Time-resolved discrimination of audio-visual emotion expressions

Federica Falagiarda(1) & Olivier Collignon(1,2)

Affiliations:

1 Institute of research in Psychology (IPSY) & Institute of Neuroscience (IoNS) - University of Louvain (UCL), Louvain, Belgium

2 Centre for Mind/Brain Studies, University of Trento, Trento, Italy

* Correspondence should be addressed to: Federica Falagiarda (federica.falagiarda@uclouvain.be) and Olivier Collignon (olivier.collignon@uclouvain.be).

Acknowledgement: This work was supported by a European Research Council starting grant (MADVIS grant #337573) attributed to OC and a Fond Special de Recherche from Université Catholique de Louvain attributed to FF. OC is a research associate at the National Fund for Scientific Research of Belgium (FRS-FNRS).

Author contributions: FF and OC designed the research; FF performed the research; FF and OC analyzed the data; FF and OC wrote the paper.

Data availability: the data, stimuli and analyses codes for this study are available at https://osf.io/yedsz/

No part of the study procedures or analyses was pre-registered prior to the research being conducted. 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.

Word count: 5474 (Title page, Abstract and References excluded)

Abstract

Humans seamlessly extract and integrate the emotional content delivered by the face and the voice of others. It is however poorly understood how perceptual decisions unfold in time when people discriminate the expression of emotions transmitted using dynamic facial and vocal signals, as in natural social context. In this study, we relied on a gating paradigm to track how the recognition of emotion expressions across the senses unfold over exposure time. We first demonstrate that across all emotions tested, a discriminatory decision is reached earlier with faces than with voices. Importantly, multisensory stimulation consistently reduced the required accumulation of perceptual evidences needed to reach correct discrimination (Isolation Point). We also observed that expressions with different emotional content provide cumulative evidence at different speeds, with "fear" being the expression with the fastest isolation point across the senses. Finally, the lack of correlation between the confusion patterns in response to facial and vocal signals across time suggest distinct relations between the discriminative features extracted from the two signals. All together, these results provide a comprehensive view on how auditory, visual and audiovisual information related to different emotion expressions accumulate in time, highlighting how multisensory context can fasten the discrimination process when minimal information is available.

Introduction

Being able to quickly and reliably discriminate facial and vocal expressions of emotion is a cognitive ability with high evolutionary importance since it helps humans and other animals engaging in successful social interactions (Darwin 1872).

The investigation of the visual display of emotions through facial expressions has dominated the research on emotions' perception. Pioneering investigations have demonstrated the existence of a set of presumably universal emotions (Ekman 1971), with distinctive patterns of facial expressions that are supposedly common across cultures (Ekman 1971; Ekman and Friesen 1986; Fridlund et al. 1987; Susskind et al. 2008, but see also Jack et al. 2012). The facial expressions linked to those universal discrete categories can already be distinguished by infants of a few months of age (Labarbera et al. 1976; Young-Browne et al. 1977) and appear to have at least partly separate neural correlates (Sprengelmeyer et al. 1998; Vytal and Hamann 2010). In our daily environment, prototypical muscular facial movements that compose emotional expressions (see the Facial Action Coding System; Ekman and Friesen 1978), also called action units (AU, e.g. eyebrow raising, mouth opening, etc), are gradually displayed on faces, hence making the vast majority of our exposition to facial expressions dynamic (Jack et al. 2014). The use of dynamic stimuli, similar to what we encounter in our natural environment, has been shown to produce higher discrimination performance as well as a differential neural activation when compared to static stimuli (Grèzes et al. 2007; Trautmann-Lengsfeld et al. 2013; Davies-Thompson et al. 2018). A thorough analysis of the evolution of the AUs of the 6 basic emotions over time (Jack et al. 2014) shows which AUs are progressively displayed on our faces when we go from a neutral to an emotional expression and how the dynamic expressions of basic emotions share common AUs at specific points in their temporal evolution, providing an explanation to why some emotions are more easily confused with each other (e.g. fear being often confused with surprise, and anger with disgust, Young et al. 1997; Jack et al. 2009). However, in spite of the evident dynamicity of facial expressions, the majority of the studies investigating facial emotions relied on static images. Subsequently, the natural time course of these expressions has become a heavily overlooked aspect and the power of these studies to generalize to our daily perceptual experience remains limited.

The perception of emotions conveyed through voices has received much less attention than emotional facial expressions. It has been shown however that the six basic emotions can be successfully discriminated through vocal expressions too and that this, as for faces, seem to happen cross-culturally (Sauter et al. 2010). Much less is known however about the standardized features

used to extract information in voices (Bochorowski and Owren 1999) and whether discrete vocal signals that might be somewhat comparable to AUs for faces exist.

