See https://doi.org/10.1016/j.cortex.2021.12.017

Emotional voices modulate perception and predictions about an upcoming face

Marc David Pell[a], Sameer Sethi[a], Simon Rigoulot[a,1], Kathrin Rothermich[a,2], Pan Liu[a,3], & Xiaoming Jiang[a,4]

[a]McGill University, School of Communication Sciences and Disorders,
2001 avenue McGill College, 8th floor, Montréal, Québec, H3A 1G1, Canada

Corresponding authors:

1. *Marc D. Pell, Ph.D.*
   McGill University
   marc.pell@mcgill.ca
   Phone: (514) 398 4133

2. Xiaoming Jiang, Ph.D.
Shanghai International Studies University
xiaoming.jiang@shisu.edu.cn

Current address**:**

[1]Department of Psychology, Université du Québec à Trois-Rivières, Trois-Rivières, Canada

[2]East Carolina University, Dept. of Communication Sciences & Disorders, Greenville, USA

[3]North Dakota State University, Department of Psychology, Fargo, USA

[4]Shanghai International Studies University, Institute of Linguistics, Shanghai, China

Abstract

When we hear an emotional voice, does this alter how the brain perceives and evaluates a subsequent face? Here, we tested this question by comparing event-related potentials evoked by angry, sad, and happy faces following vocal expressions which varied in form (speech-embedded emotions, non-linguistic vocalizations) and emotional relationship (congruent, incongruent). Participants judged whether face targets were true exemplars of emotion (facial affect decision). Prototypicality decisions were more accurate and faster for congruent vs. incongruent faces and for targets that displayed happiness. Principal component analysis identified vocal context effects on faces in three distinct temporal factors: a posterior P200 (150-250ms), associated with evaluating face typicality; a slow frontal negativity (200-750ms) evoked by angry faces, reflecting enhanced attention to threatening targets; and the Late Positive Potential (LPP, 450-1000ms), reflecting sustained contextual evaluation of intrinsic face meaning (with independent LPP responses in posterior and prefrontal cortex). Incongruent faces and faces primed by speech (compared to vocalizations) tended to increase demands on face perception at stages of structure-building (P200) and meaning integration (posterior LPP). The frontal LPP spatially overlapped with the earlier frontal negativity response; these components were functionally linked to expectancy-based processes directed towards the incoming face, governed by the *form* of a preceding vocal expression (especially for anger). Our results showcase differences in how vocalizations and speech-embedded emotion expressions modulate cortical operations for predicting (prefrontal) versus integrating (posterior) face meaning in light of contextual details.

Keywords: social perception, priming, face typicality, P200, Late Positive Potential, anger.

1. Introduction

The meanings we derive from facial expressions are influenced by the context in which they appear (Barrett, Mesquita, & Gendron, 2011; Wieser and Brosch, 2014). People assign different value to emotional displays when they possess knowledge about the poser's traits or intentions (Clark, McNeel, Bigelow, & Enticott, 2020; Rischer et al., 2020) or when facial expressions are embedded in visual scenes or verbal situations (Hess, Dietrich, Kafetsios, Elkabetz, & Hareli, 2020; Hietanen & Astikainen, 2013; Righart & de Gelder, 2008). Often, these context effects are driven by *congruency* relations; evaluative decisions about a face are facilitated when contextual details share the same affective tone (e.g., hedonic valence) or discrete emotional properties than when they somehow conflict with the face target and/or its evaluative goals (Aguado, Parkington, Dieguez-Risco, Hinojosa, & Itier, 2019; Hinojosa, Carretié, Méndez-Bértolo, Míguez, & Pozo, 2009).

Neurophysiological studies provide valuable insights as to how context shapes cortical face processing at different functional stages and time points in the visual processing stream (see Schweinberger & Neumann, 2016 for a model). Structural (visual-perceptual) encoding of a face can be inferred from two overlapping temporo-occipital components in the event-related potential (ERP): the N170, which reflects encoding of primary visual features to detect a face (Hinojosa, Mercado, & Carretié, 2015); and the less-studied P200, which reflects deeper analysis of a face's second-order spatial relations (Latinus, VanRullen, & Taylor, 2010). In parallel, distinct neural mechanisms register *affective* qualities of the facial expression, producing sustained brain activity at longer latencies as emotional stimulus properties are elaborated and associative mechanisms come into play (Eimer & Holmes, 2002;

Vuilleumier & Pourtois, 2007). Intrinsically motivating stimuli (e.g., faces high in arousal) evoke an Early Posterior Negativity (EPN) ~150-350ms post-stimulus, taken as evidence of enhanced perceptual encoding of biologically salient events (Schupp, Flaisch, Stockburger, & Junghofer, 2006; Schupp & Kirmse, 2021). These early stages of perception and attentional orienting to facial expressions are often sensitive to top-down influences of a preceding context or prime stimulus, at least in certain conditions (see Aguado et al., 2019 for a critical analysis).

At longer latencies, post-perceptual processing of a face is strongly influenced by what comes before the target. When primed by verbal descriptions of emotion, the N400 component increases for incongruent vs. congruent faces (Diéguez-Risco et al., 2015a; Dozolme, Brunet-Gouet, Passerieux, & Amorim, 2015; Krombholz, Schaefer, & Boucsein, 2007), highlighting that congruency relations modulate access to semantic details associated with a face in memory (Kutas & Federmeier, 2011). Context-face effects are particularly evident on the Late Positive Potential (LPP), a centroparietal positive-going wave which reflects sustained motivated attention to particular targets ~350-700ms post-onset of a face (Duval, Moser, Huppert, & Simons, 2013; Schupp et al., 2006). The LPP tends to increase when faces are affectively incongruent or unexpected based on preceding events (Dozolme et al., 2015; Rischer et al., 2020; Stolz, Endres, & Mueller, 2019). Arguably, these conditions tax mechanisms for constructing an internal model of face meaning at longer processing latencies, especially when expectancies formed by a preceding event yield prediction errors about the target input (Bornkessel-Schlesewsky & Schlesewsky, 2019).

1.1 Voice-face effects during emotion processing

We possess only a rudimentary understanding of how *auditory* information—music, environmental sounds, voices, etc.—affects neurophysiological responses to an upcoming face (see Gerdes et al., 2013; Puce, Epling, Thompson, & Carrick, 2007 for examples). Among these different auditory events, the link between faces and human *vocal* displays is arguably most salient; expressions in these two channels are highly integrated at the neurofunctional level during speech, affect, and person perception (Young, Frühholz & Schweinberger, 2020) and seem to form privileged memory associations (Bülthoff & Newell, 2017). This privileged relationship motivated the focus of the current study on voice-face effects.

In daily life, people encounter vocal and facial cues in tandem (multi-modal processing) or in sequence (cross-modal processing, e.g., when orienting to a person's face after hearing a voice). Both situations lead to mandatory interactive effects of expressions in each modality on cognition (Baart & Vroomen, 2018; de Gelder & Vroomen, 2000; Lavan, Lima, Harvey, Scott, & McGettigan, 2014; Massaro & Egan, 1996; Pell, 2005a). Most studies have focused on multimodal integration of emotion from voice-face expressions, for example when a static or dynamic face is temporally aligned with an emotionally inflected (pseudo)word or vocalization (Föcker & Röder, 2019; Ho, Schröger, & Kotz, 2015; Jessen & Kotz, 2011; Kokinous, Kotz, Tavano, & Schröger, 2015; Liu et al., 2012; Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000). When temporally conjoined, affective congruency effects are observed in the ERP by 200ms post-stimulus, modulating the anterior P200, emphasizing that emotional features of both events are quickly registered by the

brain. However, these data do not inform how vocal expressions modulate cortical activity evoked by an *upcoming* face (Garrido-Vásquez, Pell, Paulmann, & Kotz, 2018). For instance, would neural responses to an angry face differ after hearing someone talk in a happy versus angry voice, and how does the preceding vocal expression impact at different functional stage(s) of face perception?

This question was partially addressed in an ERP study by Paulmann & Pell (2010), who presented emotionally inflected pseudo-speech (*The plackter jabbored the tozz*) that was congruent or incongruent with a subsequent facial expression. After hearing an utterance, participants viewed an emotional face (anger, fear, happy) or a non-emotional facial expression and judged whether the target represents a "true" display of emotion (Facial Affect Decision Task, Pell, 2005a; see Figure 1 for examples). This task draws attention to emotional qualities without requiring participants to categorize/name the face (see Pell, 2005b for details). Results showed that emotional faces were sensitive to the congruency of the vocal prime, modulating the N400 amplitude to faces even when speech excerpts were brief (200-400ms in duration, see Paulmann & Pell, 2010 for precise details). These patterns suggest that voice congruency could influence how face meaning is constructed at the N400 and/or LPP stages, particularly when tasks require emotional analysis of the incoming face (Diéguez-Risco et al., 2015a; Krombholz et al., 2007). However, since Paulmann and Pell (2010) restricted analyses to the N400 time window, their data offer only a hint of the temporal and spatial neural dynamics that likely characterize voice-face effects during cross-modal processing of these stimuli in daily life (see Ethofer et al., 2006; Watson et al., 2014 for fMRI data).

While clarifying how voice-face congruency influences face perception, the physical *form* of a vocal signal could further alter expectancies about an upcoming face (Bryant, 2021). Humans express vocal emotions in two distinct ways: in running speech (*prosody*), or in a non-linguistic form (grunts of anger, laughter, crying, etc.). Non-linguistic vocalizations are relatively short signals that lack the linguistic structure and acoustic constraints of speech-embedded emotional cues (Agnew, Ward, McGettigan, Josephs, & Scott, 2017); spontaneous productions of these signals likely emanate from more primitive (limbic-based) vocal call systems shared by other mammals (Owren et al., 2011). Even when volitionally produced, vocalizations communicate emotion with greater perceptual clarity than speech (Hawk, van Kleef, Fischer, & van der Schalk, 2009) and are preferentially attended by the neurocognitive system as distinct auditory events with heightened motivational relevance (Agnew et al., 2017; Pell et al., 2015; Sauter & Eimer, 2010).

It is unclear from existing research how the form of a vocal expression alters face processing, for example, by priming or modulating predictions about the nature of an upcoming face. In a study that compared ERPs evoked from the onset of emotional vocalizations versus speech, the form of the vocal expression led to cortical differentiation of both the N100 and P200 (centroanterior) response. Vocalizations increased the P200 amplitude which displayed an earlier peak than speech-embedded expressions of the same emotion, demonstrating that the brain registers vocalizations as more salient acoustic events than speech at early processing stages (Pell et al., 2015). These data allow us to hypothesize that in addition to voice-face congruency, information about the manner in which vocal expressions are produced

will furnish relevant cues that shape expectancies and cortical responses to an incoming face.

1.2 Objectives

To cast light on how the human voice alters perception of a subsequent face, we recorded electrocortical responses to angry, sad and happy faces when they followed emotional voices varying in congruency and in form (speech, vocalizations). Following Paulmann & Pell (2010), our task required participants to conduct an emotional analysis of each face to render a prototypicality judgement (facial affect decision, Pell, 2005a) as the EEG was recorded. To avoid assumptions about how and when vocal features influence face perception, here neurophysiological responses were identified in a data-driven manner using temporal-spatial Principal Component Analysis (Dien, 2012; Pourtois, Delplanque, Michel, & Vuilleumier, 2008). ERP analyses focused on how attributes of the vocal prime influence neurocomputational processes underlying each type of facial expression over time.

