processed in early audiovisual integration and support the view that audiovisual emotion integration may be stronger and qualitatively different from other kinds of audiovisual integration (see also Baart *et al.*, 2014).

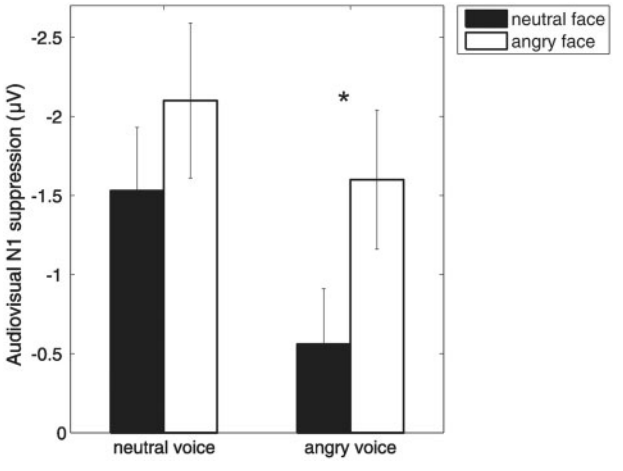

Fig. 4 The audiovisual N1 suppression effect at electrode Fz computed as the difference between the auditory-only and the audiovisual conditions, plotted separately for neutral (left) and angry (right) vocalizations.

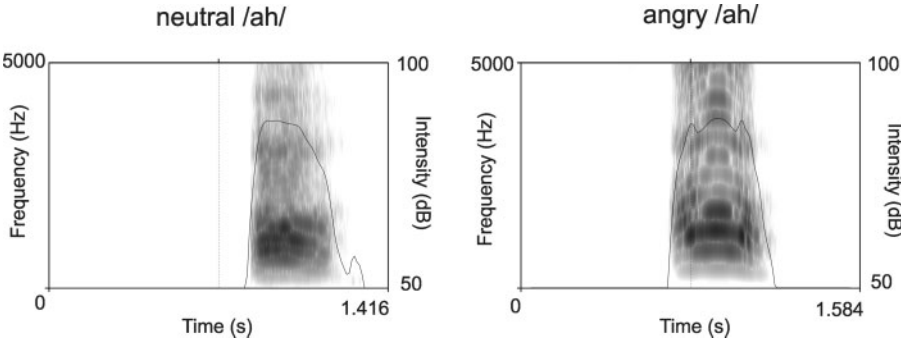

Fig. 5 Frequency spectrum and intensity contour of an example neutral/ah/(left) and an example angry/ah/(right).

In agreement with previous studies (Pourtois *et al.*, 2002; Stekelenburg and Vroomen, 2007, 2012; Vroomen and Stekelenburg, 2010; Jessen and Kotz, 2011) but contradicting others (Klucharev *et al.*, 2003; van Wassenhove *et al.*, 2005) audiovisual emotional incongruity led to discriminative effects on the auditory N1 and P2. The described interaction of emotion and visual context that was found for N1 amplitude was not found for P2 amplitude. On the other hand, no global effect of congruency was observed on N1 amplitude, whereas P2 was reduced in response to incongruent compared with congruent audiovisual expressions (see also Stekelenburg and Vroomen, 2007), possibly reflecting a reduced information gain to incongruent stimuli. Enhanced audiovisual P2 suppression in the incongruent condition is difficult to explain using the predictive coding hypothesis. Knowland *et al.* (2014) have suggested that audiovisual P2 suppression reflects competition between different multisensory inputs with greater competition for incompatible stimulation. Decreased P2 amplitudes have also been linked to increases in attention (Crowley and Colrain, 2004) and P2 suppression to audiovisual incongruity could imply attentional capture by mismatching vocalizations. In line with this, visual inspection of our EEG data revealed P3a elicitation for both neutral and emotional vocalizations in the incongruent audiovisual situation ($\sim$240–370 ms post-stimulus), suggesting alerting processes and as a consequence possibly involuntary orienting of attention toward the visual modality (see also Squires *et al.*, 1975). Altogether, the traditional view of the N1 and P2 component as part of the 'vertex potential' or N1-P2 complex has been challenged by a number of studies suggesting that both components, as well as their latency and amplitude effects, can be functionally dissociated during multisensory integration (e.g. van Wassenhove *et al.*, 2005; Vroomen and Stekelenburg, 2010; Baart *et al.*, 2014).

CONCLUSION

We studied early audiovisual integration of dynamic angry and neutral expressions (face–voice pairs) by measuring the suppression of the auditory N1 and P2 responses in audiovisual compared with auditory-only conditions. Emotional congruency of facial and vocal expressions was manipulated to investigate how audiovisual response suppression is influenced by a predictive *vs* non-predictive visual context. Consistent with our hypothesis but in contrast to previous work, audiovisual N1 suppression in the present study was both congruency- and emotion-specific. We suggest an advantage of dynamic emotion signals in early audiovisual integration and emphasize the importance of biological motion for multimodal perception, providing a valuable starting point for a more ecologically valid understanding of multimodal emotion expression perception.

SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

REFERENCES

Ambadar, Z., Schooler, J.W., Cohn, J.F. (2005). Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, *16*(5), 403–10.