When comparing the performance of discriminating emotion expression either delivered through faces or voices, it has been consistently shown that the perception of facial expressions generates higher discrimination rates and lower reaction times compared to the one of vocal expressions (Collignon et al., 2008; 2010; Charbonneau et al., 2013). Moreover, instead of being processed as separate channels, it seems that the information that progressively unfolds from the face and the voice gets seamlessly integrated into a coherent percept. For instance, it has been shown that the performance in a unimodal task is disrupted or facilitated when a respectively incongruent or congruent stimulus is presented simultaneously in a task-irrelevant modality, hence supporting the hypothesis of an automatic integration of the two expression types (Collignon et al., 2008; Dolan et al., 2001; de Gelder et al., 2005). The multisensory investigation of emotions therefore constitutes a more ecological approach to what is our daily experience.

In spite of the growing multisensory approach and use of dynamic stimuli, it has not yet been investigated how discriminatory decisions are reached over the time of the unfolding of emotional information in these expressions across the senses. Indeed, emotion expression from the face and voice are intrinsically time-embedded and the accumulation of sensory evidence allowing for a reliable decision about which emotion is being displayed may vary across the senses and across emotions. The present study has therefore been designed with the scope of evaluating how observers accumulate informational evidence at different time points during the unfolding of dynamic visual, auditory and bimodal emotional signals, using a gating paradigm (Grosjean 1980; Jesse and Massaro 2010; Sánchez-García et al. 2018). Building on existing cognitive models of speech perception (Marslen-Wilson and Welsh 1978; Marslen-Wilson 1987; Davis et al. 2002), we assume that when a perceiver performs the extraction of emotional information from the face and/or voice of an interlocutor, discrete stored properties of each emotion matching the incoming expression are rapidly and partially activated, similarly to the activation of the "word initial cohort" when hearing incoming speech (Marslen-Wilson 1987). The term "initial cohort", used in Marslen-Wilson model, refers to all the words of our lexicon that are activated when hearing incoming speech sounds; following the model, as we hear more of the speech, the possible matching candidates activated in our lexicon decrease in number, until an Isolation Point is reached, where only one candidate remains and can be compared with the perceived word. We hypothesize that the aforementioned model may apply to emotion discrimination, where, analogously to online

speech perception, such parallel activation forms a transient competition between emotion candidates that diminishes over time as the expression continues to unfold and therefore sensory evidence accumulates. Eventually, an isolation point is reached where the pool of candidates narrows to a single target emotion. This set of processes may constitute an optimally efficient system through which an emotion expression is identified as soon as it can be reliably differentiated from the others within an "emotion initial cohort". Therefore, the principal objective of our study was to investigate how the discrimination process of five basic emotions (fear, disgust, anger, sadness, happiness) evolves depending on the exposure time to facial, vocal and audiovisual stimuli. We hypothesize that a multisensory context in which both facial and vocal information is present will lead to a successful recognition with shorter exposure time than in any unimodal condition. In order to investigate this hypothesis, we empirically determined the threshold for discrimination, or Isolation Point (IP), through a psychophysical gating experiment (Grosjean 1980; Sánchez-García et al. 2018) in which the perceiver had to distinguish the emotion contained in increasingly larger segments of vocal, facial or bimodal congruent stimuli. In addition, we investigated whether discrimination follows a similar confusion pattern across our emotional set between the two modalities, and whether the confusion pattern in the audiovisual setting could be explained by the ones in the unimodal perception conditions.

Methods

Participants

Thirty-six healthy volunteers took part in the experiment (all right handed and Italian native speakers; 18 females; age range: 18-39. We aimed at a sample with a number of participants similar or above the ones used in previous studies on facial and vocal emotion perception, e.g. Collignon et al. 2008; Sauter et al. 2010). All participants needed to have normal or corrected-to-normal vision, self-reported normal audition and were naïve to the experimental hypothesis. The protocol was approved by the institutional review boards of the University of Trento and the Center for Mind/Brain Sciences (CIMeC). All subjects provided a written informed consent prior to the experiment and received course credits or a monetary reimbursement for their participation.