We expected that facial affect decisions would place high demands on the structural decoding of faces (N170, P200), which may be differentially modulated by the congruency of a preceding voice (Diéguez-Risco et al., 2015). Congruency should also impact on late ERP components which reflect semantic and affective analysis of contextualized face meaning (N400, LPP); notably, difficulties integrating voice meaning with emotionally incongruent faces should increase the posterior LPP response over congruent faces in the 350-700ms time window. The effects of voice *form* on face-related ERPs could not be predicted with certainty. We speculated that

the perceptual clarity and higher motivational salience of vocalizations should alter expectancies about an upcoming face when compared to speech, for example, by modulating mechanisms for emotional vigilance or stereotyped reactions towards the input (Bryant, 2021). Finally, we anticipated that top-down influences of vocal features would interact with endogenous properties of each type of facial expression in different ways; this could alter the nature and time course of cortical responses evoked by particular facial expressions (especially angry targets, which are associated with threat and avoidance tendencies, Rellecke, Sommer, & Schacht, 2012; Schutter, De Haan, & Van Honk, 2004).

1. Methods

The research protocol was pre-approved by the McGill University Faculty of Medicine Institutional Review Board. No part of the study procedures or analyses was pre-registered before undertaking the work. We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Study data, experimental code, and examples of auditory speech stimuli can be accessed online (https://osf.io/rjf4v/).

2.1 Participants

Twenty-four right-handed native speakers of English (12 female/12 male, Mean Age = 22.9 years, SD = 3.5) without history of major neurological or psychiatric illness were recruited via online advertisements at McGill University (Montreal, Canada).

Each participant provided written consent to participate prior to the study. A priori power analysis was conducted to determine the minimal sample size using a simulation-based approach using simr package in R (Brysbaert & Stevens, 2018; Kumler et al., 2021). The simulation started with five participants. The intercept and slopes for the fixed factors (Congruency, Voice Form) and their interactions, the random intercepts for each random factor (participant, electrode), and the residual variance were specified based on previous reports and a preliminary analysis based on five randomly sampled participants (Kumler et al., 2021; Paulmann & Pell, 2010). Power for the effect of Voice Form and for Congruency were each separately detected with 20 simulations; the analysis revealed a power of 75% for the effect of Voice Form and a power of 90% for Congruency. Power reached 100% if the sample size increased to 15 for the effect of Voice Form or if the sample size reached 20 for Congruency for the simulated dataset. This ensured that our sample of 24 participants yielded high power. All participants reported good hearing and normal vision. Prior to the EEG experiment, each participant completed questionnaires to assess their state of alertness and anxiety level (State-Trait Anxiety Inventory, Spielberger et al., 1983).

2.2 Materials

The experiment was composed of short audio recordings (vocal primes) and static photographs (face targets) paired for cross-modal presentation in the Facial Affect Decision Task (FADT, Pell, 2005a). Vocal expressions took two forms: half were non-linguistic vocalizations and half were emotionally inflected utterances (henceforth

referred to as *vocalizations* and *speech*). Vocal and facial displays were selected from published inventories and conveyed one of three emotions (angry, sad, happy).

2.2.1 Vocal expressions – The vocal stimuli were selected for a previous study of vocal expression processing (Pell et al., 2015) and were all produced in a laboratory setting. Following Paulmann & Pell (2010), the speech stimuli were short pseudo-utterances (*He placktered the tozz*) produced to express anger, sadness or happiness by ten speakers of Canadian English (6 female/4 male; taken from Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). Vocalizations were selected from the Montreal Affective Voices database (Belin, Fillion-Bilodeau, & Gosselin, 2008) and consisted of growling sounds (anger), crying (sad), and laughter (happy) produced by ten distinct speakers (5 female/5 male). Selected items were recognized for their target meaning at high accuracy rates when judged by separate listener groups in each validation study (>75% correct, or minimum five times chance level) and mean recognition accuracy for the three emotion categories was roughly similar. A brief pilot study was run to characterize perceived affective differences between the speech and vocalization stimuli when judged by a single group of listeners.[1]

[1] Fourteen participants who did not take part in the EEG study classified the emotion of each vocal expression and separately rated them for arousal and valence. In summary, vocalizations were recognized more accurately than speech (M = 91% vs. 74%), and happy expressions (*M* = 91%) were identified better than anger (*M* = 72%) and sadness (*M* = 82%) irrespective of voice form. Arousal ratings of anger, sad, and happy vocalizations did not differ, whereas angry and happy speech were perceived as more aroused than sad speech. Arousal differences in voice form were only observed for happiness (vocalizations > speech). The three emotional expressions differed significantly in perceived pleasantness (happy > anger > sad). Interestingly, positive emotions (happy) were judged more pleasant when expressed as a vocalization vs. speech, whereas negative emotions (anger, sad) were judged *less* pleasant in the form of vocalizations versus speech. Complete details and statistical reporting are provided by Pell et al. (2015).

In total, 30 unique speech stimuli and 30 unique vocalizations were used in the experiment (3 emotions x 10 speakers x 2 voice forms = 60 vocal stimuli). For each form of expression, vocal stimuli varied naturally in duration (from 900-2000ms, see Pell et al., 2015 for acoustic details of the stimuli). To control for the amount of acoustic stimulation participants received as a function of voice form, individual pseudo-utterances were paired with a vocalization conveying the same emotion and truncated to the exact same duration (Mean duration: anger = 924ms, sadness = 1990ms, happiness = 1435ms). The duration of speech and vocalizations did not significantly differ ($p>.05$). All stimuli were normalized to a peak intensity of 75 dB to mitigate gross differences in perceived loudness across stimuli/datasets.

2.2.2 Facial expressions – Target stimuli were 13.5 x 19 cm colour images of cropped facial expressions conveying anger, sadness, or happiness (Pell & Leonard, 2005; Pell, 2002). In addition, non-emotional displays of affect posed by the same actors were used to facilitate facial affect decisions (Figure 1). Previous work shows that these non-emotional faces are not recognized as basic emotions, are not subject to behavioral priming effects by a preceding voice (Pell, Jaywant, Monetta, & Kotz, 2011), and are differentiated from (prototypical) emotional faces early during neurocognitive processing (Paulmann & Pell, 2009). Expressions posed by ten different actors (6F/4M) of various ages and ethnicities were selected. Each actor posed three unique expressions per emotion (3 emotions x 10 actors x 3 poses = 90 emotional faces total) and nine unique non-emotional expressions (10 actors x 9 poses = 90 total). This resulted in 180 distinct target expressions. Mean emotion

recognition accuracy based on a seven forced-choice validation study (Pell, 2002) was high for all emotions (anger = 92%, sad = 88%, happy = 99%). All non-emotional expressions were rejected as valid exemplars of emotion at rates exceeding 90%.

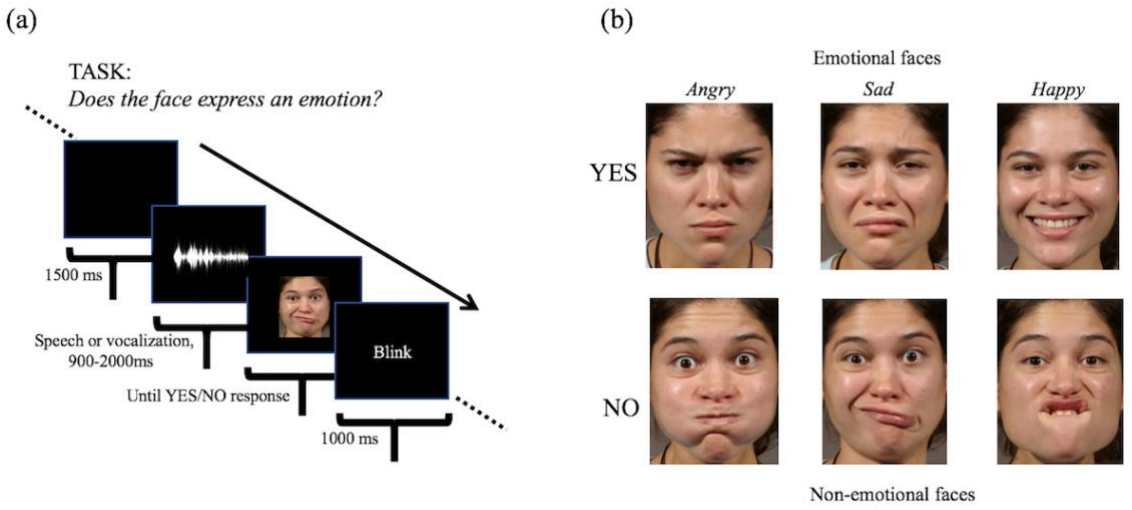


Figure 1. Example of a) trial sequence and b) emotional and non-emotional face targets presented in the experiment.

2.3 Experimental design

Each trial was composed of a voice (speech or vocalization) followed immediately by a face target (emotional or non-emotional expression). Participants made a timed yes/no decision about whether the face expresses an emotion (*yes* response for angry, sad, and happy, *no* response for non-emotional faces). Similar to lexical decisions, facial affect decisions require participants to access emotion-related semantic details about faces without having to map this information onto emotion words (Pell, 2005a).

A total of 1,080 trials were presented, reflecting an equal number of targets resulting in a *yes* (n = 540) and *no* (n = 540) decision.

Current hypotheses are informed only by voice-related effects on emotional faces (*yes* trials). For these critical targets, 180 trials (2 voice form x 3 emotions x 30 items) were composed of an emotionally *congruent* voice and face (anger-anger, sad-sad, happy-happy, congruent = 17% of all experimental trials). Another 360 trials consisted of incongruent pairings of the same stimuli (anger-sad, anger-happy; sad-anger, sad-happy; happy-anger, happy-sad). Non-emotional expressions were paired with the same vocal primes to mimic the structure of *yes* trials (2 voice form x 3 emotions x 90 non-emotional faces = 540). Speech and vocalizations were paired with a face of the same sex, although there was no strict association between the speaker/ poser identity in the experiment as a whole. In total, each vocal stimulus appeared nine times paired with emotional faces (i.e., three times with different congruent faces and six times with different incongruent faces) and nine times paired with different non-emotional expressions. Each facial display appeared six times in total.

2.4 Testing and EEG recording procedure

Participants were tested in an electrically shielded, sound-attenuating chamber seated 65cm from a computer monitor. Vocal stimuli were delivered through earphones at a comfortable volume. Each trial began with the vocal stimulus accompanied by a fixation cross in the middle of the computer screen; the target face was presented directly at the offset of the auditory event. Participants were told that they would hear people express themselves in different ways followed by a facial

expression, and that some of the sounds would not "make sense". They were told to direct their attention to the face and decide whether it conveys an emotion as accurately and quickly as possible. The target remained on the screen until the participant pressed *yes* or *no* on a response box. Once a response was recorded, the instruction "Blink" was presented on the screen for 1000ms followed by a 1500ms inter-trial interval before the next trial.

The experiment was divided into six presentation blocks of 180 (90 *yes,* 90 *no*) trials, with an equal proportion of trials containing speech and vocalizations pseudo-randomized within blocks. No face appeared twice in the same block and no vocal stimulus was repeated within 12 consecutive trials in a block. Block presentation order was counterbalanced across participants, who started with two practice blocks each consisting of 10 trials that did not appear in the experiment. To become familiar with the prototypicality judgement, the first practice block presented emotional and non-emotional faces without vocal context; participants made a facial affect decision and received written feedback after each trial ("Correct", "Incorrect"). During a second practice block, the same faces were presented but with a preceding vocal prime, followed again by feedback on decision accuracy. Following the two practice blocks, no additional feedback was provided to participants. The experiment lasted approximately 3 hours (including questionnaires and EEG preparation). To limit fatigue, two breaks were programmed within each block (after every 60 trials) and a mandatory rest break was imposed between blocks. Participants received $40 CAD upon completion of the study.