Arnal, L.H., Morillon, B., Kell, C.A., Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(43), 13445–53.

Baart, M., Stekelenburg, J.J., Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115–21.

Bassili, J.N. (1979). Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, *37*(11), 2049–58.

Besle, J., Fort, A., Delpuech, C., Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *The European Journal of Neuroscience*, *20*(8), 2225–34.

Bertelson, P., Gelder, B. de (2004). The Psychology of Multimodal Perception. In: Spence, C., Driver, J., editors. *Crossmodal Space and Crossmodal Attention*. Oxford: Oxford University Press.

Bradley, M.M., Lang, P.J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., et al. (1997). Activation of auditory cortex during silent lipreading. *Science (New York, N.Y.)*, Vol. 276(5312), 593–6.

Campanella, S., Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–43.

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

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–35.

Crowley, K.E., Colrain, I.M. (2004). A review of the evidence for P2 being an independent component process: age, sleep and modality. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *115*(4), 732–44.

De Gelder, B., Böcker, K.B., Tuomainen, J., Hensen, M., Vroomen, J. (1999). The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neuroscience Letters*, *260*(2), 133–6.

De Gelder, B., Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289–311.

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

Dietrich, S., Ackermann, H., Szameitat, D.P., Alter, K. (2006). Psychoacoustic studies on the processing of vocal interjections: how to disentangle lexical and prosodic information? *Progress in Brain Research*, *156*, 295–302.

Dietrich, S., Hertrich, I., Alter, K., Ischebeck, A., Ackermann, H. (2008). Understanding the emotional expression of verbal interjections: a functional MRI study. *Neuroreport*, *19*(18), 1751–5.

Ekman, P., Friesen, W.V. (1976). *Pictures of Facial Affect*. Palo Alto. CA: Consulting Psychologists Press.

Ethofer, T., Anders, S., Erb, M., et al. (2006). Impact of voice on emotional judgment of faces: an event-related fMRI study. *Human Brain Mapping*, *27*(9), 707–14.

Föcker, J., Gondan, M., Röder, B. (2011). Preattentive processing of audio-visual emotional signals. *Acta Psychologica*, *137*(1), 36–47.

Frijda, N.H. (1986). *The Emotions*. Cambridge, UK: Cambridge University Press.

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(20), 5004–12.

Ho, H.T., Schröger, E., Kotz, S.A. (2014). Selective attention modulates early human evoked potentials during emotional face-voice processing. (In press).

Jessen, S., Kotz, S.A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, *58*(2), 665–74.

Jessen, S., Kotz, S.A. (2013). On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, *7*, 369.

Jessen, S., Obleser, J., Kotz, S.A. (2012). How bodies and voices interact in early emotion perception. *PloS One*, *7*(4), e36070.

Kiesel, A., Miller, J., Jolicoeur, P., Brisson, B. (2008). Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–74.

Klasen, M., Chen, Y.-H., Mathiak, K. (2012). Multisensory emotions: perception, combination and underlying neural processes. *Reviews in the Neurosciences*, *23*(4), 381–92.

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.

Knowland, V.C.P., Mercure, E., Karmiloff-Smith, A., Dick, F., Thomas, M.S.C. (2014). Audio-visual speech perception: a developmental ERP investigation. *Developmental Science*, *17*(1), 110–24.

Kotz, S.A., Paulmann, S. (2011). Emotion, Language, and the Brain. *Language and Linguistics Compass*, *5*(3), 108–25.

LaBar, K.S., Crupain, M.J., Voyvodic, J.T., McCarthy, G. (2003). Dynamic Perception of Facial Affect and Identity in the Human Brain. *Cerebral Cortex*, *13*(10), 1023–33.

Miller, J., Ulrich, R., Schwarz, W. (2009). Why jackknifing yields good latency estimates. *Psychophysiology*, *46*(2), 300–12.

Paulmann, S., Jessen, S., Kotz, S.A. (2009). Investigating the Multimodal Nature of Human Communication. *Journal of Psychophysiology*, *23*(2), 63–76.

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport*, *11*(6), 1329–33.

Pourtois, G., Debatisse, D., Despland, P.-A., de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Brain Research. Cognitive Brain Research*, *14*(1), 99–105.

Schirmer, A., Kotz, S.A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30.

Schubert, E. (1999). Measuring emotion continuously: validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154–65.

Schupp, H.T., Ohman, A., Junghöfer, M., Weike, A.I., Stockburger, J., Hamm, A.O. (2004). The facilitated processing of threatening faces: an ERP analysis. *Emotion (Washington, D.C.)*, *4*(2), 189–200.

Squires, N.K., Squires, K.C., Hillyard, S.A. (1975). Two varieties of long-latency positive 35 waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*(4), 387–401.

Stekelenburg, J.J., Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–73.

Stekelenburg, J.J., Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*, 26.

Talsma, D., Doty, T.J., Woldorff, M.G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral Cortex (New York, N.Y.: 1991)*, Vol. 17(3), 679–90.

Van Wassenhove, V., Grant, K.W., Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–6.

Vroomen, J., Stekelenburg, J.J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–96.

Wagner, H.L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3–28.