Stimuli and design

The stimuli were selected from a set of dynamic stimuli in which an actor/actress performs the prototypical facial and vocal (no linguistic content) expressions of the basic emotions and the pain expression, in different intensities (Belin et al., 2008 and Simon et al., 2006). The stimuli used for

the final version of the experiment contain five emotions (anger, disgust, fear, happiness, sadness) portrayed by two actors and two actresses which were selected through a pilot study. Starting with an original pool of 19 actors/actresses, audio and video components of each file were separated, resulting in 190 files containing visual information only or auditory information only (19 actors/actresses x 5 emotions x 2 modalities). The files were shuffled and presented to a group of raters (N=9), who watched/listened to every file twice and were instructed to report the emotion that they thought was being expressed in the stimulus through a forced choice among the five potential emotions, as well as a confidence rating of their response on a 7 points likert scale. The two male actors and two female actresses that were best discriminated were chosen to be used in the study; in case of equal discrimination rates, the actors/actresses with higher confidence rates were chosen.

The bimodal stimuli are in the form of .avi files with a frame rate of 29.97 fps and audio sampling at 48 kHz. The unimodal auditory and unimodal visual stimuli are the audio track and the visual frames respectively, of the original multimedia file. The visual frame presented to the subjects in the bimodal and the visual stimuli is 332 pixels wide and 424 pixels, high sustaining 7.7 degrees of visual angle at the participant’s distance to the screen (about 70 cm). The auditory stimulus was presented through headphones (Sennheiser HD 201) at a comfortable self-adjustable volume level.

We used a behavioral gating task (Grosjean 1980; Tyler and Wessels 1985) to determine the Isolation Point (IP) of each expressed emotion in the unisensory and bisensory contexts. The IPs are defined as the gate where participants reach a 60% correct response as predicted by a psychophysical function fitted to the data. The shortest stimulation segment, or *gate*, is constituted by one frame, hence non-moving, portraying a neutral expression and lasting 100ms. Such neutral frame has been taken by a different, neutral video recorded by the actors, and used in all stimuli. Progressively longer gates are built through an increment of ~33.37ms to reach a maximum duration for stimulus presentation of 400ms (i.e. gate 1: 100ms, gate 2: 133ms, gate 3: 167ms, gate 10: 400ms; see fig. 1). The so edited multimedia files result in a constant stimuli psychophysical experiment with three within-subjects factors and the following design: 10 (gates) x 5 (emotions) x 3 (modalities).

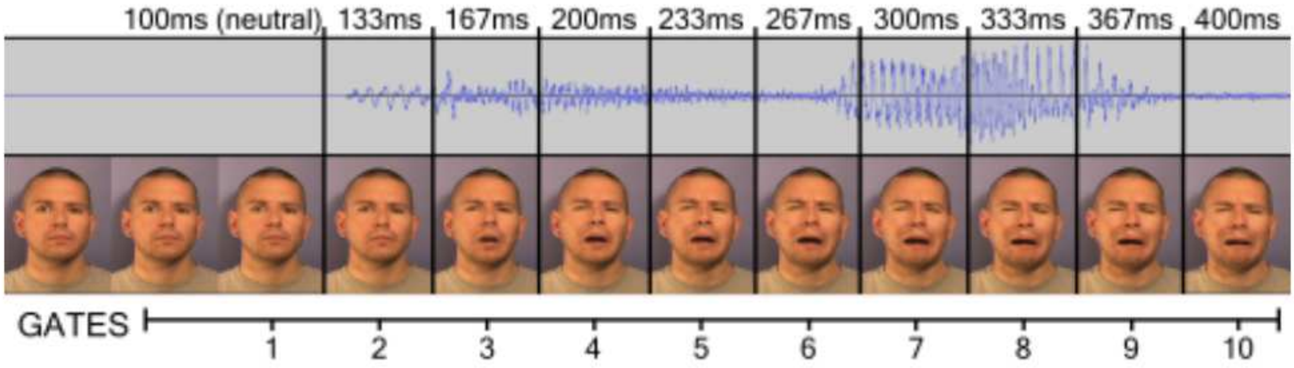

Fig. 1 – An illustration of the gating procedure applied to the stimuli.

Procedure

Participants underwent one session in the laboratory, lasting about one hour. They sat in a dimly lit room and were asked to identify the emotion expressed in the stimulus by pressing one of five buttons, each button identified with a label. Fast responses were discouraged but a time limit was set to 4s and participants were asked to be as accurate as possible within the given time window; the response had to be provided with one preferred finger. Each trial began with a fixation cross appearing for a random duration between 500 and 1000ms, followed by a stimulus lasting between 100 and 400ms (see fig. 1) and randomly selected across the auditory, visual or audio-visual conditions. The response had to be reported post-stimulus, when a question mark appeared in the center of the screen. Four trials were presented for each emotion in each modality and each gate, for a total of 600 randomized trials presented to every participant. The 600 trials were divided into twelve blocks to allow for breaks of self-determined duration throughout the experimental session. The experiment was implemented on Matlab 2013 with the Psychtoolbox extension (PTB-3) and GStreamer open source multimedia framework, on a MacBookPro with a 13.3-inches monitor and a refresh rate of 60Hz. The program collected the subjects' responses and reaction times.