The EEG was recorded by 64 active Ag/AgCl electrodes mounted in an elastic cap (actiCAP, Brain products) according to the expanded 10-20 system. Four additional electrodes were placed for vertical and horizontal electro-oculogram recording: two at the outer canthi of eyes and one above and below each eye. The signal was recorded continuously with a band pass between DC and 125 Hz, digitized at a sampling rate of 250 Hz, maintaining electrode impedances below 5 KΩ. Data were re-referenced offline to the average of the electrodes and then filtered with a bandpass of 0.01 and 30 Hz using EEGLab (Delorme & Makeig, 2004). Data from channels that were consistently poor (> 20% of trials for a given participant per channel) were replaced through spherical interpolation. Rejection of artifacts (e.g., eye blinks) and drifts was performed by automatically rejecting VEOG-artifacts above 75 μV, and voltage deflections exceeding 200 μV at any other electrode, followed by visual inspection and manual rejection/correction of trials.

2.5 ERP data analysis

ERP analyses looked at evoked responses time-locked to the onset of faces correctly identified as conveying an emotion. After pre-processing and elimination of error trials, an average of 31 trials (range = 19-45) entered into the analyses per face/congruency/voice form condition. To manage ERP waveform blurring caused by trial-to-trial latency jitter (Ouyang, Sommer, & Zhou, 2016), the ERP data first entered RIDE (residue iteration decomposition); this method showed benefits in revealing the time-resolved dynamics that different sources play in face perception, especially when late ERP components are studied (Kashyap et al., 2016; Meyer et al., 2021). If uncorrected, trial-to-trial latency variability can attenuate ERP component

amplitudes and amplitude differences between conditions in relation to noise, diminishing the size of experimental effects and statistical test parameters (among other potential issues). RIDE mitigates these problems by utilizing the latency variability and time markers to separate ERP components into a stimulus-locked (S), a response-locked (R) and an intermediate component cluster (Stürmer, Ouyang, Zhou, Boldt, & Sommer, 2013).[2]

To avoid bias and assumptions associated with conventional ERP component labelling, temporal-spatial Principal Component Analysis (PCA) was used to decompose ERP amplitudes according to their temporal and spatial correlational structure (ERP PCA Toolkit, Version 2.54, Dien, 2010, 2012). Based on subject-level ERP data, averaged from artifact-free trials and corrected for inter-trial latency variability, the PCA identifies independent spatial component(s) associated with the experimental factors within a certain temporal window and determines if the component response is positive or negative. The temporal and spatial variances entered the PCA by including the sampling points of the entire ERP segment (100ms before the onset and 1000ms after the onset of the face) on all EEG channels.

[2] RIDE was performed independently on all EEG epochs for the face stimuli per participant and per condition. The time window for extracting the "S" component was 0-500ms, the "C" component was 100-900ms, and the "R" component was -300-300ms around the response. The latency of S for each trial was set to be locked to the stimulus onset. The latency of C component was first estimated by Woody's method within the time windows. After the latencies for the three component clusters were obtained, the data were subjected to RIDE composition into three component clusters associated with the three latency sets. The decomposition step and latency updating step were iterated until convergence (Ouyang, Herzmann, Zhou, & Sommer, 2011; Ouyang et al., 2016). The potential sub-component clusters defined by RIDE were re-synchronized to its own latency across single trials and located at the most probable latency. The ERPs were reconstructed by compensating for the trial-to-trial latency variability.

Individual variance was accounted for by including averaged ERPs for each subject into the PCA. The expression form entered the PCA as a condition variance. A temporal PCA was first performed and the retained factors were determined by a Parallel Scree Test (Horn, 1965). A subsequent spatial PCA was then performed on each temporal factor which survived the test. Oblique rotations (Promax method) were performed to achieve the largest representation of a factor as one ERP component in the temporal and the spatial PCA (Dien, 2012). Factor loadings were rescaled to microvolts by converting them into covariance loadings (Dien, 2006).

2.6 Statistical analysis

We analyzed covariance loadings of the peak channel at peak time point for each temporal-spatial (T-S) factor that explained more than .5% of unique variance, with at least one electrode above .5 threshold factor loading (time-locked to the face). Trials were labeled according to voice Form (speech, vocalization), Congruency (congruent, incongruent), and Face (angry, happy, sad). For each T-S factor, we built linear mixed effects models (LMM) to evaluate the significance of the main condition effects and their interactions (e.g., Jiang & Pell, 2015). Maximal random effect structures were kept diminishing type I error (Barr, Levy, Scheepers, & Tily, 2013). To clearly exemplify the effects of vocal prime characteristics, LMMs were built separately for each facial expression taking Form and Congruency as fixed factors. Electrode was included as a fixed factor, selecting all channels whose factor loading on a T-S factor was greater than .6. To limit sources of "within-perceiver" contextual variability on our results (Barrett et al., 2011), all models included STAI trait anxiety

scores (Pell et al., 2015; Wieser & Moscovitch, 2015) as an additional fixed factor to participant as the random factor, and age and gender were included as control factors.

2. Results

3.1 Behavioural data

Behavioral performance in the experiment is summarized in Table 1. Participants were highly accurate at discriminating whether faces expressed an emotion (Yes trials: M =91.0%, SE =12.3% correct; No trials: M = 91.9%, SE =9.1% correct). Decision times tended to be shorter for emotional (*M*=802ms, SE =351ms) than non-emotional (*M*=846ms, SE=404ms) faces. To account for the speed-accuracy-trade-off and unambiguously quantify behavioral performance for two-choice speeded decisions, drift rates were calculated via the EZ-diffusion model (Wagenmakers, Van der Maas, Grasman, 2007). The drift rate values were calculated in R based on the mean RT, variance RT, and mean accuracy per experimental condition and participant. An LMM was built on the drift rates with Face (angry, sad, happy), voice Form (speech, vocalization), and Congruency (congruent, incongruent) as fixed effects, participant age, trait anxiety, and sex as controlling factors, and participant as the random factor. The model revealed a significant effect of Face ($F(2, 238)=49.18$, $p<.0001$) and Congruency ($F(1, 238)=5.59$, $p=.02$). Angry and sad faces had a lower drift rate than happy faces ($b=-.05$, $t=-4.89$, $p<.0001$; $b=-.03$, $t=-3.26$, $p=.001$), indicating that happy targets were judged to be emotional expressions more quickly and accurately overall. The drift rate was lower for incongruent vs. congruent trials ($b=-.02$, $t=-1.84$, $p=.06$), meaning that incongruent faces tended to elicit slower and less accurate decisions.

There was no evidence that the *form* of vocal expression (speech, vocalization) influenced behavioral decisions about face targets.

Table 1. Mean ± standard error for accuracy (%) and latency (milliseconds) of facial affect decisions to emotional targets according to characteristics of the vocal prime.

		Voice prime			
		Speech		Vocalization	
	Face target	*Congruent*	*Incongruent*	*Congruent*	*Incongruent*
Accuracy (%)					
	Angry	86.7±14.1	86.1±15.9	88.0±13.4	85.2±16.1
	Happy	96.5 ± 5.2	96.1 ± 8.0	97.1 ± 5.2	94.3 ± 9.5
	Sad	89.9±11.4	91.3±12.3	91.9±11.7	89.9±13.9
Response Time (milliseconds)					
	Angry	835 ± 366	849 ± 448	787 ± 336	855 ± 412
	Happy	753 ± 274	745 ± 318	736 ± 313	749 ± 309
	Sad	802 ± 361	825 ± 383	830 ± 355	821 ± 362

3.2 PCA results and selection of ERP components

The PCA on emotional faces revealed six temporal factors that passed the Scree test (96% of total variance) and six spatial factors on each temporal component (76% of the variance). Thirteen T-S components, each accounting for at least .5% of unique variance in the EEG data, were retained (see Appendix). Condition effects were evident in three temporal factors (Figure 2). The first factor (1.3% of total variance)

was characterized by a bilateral posterior response (12 channels: P6, P7, P09, P07, P03, POz, P04, P08, PO10 O1, Oz, O2, maximal at 02) peaking at 172ms. Visual inspection of the waveform shows a positive shift between ~150-250ms post-onset of the face. The PCA shows that this was a *positive* deflection with latency and distributional properties resembling the occipitotemporal P200 component (Schweinberger & Neumann, 2016). A second temporal factor (5.0% of total variance) peaked at 380ms at prefrontal recording sites (Fp1, Fp2, AF8, maximal at Fp1). This component was defined by a broad slow negative wave to angry faces lasting from ~200-750ms. Although we expected contextualized faces to engage semantic/ associative processes in the 350-700ms latency window (e.g., modulating the N400 or LPP), the prefrontal topography, protracted evolution, and specificity of this component to angry faces imply that distinct neural mechanisms were at play. We explored these data as the *Frontal Negativity* factor.

Inspection of the third temporal factor reveals a delayed and sustained positivity beginning ~450-500ms post-onset of the face until the end of the analysis window, peaking at 940ms, consistent with the Late Positive Potential. This temporal factor was divided into two spatially independent components: a centro-posterior response (CP4, P5, P3, P1, Pz, P2, P4, PO7, PO3, POz, PO4, O1, Oz, O2, maximal at P3, 8.6% total variance) elicited by each facial expression; and a prefrontal response (Fp1, AF8, 4.2% of total variance) elicited by sad and happy faces in relation to our vocal conditions. While temporally misaligned and distinct in polarity, the Frontal Negativity (to angry faces) and the frontal component in the LPP temporal factor (to happy and sad faces) exhibited nearly identical spatial properties. We modelled the

effects of vocal expressions (Congruency, Form) on each ERP component identified by the T-S PCA (P200, Frontal Negativity, LPP) separately for each facial expression.[3]

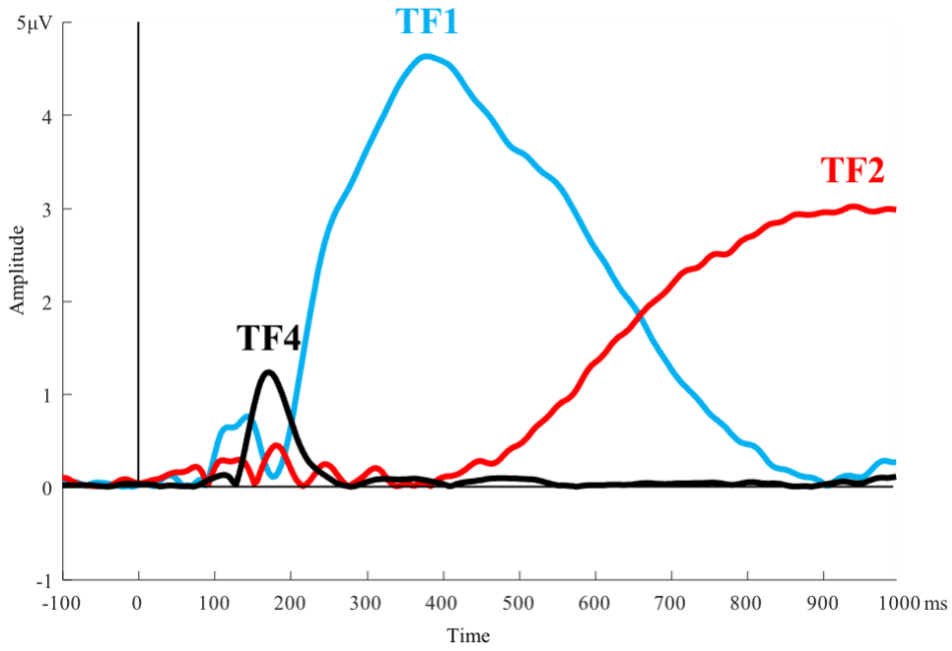


Figure 2. Temporal factors which demonstrated a significant effect of vocal primes during facial expression processing based on temporal-spatial principal component analysis of latency-corrected ERP data.

3.2.1 P200 – Overall, emotional speech increased the P200 to each facial expression when compared to vocalizations (Form: $F_{ANGRY}(1, 1025) = 71.99$, $p<.0001$; $F_{SAD}(1,$

[3] While current hypotheses are centred on emotional faces, we re-ran the t-s PCA including non-emotional expressions as a fourth face type to compare results when all targets in the experiment were included. The PCA produced seven temporal and six spatial factors per temporal component, accounting for 97% and 77% of total variance, respectively. Similar to what is reported for emotional faces, 12 unique T-S components were defined by three major temporal factors, peaking at 176ms, 376ms, and 940ms.

1025) = 54.40, p<.0001; F_{HAPPY}(1, 1025) = 130.36, p<.0001). Voice-face congruency also modulated the P200 amplitude (F_{ANGRY}(1, 1025) =4.81, p=.03; F_{SAD}(1, 1025)= 20.06, p<.0001, but F_{HAPPY}(1, 1025) =2.35, p=.13, *ns*), often in combination with the voice's form (Congruency x Form: F_{ANGRY}(1, 1025)=29.70, p<.0001; F_{HAPPY}(1, 1025) = 11.22, p=.0008). In general, the P200 to incongruent faces was larger than to congruent faces, a pattern that was consistent when participants heard vocalizations but more variable in the speech condition. Effects of the vocal expression's form on P200 amplitude (speech > vocalization) were greater when the face was contextually congruent vs. incongruent (Figure 3).

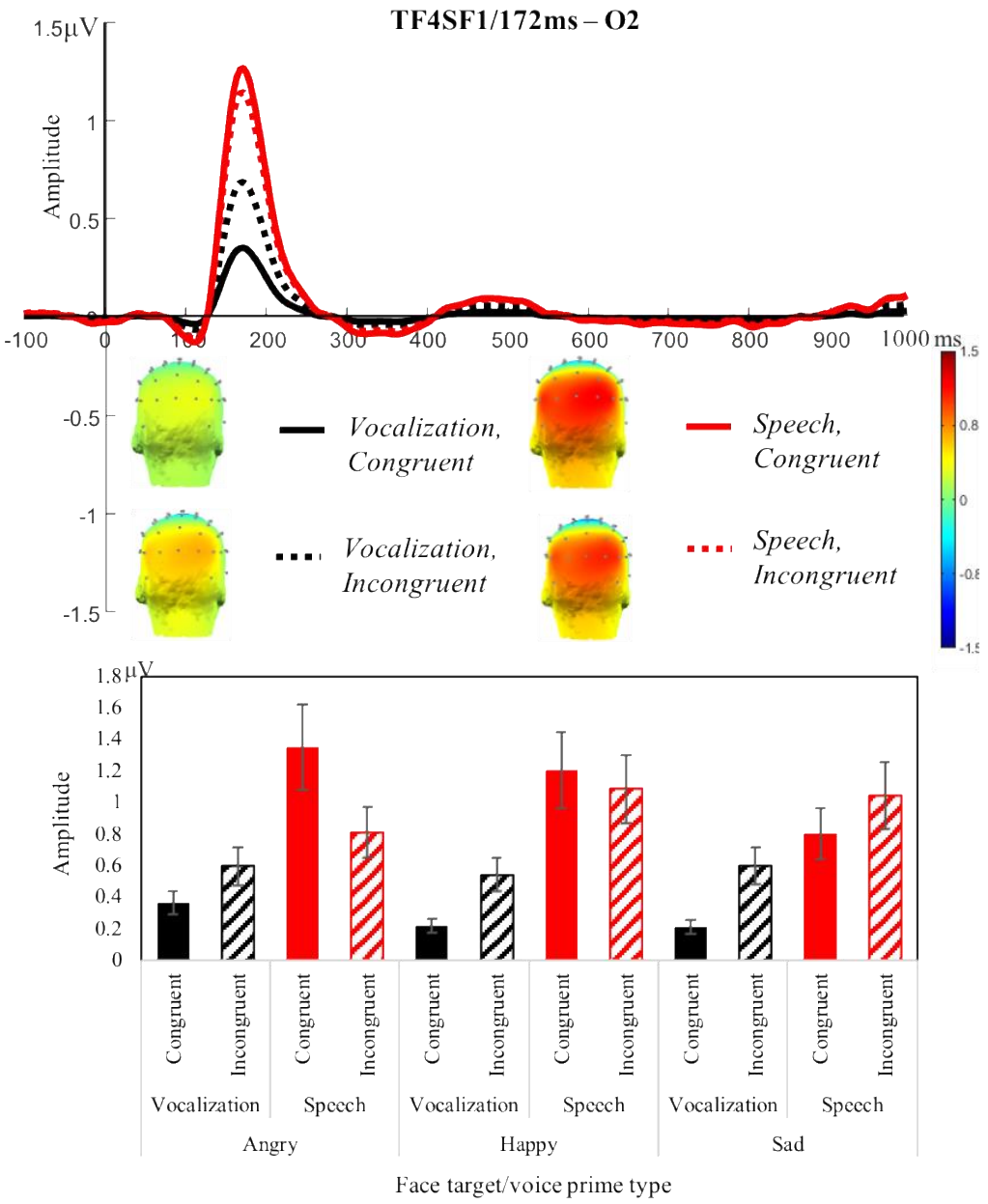


Figure 3. Effects of vocal expression congruency and form on the P200 amplitude evoked by face targets at occipital-temporal channels.

3.2.2 Frontal Negativity – Analysis of the frontal negativity (peaking at 386ms) indicated that this component was uniquely sensitive to the *form* of the preceding vocal expression only when angry targets were presented. Angry faces displayed an

increased negativity following emotional speech than vocalizations (Vocal Form: $F_{ANGRY}(1, 233) = 6.55$, p=.01, Figure 4). The impact of vocal Form on the other facial expressions was not significant ($F_{HAPPY}(1, 233) = .47$, p=.50; $F_{SAD}(1, 233) = .28$, p=.60). The Frontal Negativity was not influenced by contextual Congruency in any way (p's>.15).

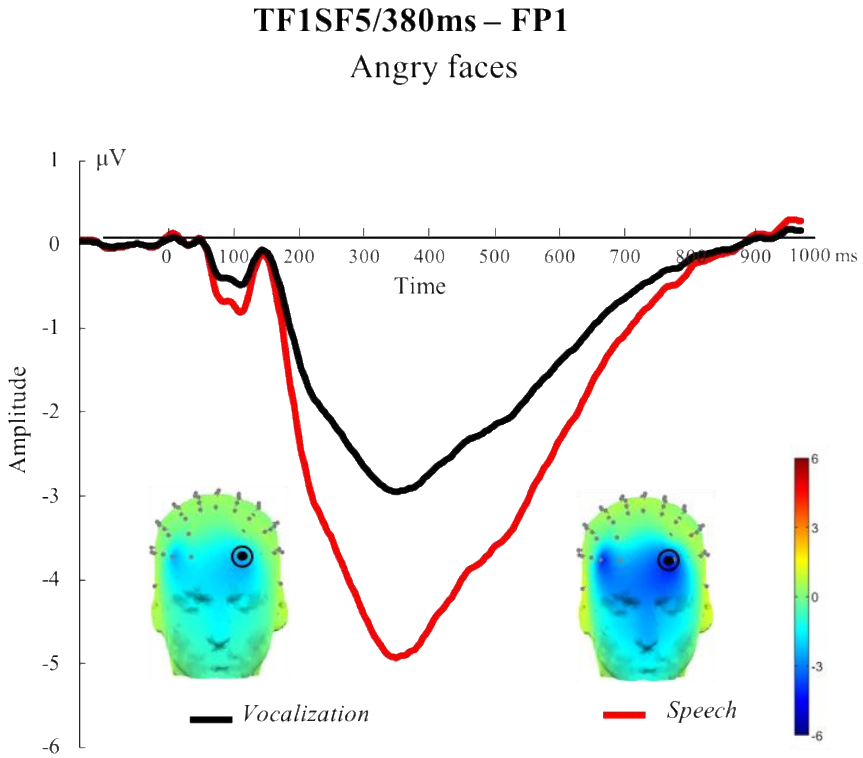


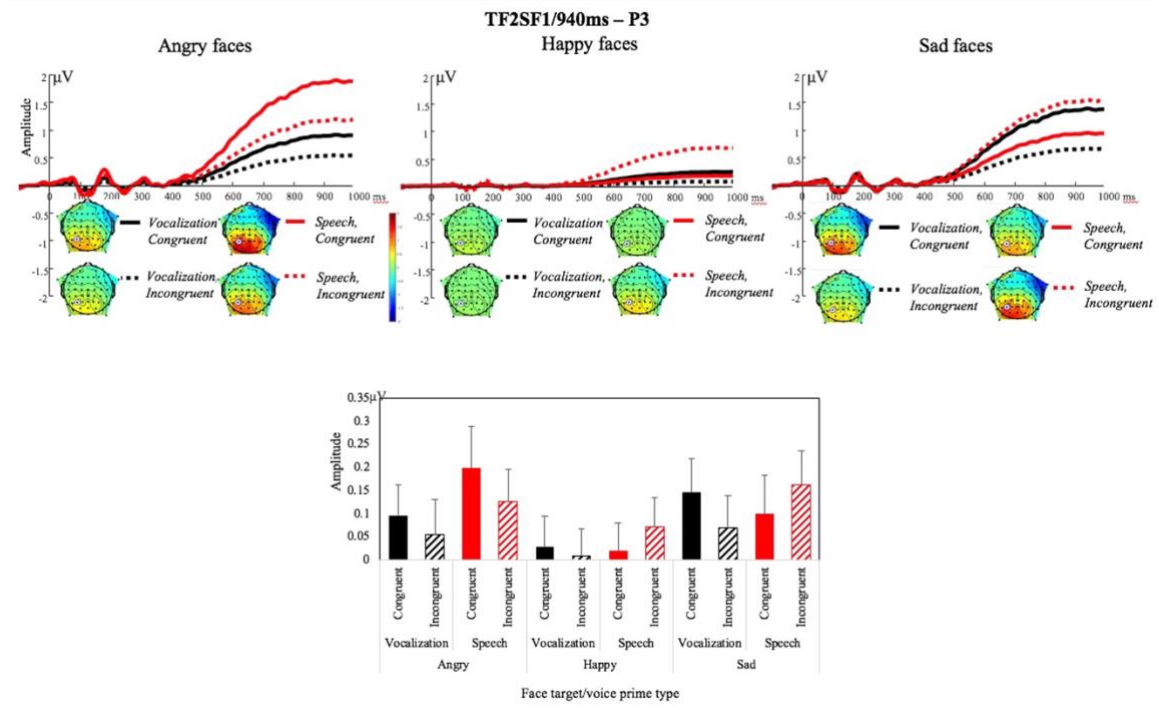
Figure 4. Modulation of a prefrontal negativity evoked by angry faces according to the form of a preceding vocal expression.