Data analyses and results

Trials in which the computer failed to present the stimuli with the correct timing, with a tolerance of ±12ms around the expected duration since the video frame rate is not an exact multiple of the screen refresh rate, were discarded. Trials in which the subjects reached the time out limit, pressed a key other than the five labeled ones or responded faster than 150ms were also discarded. These

constituted 1.7% of the total number of trials. The remaining data were analyzed through psychometric curve fitting in order to determine the IP of discrimination of each emotion in each modality. Additionally, sensitivity indices (d’) were calculated to provide a reliable bias-free general measure of the performance in the task.

Psychometric fitting and Isolation Points

For each emotion and each modality, logistic psychometric curves were fitted to the proportion of correct responses as a function of the gates, through maximum likelihood estimation (fitting procedures were performed in R through the *Quickpsy* package, no parameter of the curves has been fixed *a priori*; Linares and López-Moliner 2016; R_Core_Team 2017). For illustration purposes only, see the representations of the fittings at a group level in figure 2. IPs were extracted from the models at the conventional value of 0.6 proportion of correct responses (p(threshold) = (1 + p(chance)) / 2)). The data relative to Sadness had to be excluded from the fitting procedure due to the presence of a reliable response bias in the shortest gates: participants tended to consistently report Sadness as their response to the shortest static neutral facial stimulus (see confusion matrices in fig. 5 and supplemental material). This causes the proportion of correct responses to be highly above chance level for Sadness in the visual and bimodal conditions for the shortest gate; which subsequently leads the IP values from sigmoid functions fitted to these data to be consistently lower than for all other emotions. Such thresholds extracted from biased models do not carry theoretical meaning when compared to other emotions.

A priori we identified three types of outlying IP values, they were excluded from further analyses (6% in total). The first is constituted by the conditions for which a poor goodness of fit, evaluated through a bootstrapping procedure, resulted from the psychophysical models (0.9%). The second type of outlying IPs are the ones that are extracted from decreasing curves or that have a negative value (3.2%); due to strong a priori expectations on the fact that performance, therefore the psychometric curves, should be growing with the growing exposure to the stimuli, a decreasing curve or a negative IP cannot reflect the psychological processes that we aim to investigate; these values are likely to be the outcome of distraction or other response biases in the task and need therefore to be removed from the analyses. Lastly, we considered outliers the IPs deviating >3SDs from the mean of their condition (1.9%).