3.2.3 Late Positive Potential (LPP) – Within the third temporal factor (peaking at 940ms), the LPP could be decomposed into two independent spatial factors: a posterior (centro-parietal) component and an anterior (prefrontal) component.

Analysis of the posterior LPP revealed interactive effects of vocal Congruency and Form on each facial expression ($F_{ANGRY}(1, 1025) = 4.39$, p=.03; $F_{SAD}(1, 1025) = 91.76$, p<.0001; $F_{HAPPY}(1, 1025) = 32.98$, p<.0001). Incongruent faces produced a larger, sustained positive-going response when participants heard speech than vocalizations (angry: b=.05, t=7.76, p<.0001; sad: b=.07, t=9.24, p<.0001; happy: b=.05, t=8.39, p<.0001). The posterior LPP to congruent faces did not systematically vary by Form of vocal expression (angry congruent: speech>vocalization; sad congruent: vocalization>speech; happy congruent: *ns*). Congruent faces increased the LPP over incongruent faces when they followed vocalizations (angry: b=-.30, t=-4.58, p<.0001; sad: b= -.06, t=-8.14, p<.0001; happy: b= -.01, t=-2.38, p=.02) and angry speech, but this pattern was reversed for sad and happy faces primed by speech (incongruent > congruent, sad: b= .05, t= 5.83, p<.0001; happy: b= .04, t=6.71, p<.0001). These patterns can be seen in Figure 5a (top panel).

Analysis of the anterior LPP response showed that irrespective of congruency, sad faces produced a larger positive-going response following speech than vocalizations (Form: $F_{SAD}(1, 145) = 7.33$, p=.008). For happy faces, the frontal LPP varied in a more complicated manner by vocal Congruency and Form ($F_{HAPPY}(1, 145) = 4.38$, p<.05). Incongruent happy faces displayed a stronger positive shift following vocalizations than speech (b= -.24, t= -2.90, p=.005). Also, following emotional speech the frontal LPP exhibited a more positive deflection for congruent vs. incongruent

a)



b)

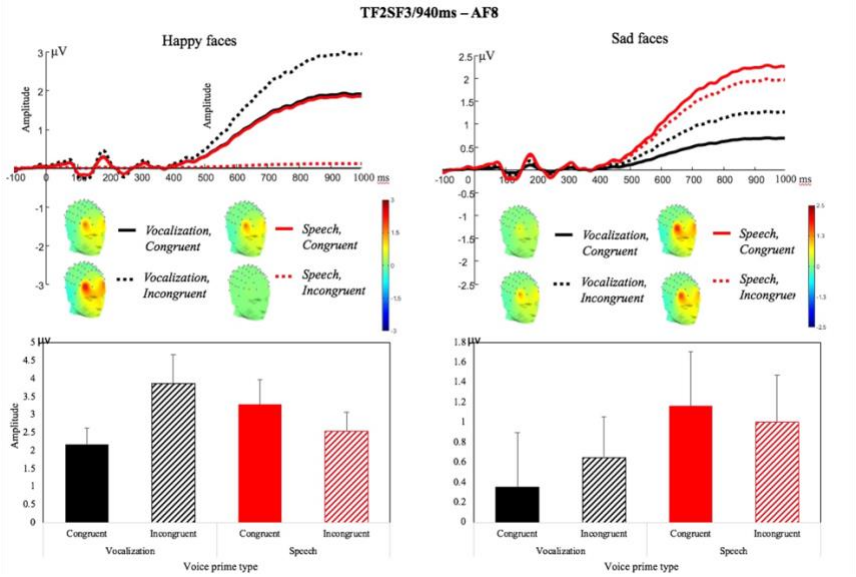


Figure 5. Effects of vocal primes on the Late Positive Potential evoked by a) each face target at posterior channels; and b) happy and sad face targets at prefrontal channels.

happy faces (b= -.15, t= -2.50, p=.01). Vocal features did not significantly modulate the frontal LPP for angry faces in this latency range. The impact of vocal expression

form and congruency on the anterior LPP response is shown in Figure 5b (bottom panel).

3. Discussion

In this study, participants evaluated the emotional status (*prototypicality*) of facial expressions, a process that is influenced by the relationship of a preceding voice (Jaywant & Pell, 2012; Pell, 2005a; Pell et al., 2011). Decisions about congruent vs. incongruent faces were faster and more accurate, providing new evidence that voices facilitate processing of facial expressions that share the same emotional quality (Aguado et al., 2019; Dozolme et al., 2015; Herring, Taylor, White, & Crites, 2011; Hietanen & Astikainen, 2013b; Paulmann & Pell, 2010). Interestingly, the *form* of vocal expression (speech, vocalization) did not influence facial affect decisions. This implies that on-line effects of voice form witnessed at the cortical level had little impact when decisions about the meaning of the face target were actually executed (~ 800ms post-onset of the face).

Using objective methods that avoided assumptions about the impact of vocal primes on face-related ERPs (Paulmann & Pell, 2010), our data show that the neurophysiological response to faces was differentiated in three consecutive latency ranges: first, vocal expression congruency *and* type modulated the occipitotemporal P200 (150-250ms) to all three emotional faces; then, there was an effect of vocal expression form (but not congruency) on angry faces, which evoked a slow prefrontal negativity in the 200-750ms time interval; and at a final stage, combined differences in congruency and voice form altered the LPP response to each facial expression

(sustained central-posterior positivity). The posterior LPP was accompanied by a prefrontal positivity evoked by sad and happy faces in the same temporal factor (450-1000ms). These results show that cortical face processing is sensitive to attributes of a preceding voice at early structure-building phases (P200) and late stages of meaning elaboration (LPP), establishing that a broader set of neurocomputations underlie voice-face effects than those first described by Paulmann and Pell (2010).

4.1 Early effects: Perception of face typicality (P200)

One of the novel contributions of our study is to demontrate that emotional voices modulate structure-building phases of cortical face processing, here the perception of *second-order* spatial features. When a face appears, visual-perceptual processes decode its spatial characteristics which can lead to recognition (i.e., activation of familiar faces stored in long-term memory, Burton, Jenkins & Schweinberger, 2011). In parallel, perceptual operations decode emotional information beginning ~120ms post-onset of the face (Vuilleumier & Pourtois, 2007). Most research on how context affects face perception have examined the N170 component, an index of detection/ holistic representation of a face (Hinojosa, Mercado, & Carretié, 2015). Sequential priming studies report that visual and situational context modulate demands on holistic structure-building processes, increasing the N170 amplitude to emotionally incongruent vs. congruent faces (Diéguez-Risco et al., 2015; Hietanen & Astikainen, 2013; Righart & de Gelder, 2008; cf. Aguado et al., 2019 for a critical analysis). For our task, the PCA uncovered no evidence that vocal primes modulated the N170; rather, this effect was observed on the subsequent P200, a spatially overlapping positivity

which encodes "second-order" spatial relations once a face has been detected (Latinus & Taylor, 2006; Schweinberger & Neumann, 2016).

The face-sensitive P200 gauges the distinctiveness or *typicality* of unfamiliar faces, a dissociable process that encodes a face's perceptual similarity to learned exemplars within a multidimensional psychological 'face-space' (Valentine, Lewis, & Hills, 2016; Wuttke & Schweinberger, 2019). P200 amplitudes are modulated when people view spatially distorted facial caricatures (Kaufmann & Schweinberger, 2012) and when faces blend emotional features from two expressions (e.g., a happy mouth and angry eyes, Calvo, Marrero, & Beltrán, 2013). Recent data demonstrate that P200 amplitudes evoked by unfamiliar faces parametrically increase according to their perceptual similarity to a prototype, their 'distance-to-norm' in hypothesized face-space (Wuttke & Schweinberger, 2019). Along these lines, the P200 increases when people view same- vs. other-race faces (Stahl, Wiese, & Schweinberger, 2008; Tortosa, Lupianez, & Ruz, 2013), given that other-race faces are experienced less often and are more perceptually distant to the observer.

While P200 studies rarely draw participants' attention to facial *expressions* (as opposed to identity features, cf. Calvo et al., 2013), our task appears to place high demands on processes for analyzing individual second-order spatial information during *emotional* analysis of a face. Facial affect decisions require participants to compare the metric layout of emotional and non-emotional faces—e.g., natural variations in lip shape, eye, and brow movements—to a mental prototype of how emotions are typically expressed (Paulmann & Pell, 2009; Pell, 2005a). Arguably, perceptual learning about how facial emotion is communicated across individuals,

and in different social contexts reflects an unexplored dimension of variability in face-space that alters similarity/prototypicality judgements (Valentine et al., 2016), one that is uniquely engaged by our task. Alternatively, by drawing attention to second-order visual features diagnostic of a face's emotional status, our task may have highlighted perceptual variability in the *identity* of the posers, which included individuals from various ethnic and racial backgrounds; these conditions would tax P200- (rather than N170-) related neural mechanisms (Wuttke & Schweinberger, 2019). While these ideas await validation, our paradigm may be ideally suited for probing visual-perceptual operations underlying the P200 in an emotional context.

The observation that P200 amplitudes increased for incongruent faces and following speech-embedded emotions supplies the first evidence that early (pre-categorical) operations are sensitive to the emotional quality of a preceding voice *and* its form of acoustic expression. These results underscore the impact of vocal expressions on visual perception and attentional orienting to faces (Liu, Rigoulot, Jiang, Zhang, & Pell, 2021; Paulmann, Titone, & Pell, 2012; Rigoulot & Pell, 2014; Schirmer, Wijaya, Wu, & Penney, 2019). They also suggest that emotional voices shape early face structure-building procedures beyond the extraction of primary spatial information (N170, Diéguez-Risco et al., 2015; Righart & de Gelder, 2008), modulating the extent to which people assess secondary facial characteristics relevant to perception. Interestingly, the P200 peaked earlier in our data (172ms post-onset of the face) than reported elsewhere; latency differences may reflect our participants' novel attention to *emotional* properties of a face, which could modulate P200-related perceptual processes more rapidly than other forms of typicality

judgements based on caricatures or schematic faces (Kaufmann & Schweinberger, 2012; Schulz et al., 2012).

4.2 Mid-Late effects: Intrinsic and external effects on meaning processes

Another major contribution of our report is to show that neurophysiological activity reflecting how face meaning is incrementally generated is sensitive to the congruency *and* form of a preceding vocal expression. The Late Positive Potential (LPP) is highly sensitive to intrinsic properties of an emotionally arousing target (face, picture) and external effects of its context (Hajcak, MacNamara, & Olvet, 2010). In our study, the congruency of vocal expressions *and* cues about their form (speech, vocalization) produced sustained differences in the LPP; this effect started 450-500ms post-onset of the face and endured beyond our analysis window (peaking at 940ms).