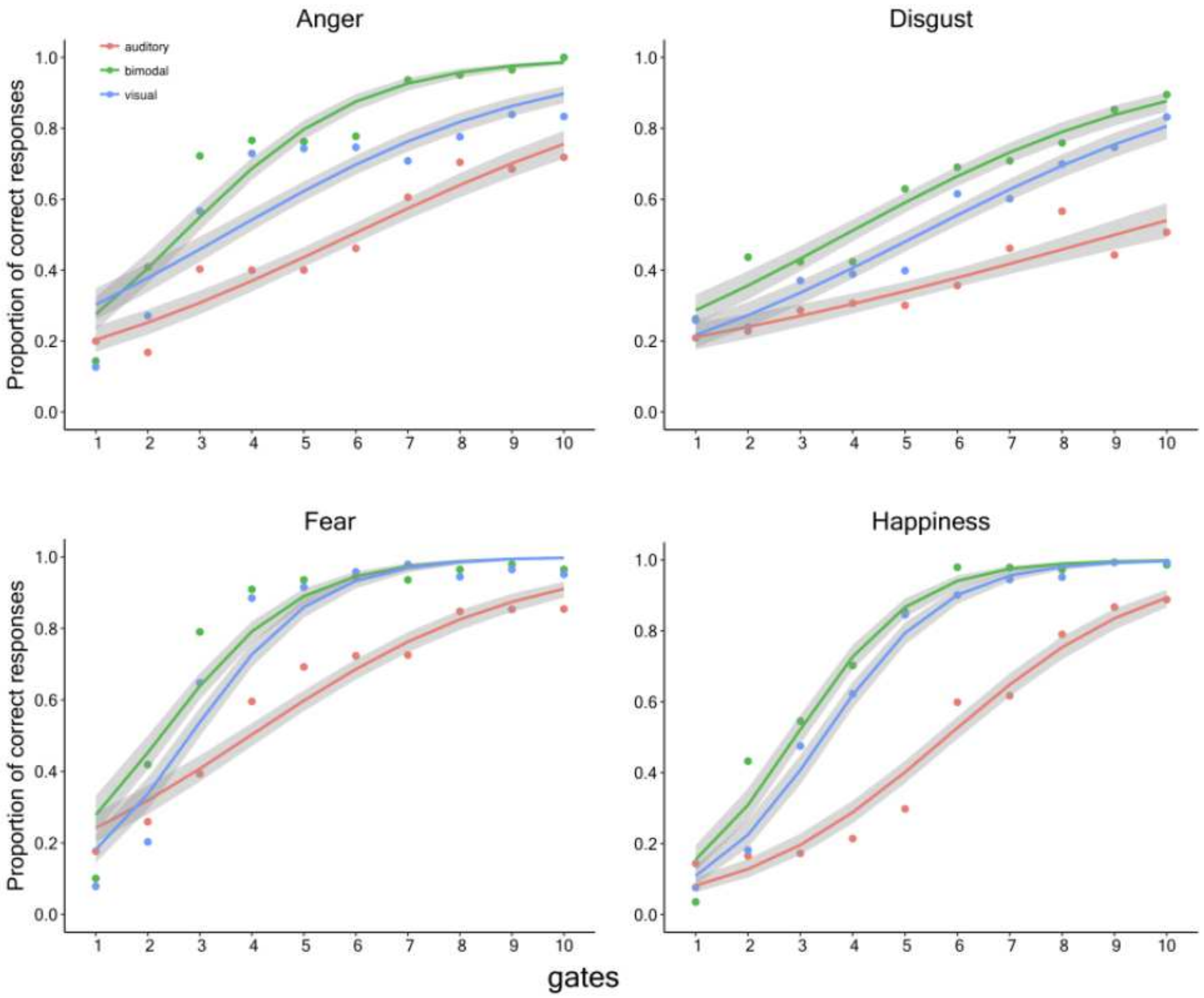

Fig 2 – Psychometric logistic models fitted to the data of the individual emotions and modalities (fitting representation at the group level).

The IP values extracted from the logistic curves were submitted to a generalized linear mixed model (GLMM) with Modality and Emotion as fixed effects and the subjects as random effect (marginal $R^2 = .62$; conditional $R^2 = .67$). An analysis of variance (Kenward-Rogers degrees of freedom approximation method) on the model was computed in order to evaluate the global effects of the predictors. This analysis revealed a main effect of Modality ($F = 224.9$, $p < .0001$) and Emotion ($F = 100.4$, $p < .0001$), as well as an interaction between the two factors ($F = 6.9$, $p < .001$). Post-hoc comparisons showed that the IP in the bimodal condition was significantly lower than either unimodal conditions and that the visual IP was significantly lower than the auditory one (all ps < 0.0001, Bonferroni corrected; fig. 3A). Post-hoc comparisons on the main effect of Emotion showed that the IPs of the analyzed emotions appeared to be all significantly different from each other, as

follows: the IP for Fear was the lowest, followed by Happiness, Anger, and finally Disgust (Fear vs Happiness, $p = .0002$; Fear vs Anger, $p < .0001$; Fear vs Disgust, $p < .0001$; Happiness vs Anger, $p = .019$; Happiness vs Disgust, $p < .0001$; Anger vs Disgust, $p < .0001$; Bonferroni corrected; Fig 3B). These results show how, regardless of the modality of presentation, discrimination is reached earlier for some emotion compared to others. Lastly, the interaction between Modality and Emotion suggests that the way each modality contributes to an earlier IP when going from vocal to facial and finally to the bimodal information, is different across emotions. For example, the auditory information contributes to a larger extent in producing a bimodal advantage when compared to the visual modality in Anger and Disgust than in Fear or Happiness (see fig. 3C). Despite the thorough interpretation of such interaction being beyond the scope of the experiment, it shows that the information in the two channels has different contribution and potentially different saliency across emotions.