At posterior scalp regions, the LPP tended to increase for incongruent faces (Diéguez-Risco et al., 2015b; Dozolme et al., 2015; Stolz et al., 2019) and following speech-embedded emotions (compared to vocalizations), with each face target eliciting distinct patterns of brain activity as a function of the preceding voice in the mid-late latency range (review Figure 5). [4] Increased LPP amplitudes reflect difficulties *integrating* details about a face in working memory with preceding events, causing sustained motivated attention to the target and increased cognitive effort (Diéguez-Risco et al., 2015b; Schupp et al., 2006). When contextual features increase

[4] There was no evidence that our prime manipulations modulated *access* to emotional face representations in semantic memory; such differences would produce a centroparietal negativity (N400-like effect) in the early phase of the LPP latency range, which was not observed for our paradigm (cf. Paulmann & Pell, 2010).

uncertainty about the conceptual relevance of a target (face, word) or highlight *implausible* stimulus relationships, late positive brain potentials (LPP, LPC, or P600-like effects) are evoked at posterior midline regions of the scalp. These deflections gauge prolonged attempts by the cognitive system to synthesize or bind information about the input within its embedded context (Kuperberg, Brothers, & Wlotko, 2019; Kutas & Federmeier, 2011), part of a domain-general attentional reorienting process (Sassenhagen & Bornkessel-Schlesewsky, 2015). Differences observed here in the LPP imply that incongruent voices and speech-embedded expressions both promoted uncertainty about the target and hampered processes for integrating voice-face meaning over an extended timeframe, beginning at the P200 stage. However, broader inspection of the late ERP effects suggests that additional mechanisms were at play, and that analysis of voice-face meaning engaged multiple operations associated with different onsets and distinct sources of brain activity.

4.2.1 Neurocomputational divisions in the Late Positive Potential

The LPP factor was composed of synchronized positive deflections at posterior and prefrontal recording sites (Aguado et al., 2019; Dozolme et al., 2015; Moratti, Saugar, & Strange, 2011). Although conventionally measured at posterior scalp regions, the LPP is increasingly conceived as having functionally distinct temporal and/or spatial divisions which produce overlapping effects along the midline of the brain (Foti, Hajcak, & Dien, 2009; MacNamara, Foti, & Hajcak, 2009; Moratti et al., 2011). This claim is supported by objective, whole-brain analysis of our ERP data, which confirmed that the posteriorly distributed LPP was temporally coupled with an

independent frontal LPP component (for similar PCA-based findings when participants viewed emotional pictures, see Foti et al., 2009; MacNamara et al., 2009).

Effects of emotional voices on the posterior versus frontal LPP response to faces were unique and did not affect each target face in the same way (vocal modulation of the posterior LPP was observed for all faces, whereas modulation of the frontal component was restricted to sad and happy targets). According to predictive coding accounts of how anticipatory processes influence bottom-up processing of sensory input (Bornkessel-Schlesewsky & Schlesewsky, 2019), cortical divisions in the LPP could reflect neural mechanisms for semantic integration (posterior LPP) versus semantic *prediction* (frontal LPP) as a representation of face meaning is incrementally constructed. Along similar lines, neuroimaging studies have linked bottom-up processes for low-level perception and integration of sensory input (speech, faces) to a posterior (temporally based) cortical network, whereas top-down predictions about these events engage prefrontal mechanisms within an anterior (frontal-parietal) network (Mechelli et al., 2004; Zekveld et al., 2006).

The frontal LPP in our data shared important elements with an earlier *negative* potential evoked by angry faces (beginning 200ms prior to the LPP onset). This large negative deflection was modulated by the physical form (not congruency) of a preceding voice, which markedly increased when angry faces were primed by speech compared to vocalizations. The frontal negativity to anger (200-750ms) and the frontal LPP to sad and happy faces (450-1000ms) displayed the same scalp topography in the prefrontal cortex; moreover, the frontal response to angry and sad faces was only sensitive to the form of a vocal expression (speech > vocalization).

These data furnish new evidence that the brain encodes speech and vocalizations as distinct classes of acoustic events (Pell et al., 2015) and that details about *how* a vocal expression is produced are relevant when sensing a face (Bryant, 2021). They also permit arguments that while the two frontally distributed cortical responses we observed were temporally independent and qualitatively distinct, they shared functional mechanisms linked to *expectancies* about the face target.

4.2.2 Early, enhanced processing of angry faces

Compared to other expressions, angry faces rapidly capture exogenous attention due to their association with biological threat (Dimberg & Öhman, 1996; Mogg & Bradley, 1999), which enhances perceptual (bottom-up) processing of the input. This factor could explain why the *onset* of the neural response to angry faces occurred earlier (~200ms) than the frontal LPP to sad and happy targets (Rellecke et al., 2012; Stolz et al., 2019). Arousing emotional stimuli are known to increase the EPN in the 150-350ms latency range post-onset of a face (Schupp & Kirmse, 2021), whereas here, angry targets evoked a prefrontal response with a protracted time course. Still, the *origin* of the frontal negativity is likely to reflect intrinsic capturing of attention by highly arousing negative faces in our dataset, as cognitive resources are diverted away from the prime to assess threat-related properties of the target more deeply. It can be further argued that angry faces would have selectively taxed regulative control processes to inhibit automatic action tendencies associated with threatening targets (Folstein & Van Petten, 2008; Larson, Clayson, & Clawson, 2014), to allow participants in our experiment to adjust their behavioral responses to meet

task demands (i.e., to categorize the status of the facial expression irrespective of its emotion). These executive control mechanisms involve contributions of the anterior cingulate cortex and produce early negative deflections at prefrontal channels in the ERP (Bush, Luu, & Posner, 2000; Carretié, Albert, López-Martín, & Tapia, 2009), a possible explanation for the prolonged time course and distinct polarity of the cortical response to angry faces in the second temporal factor.

Our finding that only the *form* of a vocal prime modulated the anger-related frontal negativity (speech > vocalization) is instructive and without precedence in the face processing literature. Slow negative brain potentials have been linked to anticipatory attention (León-Cabrera, Flores, Rodríguez-Fornells, & Morís, 2019) as well as semantic processes for contextually (re)interpreting *unexpected but plausible* events (Coulson & Kutas, 2001; Wlotko & Federmeier, 2012). This activity reflects the cognitive system's attempt to "shift frames" and reassess the relevance of prior information when faced with a plausible but unanticipated continuation of an event. Along these lines, patterns in the neurophysiological data suggest that while angry faces were subject to enhanced perception and monitoring (due to intrinsic threat), they may also have been less *expected* following a speech stimulus than a vocalization (of whatever emotional quality). According to predictive coding models (Bornkessel-Schlesewsky & Schlesewsky, 2019), failure to anticipate angry targets after listening to speech would cause a large prediction error (negative ERP deflection), with costs on target processing as a generative model of the face's contextual relevance is updated. Similar operations might explain why sad targets engaged prefrontal mechanisms more extensively following speech than vocalizations but at a later time

point (review Figure 5b). While unproven, we speculate that these effects refer to different expectancies and learned constraints in the use of negative facial expressions during speech behaviour (see below). Note that since speech and vocalization primes were identical in duration, effects of voice form on the prefrontal response could not be explained by relative differences in the duration of primes for the three types of emotion targets.

4.2.3 How vocal context constrains face meaning in social settings

A potentially useful way to characterize the mid-late effects is in terms of the amount of *constraint* imposed by speech versus vocalizations when particular facial expressions are encountered. As acoustically primitive, unambiguous emotional signals, vocalizations may create a strong expectancy that the voice will be accompanied by an explicit emotional response in the face (e.g., by means of general arousal facilitation between modalities, Hinojosa et al., 2009). In contrast, interpersonal experience could dictate that emotional speech is associated with many forms of social reactions or possible continuations, and that people often mask or conceal their true feelings during speech communication (especially strong negative facial reactions, Liu et al., 2015).

Individual experience could mean that negative (especially angry) faces are highly unexpected when coupled with speech; resulting (mis)predictions produce a late frontally distributed negative or positive deflection in the ERP, depending on the intrinsic salience of the facial expression. In our study, this reinforces an important distinction between sequential expectancy effects (frontal) and semantic integration

processes (posterior) in the LPP latency period (see Dien et al., 2010 for a conceptually similar division of the N400/P400 effect). These concepts, which are well-developed in the psycholinguistic literature (Bornkessel-Schlesewsky & Schlesewsky, 2019; Brothers, Swaab, & Traxler, 2015; Kuperberg et al., 2019; Thornhill & Van Petten, 2012), are rarely applied to studies looking at priming or context use in broader aspects of social communication. Merging ideas across these literatures could prove useful as we define how prior knowledge and events dynamically shape face meaning, allowing further specificity of computational divisions in the LPP latency range.

If indeed late frontal effects reflect cases of contextually unanticipated targets, it is curious that the cortical response increased when happy faces were incongruent and followed a *vocalization* (thus showing reverse sensitivity to the form of vocal expression). If replicated, this would signify that happy faces are unexpected after listeners hear negative vocal sounds (crying or growls), but they are a more logical continuation after someone speaks in a negative tone of voice. Interestingly, positive facial expressions are often retained during conversations as a sign of support and affiliation with the speaker, even when vocal messages are negative (Crivelli & Fridlund, 2018; Van Kleef, 2009). As this research progresses, we can test whether individual experience with how particular facial expressions are coupled with emotional sounds alters expectancies we form in social communication settings. Additional work could then examine how the neurocognitive system adjusts to voice-face temporal sequencing effects in combination with situational knowledge held by the perceiver. These efforts will move us closer to an understanding of how facial

expressions are perceived and interpreted in the natural soundscape of human interactions.

4. Acknowledgements

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to Marc D. Pell (RGPIN-2016-04373).

CRediT author statement:

Pell, Marc D.: Conceptualization, Methodology, Resources, Writing – Original draft, Visualization, Supervision, Project administration, Funding acquisition. **Sethi, Sameer**: Conceptualization, Methodology, Investigation, Formal analysis, Writing- Original draft. **Rigoulot, Simon**: Conceptualization, Methodology, Investigation, Formal analysis, Writing – Review & Editing, Supervision. **Rothermich, Kathrin:** Formal analysis, Writing – Review & Editing. **Liu, Pan**: Formal analysis. **Jiang, Xiaoming:** Conceptualization, Methodology, Formal analysis, Writing- Original draft, Visualization.

5. References

Agnew, Z. K., Ward, L., McGettigan, C., Josephs, O., & Scott, S. K. (2017). Distinct neural systems for the production of communicative vocal behaviors. *BioRxiv*. https://doi.org/10.1101/107441

Aguado, L., Parkington, K. B., Dieguez-Risco, T., Hinojosa, J. A., & Itier, R. J. (2019). Joint modulation of facial expression processing by contextual congruency and task demands. *Brain Sciences*, *9*(5), 1–20. https://doi.org/10.3390/brainsci9050116

Baart, M., & Vroomen, J. (2018). Recalibration of vocal affect by a dynamic face. *Experimental Brain Research*, *0*(0), 1–8. https://doi.org/10.1007/s00221-018-5270-y

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286–290. https://doi.org/10.1177/0963721411422522

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531–539. https://doi.org/10.3758/BRM.40.2.531

Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology, 10:*298. Doi: 10.3389/fpsyg.2019.00298.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Bryant, G.A. (2021). The evolution of human vocal emotion. *Emotion Review*, *13*(1), 25-33. Doi: 10.1177/1754073920930791.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. http://doi.org/10.5334/joc.10

Bülthoff, I., & Newell, F. N. (2017). Crossmodal priming of unfamiliar faces supports early interactions between voices and faces in person perception. *Visual Cognition*, *25*(4–6), 611–628. https://doi.org/10.1080/13506285.2017.1290729

Burton, A.M., Jenkins, R. & Schweinberer, S.R. (2011) Mental representations of familiar faces. *British Journal of Psychology, 102,* 943-958. https://doi.org/10.1111/j.204408295.2011.02039.x.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–221.

Calvo, M. G., Marrero, H., & Beltrán, D. (2013). When does the brain distinguish between genuine and ambiguous smiles? An ERP study. *Brain and Cognition*, *81*(2), 237–246. https://doi.org/10.1016/j.bandc.2012.10.009

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Carretié, L., Albert, J., López-Martín, S., & Tapia, M. (2009). Negative brain: An integrative review on the neural processes activated by unpleasant stimuli. *International Journal of Psychophysiology*, *71*(1), 57–63. https://doi.org/10.1016/j.ijpsycho.2008.07.006

Clark, G. M., McNeel, C., Bigelow, F. J., & Enticott, P. G. (2020). The effect of empathy and context on face-processing ERPs. *Neuropsychologia*, *147*(August), 107612. https://doi.org/10.1016/j.neuropsychologia.2020.107612

Coulson, S., & Kutas, M. (2001). Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters*, *316*(2), 71–74. https://doi.org/10.1016/S0304-3940(01)02387-4

Crivelli, C., & Fridlund, A. J. (2018). Facial Displays Are Tools for Social Influence. *Trends in Cognitive Sciences*, *22*, 388–399. https://doi.org/10.1016/j.tics.2018.02.006

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(August 2011), 289–311. https://doi.org/10.1080/026999300378824

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*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Diéguez-Risco, T., Aguado, L., Albert, J., & Hinojosa, J. A. (2015). Judging emotional

congruency: Explicit attention to situational context modulates processing of facial expressions of emotion. *Biological Psychology*, *112*, 27–38. https://doi.org/10.1016/j.biopsycho.2015.09.012

Dien, J. (2006). Progressing towards a consensus on PCA of ERPs. *Clinical Neurophysiology*. Dien and Frishkoff. https://doi.org/10.1016/j.clinph.2005.09.029

Dien, J. (2010). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods*, *187*(1), 138–145. https://doi.org/10.1016/j.jneumeth.2009.12.009

Dien, J. (2012). Applying principal components analysis to event-related potentials: a tutorial. *Developmental Neuropsychology*, *37*(6), 497–517.

Dien, J., Michelson, C.A., & Franklin, M.S. (2010). Separating the visual sentence N400 effect from the P400 sequential expectancy effect: Cognitive and neuroanatomical implications. *Brain Research, 1355*, 126-140. Doi: 10.1016/j.brainres.2010.07.099.

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*, *20*(2), 149–182. https://doi.org/10.1007/BF02253869

Dozolme, D., Brunet-Gouet, E., Passerieux, C., & Amorim, M. A. (2015). Neuroelectric correlates of pragmatic emotional incongruence processing: Empathy matters. *PLoS ONE*, *10*(6), 1–20. https://doi.org/10.1371/journal.pone.0129770

Duval, E. R., Moser, J. S., Huppert, J. D., & Simons, R. F. (2013). What's in a face?: The late positive potential reflects the level of facial affect expression. *Journal of Psychophysiology*, *27*(1), 27–38. https://doi.org/10.1027/0269-8803/a000083

Eimer, M., & Holmes, A. (2002). An ERP study on the time course of emotional face processing. *Neuroreport*, *13*(4), 427–431. https://doi.org/10.1097/00001756-200203250-00013

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., Reiterer, S., Grodd, W., & Wildgruber, D. (2006). Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping, 27*, 707-714.

Föcker, J., & Röder, B. (2019). Event-related potentials reveal evidence for late integration of emotional prosody and facial expression in dynamic stimuli: An ERP study. *Multisensory Research*, *32*(6), 473–497. https://doi.org/10.1163/22134808-

20191332

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology*, *45*(1), 152–170. https://doi.org/10.1111/j.1469-8986.2007.00602.x

Foti, D., Hajcak, G., & Dien, J. (2009). Differentiating neural responses to emotional pictures: Evidence from temporal-spatial PCA. *Psychophysiology*, *46*(3), 521–530. https://doi.org/10.1111/j.1469-8986.2009.00796

Garrido-Vásquez, P., Pell, M. D., Paulmann, S., & Kotz, S. A. (2018). Dynamic Facial Expressions Prime the Processing of Emotional Prosody, *12*(June), 1–11. https://doi.org/10.3389/fnhum.2018.00244

Gerdes, A., Wieser, M., Bublatzky, F., Kusay, A., Plichta, M. & Alpers, G. (2013). Emotional sounds modulate early neural processing of emotional pictures. *Frontiers in Psychology,* 4:741. Doi: 10.3389/fpsyg.2013.00741.

Hajcak, G., Dunning, J. P., & Foti, D. (2009). Motivated and controlled attention to emotion: Time-course of the late positive potential. *Clinical Neurophysiology*, *120*(3), 505–510. https://doi.org/10.1016/j.clinph.2008.11.028

Hajcak, G., MacNamara, A., & Olvet, D. M. (2010). Event-related potentials, emotion, and emotion regulation: an integrative review. *Developmental Neuropsychology*, *35*(2), 129–155. https://doi.org/10.1080/87565640903526504

Hawk, S. T., van Kleef, G. a, Fischer, A. H., & van der Schalk, J. (2009). "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion*, *9*(3), 293–305. https://doi.org/10.1037/a0015178

Herring, D. R., Taylor, J. H., White, K. R., & Crites, S. L. (2011). Electrophysiological Responses to Evaluative Priming: The LPP Is Sensitive to Incongruity. *Emotion*, *11*(4), 794–806. https://doi.org/10.1037/a0022804

Hess, U., Dietrich, J., Kafetsios, K., Elkabetz, S., & Hareli, S. (2020). The bidirectional influence of emotion expressions and context: emotion expressions, situational information and real-world knowledge combine to inform observers' judgments of both the emotion expressions and the situation. *Cognition and Emotion*, *34*(3), 539–552. https://doi.org/10.1080/02699931.2019.1651252

Hietanen, J. K., & Astikainen, P. (2013). N170 response to facial expressions is

modulated by the affective congruency between the emotional expression and preceding affective picture. *Biological Psychology*, *92*, 114–124. https://doi.org/10.1016/j.biopsycho.2012.10.005

Hinojosa, J. A., Mercado, F., & Carretié, L. (2015). N170 sensitivity to facial expression: A meta-analysis. *Neuroscience and Biobehavioral Reviews*, *55*, 498–509. https://doi.org/10.1016/j.neubiorev.2015.06.002

Hinojosa, José a, Carretié, L., Méndez-Bértolo, C., Míguez, A., & Pozo, M. a. (2009). Arousal contributions to affective priming: electrophysiological correlates. *Emotion*, *9*(2), 164–171. https://doi.org/10.1037/a0014680

Ho, H. T., Schröger, E., & Kotz, S. A. (2015). Selective Attention Modulates Early Human Evoked Potentials during Emotional Face-Voice Processing. https://doi.org/10.1162/jocn_a_00734

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Jaywant, A., & Pell, M. D. (2012). Categorical processing of negative emotions from speech prosody. *Speech Communication*, *54*(1), 1–10. https://doi.org/10.1016/j.specom.2011.05.011

Jessen, S., & Kotz, S. a. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, *58*(2), 665–674. https://doi.org/10.1016/j.neuroimage.2011.06.035

Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex*, *66*. https://doi.org/10.1016/j.cortex.2015.02.002

Kashyap, R., Ouyang, G., Sommer, W., & Zhou, C. (2016). Neuroanatomic localization of priming effects for famous faces with latency-corrected event-related potentials. *Brain Research*, *1632*, 58-72. http://dx.doi.org/10.1016/j.brainres.2015.12.001

Kaufmann, J.M. & Schweinberger, S. R. (2012). The faces you remember: Caricaturing shape facilitates brain processes reflecting the acquisition of nrew face representations. *Biological Psychology, 89*(1), 21-33. https://doi.org/10.1016/j.biopsycho.2011.08.011.

Kokinous, J., Kotz, S. A., Tavano, A., & Schröger, E. (2015). The role of emotion in dynamic audiovisual integration of faces and voices. *Social Cognitive and Affective*

Neuroscience, *10*(5), 713–720. https://doi.org/10.1093/scan/nsu105

Krombholz, A., Schaefer, F., & Boucsein, W. (2007). Modification of N170 by different emotional expression of schematic faces. *Biological Psychology*, *76*, 156–162. https://doi.org/10.1016/j.biopsycho.2007.07.004

Kumle, L., Vo, M., Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, https://doi.org/10.3758/s13428-021-01546-0

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2019). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. https://doi.org/10.1162/jocn_a_01465

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP) NIH Public Access. *Annu Rev Psychol*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Larson, M. J., Clayson, P. E., & Clawson, A. (2014). Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. *International Journal of Psychophysiology*, *93*(3), 283–297. https://doi.org/10.1016/j.ijpsycho.2014.06.007

Latinus, M., & Taylor, M. J. (2006). Face processing stages: Impact of difficulty and the separation of effects. *Brain Research*, *1123*(1), 179–187. https://doi.org/10.1016/j.brainres.2006.09.031

Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience*, *11*. https://doi.org/10.1186/1471-2202-11-36

Lavan, N., Lima, C. F., Harvey, H., Scott, S. K., & McGettigan, C. (2014). I thought that I heard you laughing: Contextual facial expressions modulate the perception of authentic laughter and crying. *Cognition and Emotion*, (September), 1–10. https://doi.org/10.1080/02699931.2014.957656

León-Cabrera, P., Flores, A., Rodríguez-Fornells, A., & Morís, J. (2019). Ahead of time: Early sentence slow cortical modulations associated to semantic prediction. *NeuroImage*, *189*(January), 192–201.

https://doi.org/10.1016/j.neuroimage.2019.01.005

Lin, H., & Liang, J. (2019). Contextual effects of angry vocal expressions on the encoding and recognition of emotional faces: An event-related potential (ERP) study. *Neuropsychologia*, *132*(January), 107147. https://doi.org/10.1016/j.neuropsychologia.2019.107147

Liu, P., Rigoulot, S., Jiang, X., Zhang, S., & Pell, M. D. (2021). Unattended emotional prosody affects visual processing of facial expressions in Mandarin-speaking Chinese: A comparison with English-speaking Canadians. *Journal of Cross-Cultural Psychology*. https://doi.org/10.1177/0022022121990897

Liu, P., Rigoulot, S., & Pell, M. D. (2015). Culture modulates the brain response to human expressions of emotion : Electrophysiological evidence. *Neuropsychologia*, *67*, 1–13.

Liu, T., Pinheiro, A., Zhao, Z., Nestor, P. G., McCarley, R. W., & Niznikiewicz, M. A. (2012). Emotional cues during simultaneous face and voice processing: Electrophysiological insights. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0031001

MacNamara, A., Foti, D., & Hajcak, G. (2009). Tell Me About It: Neural Activity Elicited by Emotional Pictures and Preceding Descriptions. *Emotion*, *9*(4), 531–543. https://doi.org/10.1037/a0016251

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215–221.https://doi.org/10.3758/BF03212421

Mechelli, A., Price, C.J., Friston, K.J., & Ishai, A. (2004). Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cerebral Cortex, 14*, 1256-1265. Doi: 10.1093/cercor/bhh087.