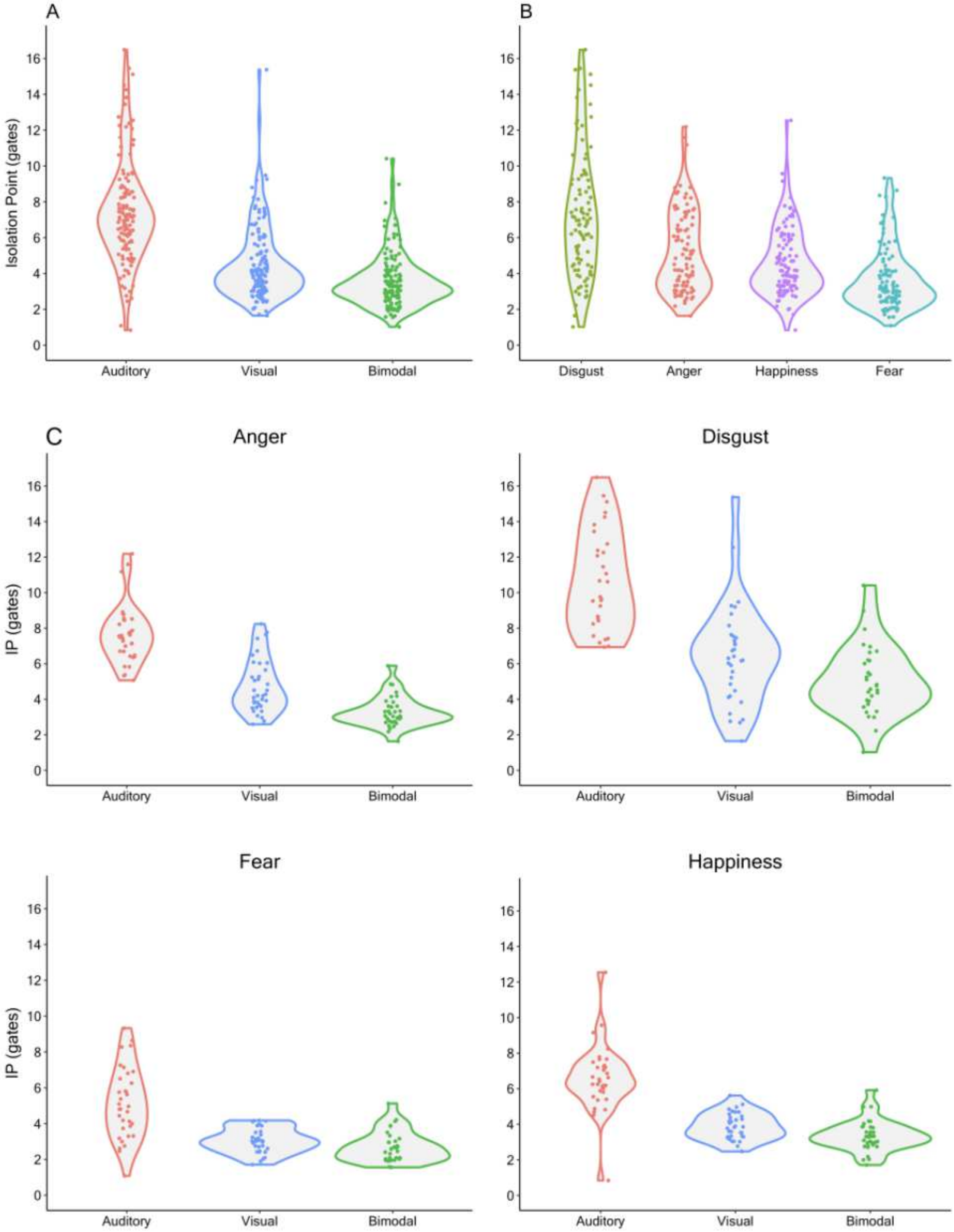

Fig. 3 – [A]: the violin plots represent the IP by modality, expressed as the gate at which 60% of accuracy is reached. Post-hoc t-tests revealed that all modalities have significantly different thresholds. [B]: the violin plots represent the IP for each fitted emotion, expressed as the gate at

which 60% of accuracy is reached. Post-hoc t-tests revealed that all emotions pairs, except Happiness vs. Anger, have significantly different thresholds. [C]: violin plots of the IPs, divided by modality and by emotion. N = 36.

Sensitivity indices

Sensitivity indices, or d' (d-primes), constitute an unbiased measure of performance since they not only consider correct responses, but incorrect responses too (Signal Detection Theory, Tanner and Swets 1954; Macmillan and Creelman 2004). Given that the fitting procedure are based on the proportion of correct responses, the calculation of d' allow us to complement and strengthen the results obtained through the analysis of the IPs, without the exclusion of Sadness. Sensitivity indices were calculated through the following formula: d' = z(Hit rate) – z(False alarm rate) (calculation performed in R, through the *sensR* package). They are shown in fig 4, divided by gate, emotion and modality. Since the shortest stimuli contained no emotional information, the correspondent d' cannot have a value significantly exceeding 0 and were removed from the following analyses of variance to avoid spurious interactions. D' values were submitted to a 9 (gates) x 5 (emotions) x 3 (modalities) repeated measures ANOVA. The ANOVA revealed a main effect of all factors (gate, $F = 487.78$, $p < .0001$, $\eta_p^2 = .93$; modality, $F = 483.47$, $p < .0001$, $\eta_p^2 = .93$; emotion, $F = 71.22$, $p < .0001$, $\eta_p^2 = .67$). Post-hoc analyses on the main effect of emotion reveals that different sensitivities arise for the different emotional expressions employed in the study: Disgust leads to a performance which is significantly lower than all other emotions; Fear leads to the highest performance, significantly higher than all other emotions except Happiness; Sadness and Anger do not differ from each other but lead to significantly higher performance than Disgust and significantly lower than Happiness and Fear (pairwise t-tests, Bonferroni corrected: Disgust vs. Sadness, $p < .0001$; Disgust vs. Anger, $p < .0001$; Disgust vs. Happiness, $p < .0001$; Disgust vs. Fear, $p < .0001$; Sadness vs. Anger, $p = 1$; Sadness vs. Happiness, $p < .0001$; Sadness vs. Fear, $p < .0001$; Anger vs. Happiness, $p < .0001$; Happiness vs. Fear, $p = 1$; see fig 4C). Post-hoc analyses on the main effect of Modality are highly consistent with the IPs results, showing how sensitivity is higher for visual expressions than auditory, but highest for bimodal expressions (pairwise t-tests, Bonferroni corrected: Bimodal vs Visual, $p < .0001$; Bimodal vs Auditory, $p < .0001$; Visual vs Auditory, $p < .0001$; see fig 4B). All the two-ways interactions and the three-way interaction from the analysis of variance on the d' are significant (gate*modality, $F = 6.69$, $p < .0001$, $\eta_p^2 = .16$; gate*emotion, $F = 10.5$, $p < .0001$, $\eta_p^2 = .23$; modality*emotion, $F = 23.66$, $p < .0001$, $\eta_p^2 = .4$; gate*modality*emotion, $F = 4.42$, $p < .0001$,

$\eta_p^2 = .11$), however the thorough interpretation of each interaction goes beyond the scope of this paper, therefore these results will not be discussed further.