Mermillod, M., Grynberg, D., Pio-Lopez, L., Rychlowska, M., Beffara, B., Harquel, S., … Droit-Volet, S. (2018). Evidence of rapid modulation by social information of subjective, physiological, and neural responses to emotional expressions. *Frontiers in Behavioral Neuroscience*, *11*(January), 1–14. https://doi.org/10.3389/fnbeh.2017.00231

Meyer, K., Rostami, H., Ouyang, G., Debener, S., Sommer, W., & Hildebrandt, A. (2021). Mechanisms of face specificity – Differentiating speed and accuracy in face

cognition by event-related potentials of central processing. *Cortex, 134*, 114-133. https://doi.org/10.1016/j.cortex.2020.10.016

Mogg, K., & Bradley, B. P. (1999). Orienting of attention to threatening facial expression presented under conditions of restricted awareness, (August 2011), 37–41. https://doi.org/10.1080/026999399379050

Moratti, S., Saugar, C., & Strange, B. A. (2011). Prefrontal-occipitoparietal coupling underlies late latency human neuronal responses to emotion. *Journal of Neuroscience*, *31*(47), 17278–17286. https://doi.org/10.1523/JNEUROSCI.2917-11.2011

Ouyang, G., Herzmann, G., Zhou, C., & Sommer, W. (2011). Residue iteration decomposition (RIDE): A new method to separate ERP components on the basis of latency variability in single trials. *Psychophysiology*, *48*(12), 1631–1647. https://doi.org/10.1111/j.1469-8986.2011.01269.x

Ouyang, G., Sommer, W., & Zhou, C. (2016). Reconstructing ERP amplitude effects after compensating for trial-to-trial latency jitter: A solution based on a novel application of residue iteration decomposition. *International Journal of Psychophysiology*, *109*, 9–20. https://doi.org/10.1016/j.ijpsycho.2016.09.015

Owren, M.J., Amoss, R.T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, *73*(6), 530-544.

Paulmann, S., & Pell, M. D. (2009). Facial expression decoding as a function of emotional meaning status: ERP evidence. *Neuroreport*, *20*, 1603–1608. https://doi.org/10.1097/WNR.0b013e3283320e3f

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: how much is enough? *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 230–242. https://doi.org/10.3758/CABN.10.2.230

Paulmann, S., Titone, D., & Pell, M. D. (2012). How emotional prosody guides your way: Evidence from eye movements. *Speech Communication*, *54*, 92–107. https://doi.org/10.1016/j.specom.2011.07.004

Pell, M.D. (2002). Evaluation of nonverbal emotion in face and voice: some preliminary findings on a new battery of tests. *Brain and Cognition*, *48*, 499–504.

Pell, M.D. (2005a). Nonverbal emotion priming: Evidence from the "Facial Affect Decision Task." *Journal of Nonverbal Behavior*, *29*(1), 45–73. https://doi.org/10.1007/s10919-004-0889-8

Pell, M.D. (2005b). Prosody-face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior*, *29*(4), 193–215. https://doi.org/10.1007/s10919-005-7720-z

Pell, M.D., & Kotz, S.A. (2011). On the time course of vocal emotion recognition. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027256

Pell, M.D., & Leonard, C. L. (2005). Facial expression decoding in early Parkinson's disease. *Cognitive Brain Research*, *23*, 327–340. https://doi.org/10.1016/j.cogbrainres.2004.11.004

Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*, 417–435. https://doi.org/10.1016/j.wocn.2009.07.005

Pell, M.D, Jaywant, A., Monetta, L., & Kotz, S.A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition & Emotion*, *25*, 834–853. https://doi.org/10.1080/02699931.2010.516915

Pell, M.D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., & Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, *111*, 14–25. https://doi.org/10.1016/j.biopsycho.2015.08.008

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), 1329–1333.

Pourtois, G., Delplanque, S., Michel, C., & Vuilleumier, P. (2008). Beyond conventional event-related brain potential (ERP): exploring the time-course of visual emotion processing using topographic and principal component analyses. *Brain Topography*, *20*, 265–277.

Puce, A., Epling, J. a., Thompson, J. C., & Carrick, O. K. (2007). Neural responses elicited to face motion and vocalization pairings. *Neuropsychologia*, *45*, 93–106. https://doi.org/10.1016/j.neuropsychologia.2006.04.017

Rellecke, J., Sommer, W., & Schacht, A. (2012). Does processing of emotional facial expressions depend on intention? Time-resolved evidence from event-related brain potentials. *Biological Psychology*, *90*(1), 23–32. https://doi.org/10.1016/j.biopsycho.2012.02.002

Righart, R., & de Gelder, B. (2008). Rapid influence of emotional scenes on encoding of facial expressions: An ERP study. *Social Cognitive and Affective Neuroscience*, *3*(3), 270–278. https://doi.org/10.1093/scan/nsn021

Rigoulot, S., & Pell, M. D. (2014). Emotion in the voice influences the way we scan emotional faces. *Speech Communication*, *65*. https://doi.org/10.1016/j.specom.2014.05.006

Rischer, K. M., Savallampi, M., Akwaththage, A., Salinas Thunell, N., Lindersson, C., & MacGregor, O. (2020). In context: Emotional intent and temporal immediacy of contextual descriptions modulate affective ERP components to facial expressions. *Social Cognitive and Affective Neuroscience*, *15*(5), 551–560. https://doi.org/10.1093/scan/nsaa071

Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex, 66,* A3–A20. https://doi.org/10.1016/j.cortex.2014.12.019

Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, *22*, 474–481. https://doi.org/10.1162/jocn.2009.21215

Schindler, S., & Bublatzky, F. (2020). Attention and emotion: An integrative review of emotional face processing as a function of attention. *Cortex, Pre-Print.* https://doi.org/10.1016/j.cortex.2020.06.010

Schirmer, A., Wijaya, M., Wu, E., & Penney, T. B. (2019). Vocal threat enhances visual perception as a function of attention and sex. *Social Cognitive and Affective Neuroscience*, *14*(7), 727–735. https://doi.org/10.1093/scan/nsz044

Schulz, C., Kaufmann, J. M., Kurt, A., & Schweinberger, S. R. (2012). Faces forming traces: Neurophysiological correlates of learning naturally distinctive and caricatured faces. *NeuroImage*, *63*(1), 491–500. https://doi.org/10.1016/j.neuroimage.2012.06.080

Schupp, H. T., Flaisch, T., Stockburger, J., & Junghofer, M. (2006). Chapter 2 Emotion and attention: event-related brain potential studies. *Progress in Brain Research*, *156*, 31–51. https://doi.org/10.1016/S0079-6123(06)56002-9

Schupp, H. T., & Kirmse, U. M. (2021). Case-by-case : Emotional stimulus significance and the modulation of the EPN and LPP. *Psychophysiology*, (December 2020), 1–13. https://doi.org/10.1111/psyp.13766

Schutter, D. J. L. G., De Haan, E. H. F., & Van Honk, J. (2004). Functionally dissociated aspects in anterior and posterior electrocortical processing of facial threat. *International Journal of Psychophysiology*, *53*(1), 29–36. https://doi.org/10.1016/j.ijpsycho.2004.01.003

Schweinberger, S. R., & Neumann, M. F. (2016). Repetition effects in human ERPs to faces. *Cortex*, *80*, 141–153. https://doi.org/10.1016/j.cortex.2015.11.001

Stahl, J., Wiese, H., & Schweinberger, S. R. (2008). Expertise and own-race bias in face processing: an event-related potential study. *Neuroreport*, *19*, 583–587.

Stolz, C., Endres, D., & Mueller, E. M. (2019). Threat-conditioned contexts modulate the late positive potential to faces—A mobile EEG/virtual reality study. *Psychophysiology*, *56*(4), 1–15. https://doi.org/10.1111/psyp.13308

Stürmer, B., Ouyang, G., Zhou, C., Boldt, A., & Sommer, W. (2013). Separating stimulus-driven and response-related LRP components with Residue Iteration Decomposition (RIDE). *Psychophysiology*, *50*(1), 70–73. https://doi.org/10.1111/j.1469-8986.2012.01479.x

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Tortosa, M. I., Lupianez, J., & Ruz, M. (2013). Race, emotion and trust: an ERP study. *Brain Research*, *1494*, 44–55.

Valentine, T., Lewis, M.B., & Hills, P.J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology, 69*(10), 1996-2019. https://doi.org/10.1080/17470218.2014.990392.

Van Kleef, G. A. (2009). How emotions regulate social life: The Emotions as Social

Information (EASI) Model. *Current Directions in Psychological Science, 18*(3), 184-188.

Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia, 45*(1), 174–194.

Wagenmakers, E-J, van der Maas, H.L.J., & Grasman, R.P.P.P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14* (1), 3-22.

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F. & Belin, P. (2014) Crossmodal adaptation in right posterior superior temporal sulcus during face–voice emotional integration. *The Journal of Neuroscience, 34*(20), 6813-1821.

Wieser, M. J., & Moscovitch, D. A. (2015). The effect of affective context on visuocortical processing of neutral faces in social anxiety. *Frontiers in Psychology*, *6* (NOV), 1–12. https://doi.org/10.3389/fpsyg.2015.01824

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage, 62*(1), 356–366.

Wuttke, S.J, & Schweinberger, S.R. (2019). The P200 predominantly reflects distance-to-norm in face space whereas the N250 reflects activation of identity-specific representations of known faces. *Biological Psychology*, *140,* 86-95. doi:10.1016/j.biopsycho.2018.11.011

Young, A.W., Frühholz, S., & Schweinberger, S.R. (2020). Face and voice perception: understanding commonalities and differences. *Trends in Cognitive Sciences*, *24* (5), 398-410.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*, 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Zekveld, A.A., Heslenfeld, D.J., Festen, J.M., & Schoonhoven, R. (2006). Top-down and bottom-up processes in speech comprehension. *NeuroImage, 32*, 1826-1836. Doi: 10.1016/j.neuroimage.2006.04.199.

6. Appendix

List of principal components which explained at least .5% of the total variance and .5% of the unique variance of the averaged ERP data time-locked to the emotional target face.

Principal component	Peak Latency (ms)[a]	Peak Channel[b]	Peak Polarity[c]	Total variance explained	Unique variance explained
TF1SF1	380	PO3	+	11.6%	1.3%
TF1SF2	380	PO9	-	8.5%	1.8%
TF1SF3	380	O2	+	7.5%	1.5%
TF1SF4	380	FT10	-	5.6%	0.9%
TF1SF5	380	Fp1	-	5.0%	1.1%
TF1SF6	380	PO7	-	2.0%	0.5%
TF2SF1	940	P3	+	8.6%	2.6%
TF2SF2	940	Cz	+	4.8%	1.4%
TF2SF3	940	AF8	+	4.2%	1.4%
TF2SF4	940	AF7	-	3.8%	1.0%
TF2SF5	940	PO8	-	2.3%	0.8%
TF2SF6	940	Fp1	-	2.2%	0.8%
TF4SF1	172	O2	+	1.3%	0.8%

Note: the PCA was based on the ERP average of each of the 275 sampling points of the entire epoch of the face stimulus for angry, sad, and happy faces preceded by different voice forms on 62 EEG channels and 23 participants.

[a]Peak Latency: the time point with the largest absolute voltage (of all the conditions after computing the grand average).
[b]Peak Channel: the channel with the greatest absolute voltage (out of all conditions after computing the grand average).
[c]Peak polarity: whether the voltage of the peak latency is positive or negative