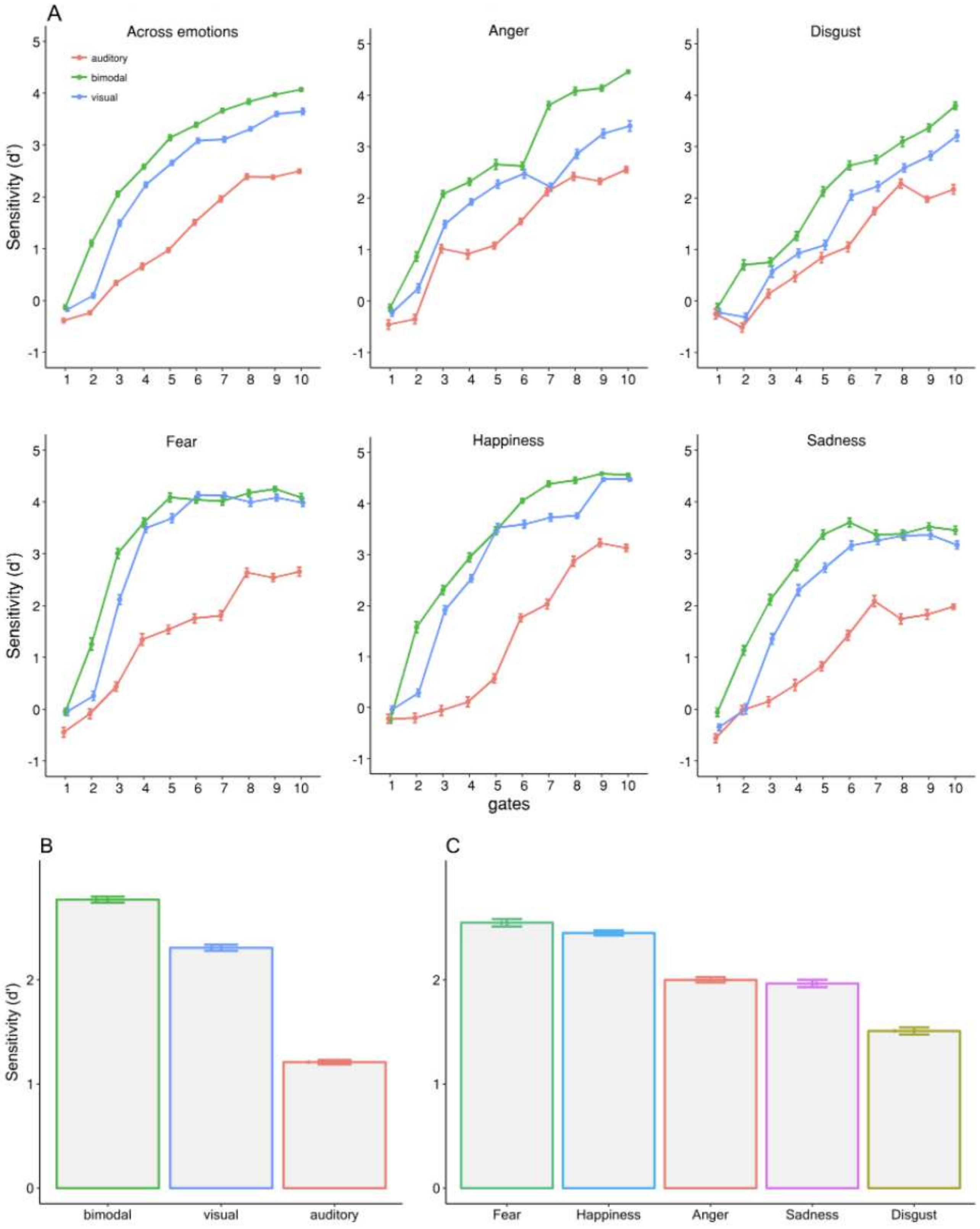

Fig. 4 – [A] Sensitivity indices are shown for each emotion, each modality, and each gate separately.

[B and C] The bar plots represent the relevant main effects, of modality and emotion respectively.

In order to exclude the possibility that the general increase in performance at the increasing duration of the stimulation segments might be due to a simple increase in conveyed energy throughout the unfolding of the facial and vocal expressions, we investigated potential correlations between change in energy in our stimuli and change in the performance of the participants. The difference in root mean square (RMS) was measured across gates in auditory files, while the instantaneous visual change (IVC, defined in Brookshire et al. 2017) was measured across frames of the visual stimuli for two separate subparts of the face, the eyes and the mouth. The differences in RMS across gates showed no positive correlation with the subjects' performance in the discrimination of vocal expressions across gates. The IVC measured for the eyes area showed a positive correlation with the performance difference across gates in the visual task for "Fear" ($r = .44$, $p < .0001$), and no positive correlation for any other emotion (see supplementary material for a detailed description and further discussion of the RMS and IVC analyses).

Confusion matrices

Behavioral confusion matrices were calculated. The matrices across subjects are represented in fig. 5 for each gate and each modality separately. The emotions contained in the stimuli are represented by rows, while the emotions of the subjects' responses are represented by columns. Correct responses therefore lay on the diagonal of the matrix, while the off-diagonal data represents incorrect responses, or *confusion patterns*. Since the visual and auditory components of the stimuli originate from one multimedia file in which an actor or actress is performing the facial expression and the vocalization simultaneously, we reasoned that a correlation at each time point between the confusion matrices originating from the two signals could be observed (e.g. when the mouth widens or narrows, the intensity and pitch of the vocal signal will likely change). In order to assess whether behavioral confusion patterns share similarities across visual and auditory information over time, we computed the correlations between the confusion patterns of visual matrices and the auditory matrix of the corresponding gate (on-diagonal data was discarded from all matrices to avoid spurious correlations; Ritchie et al. 2017). Opposite to our expectation, all the correlations were non-significant, except a weak positive correlation for gate 5 ($r = .14$, $p = .019$; all other $ps > .21$, Bonferroni corrected). These results suggest that confusion patterns of behavior might be driven by

features of the stimuli that differ between facial and vocal expressions and that are not correlated between the two channels.

Additionally, we wanted to see whether the multimodal confusion pattern could be predicted based on the confusion patterns in the visual and auditory condition unimodally, we therefore ran a linear regression (on-diagonal data discarded) with the following formula: bimodal ~ visual + auditory. A significant regression equation could be found ($R^2 = .23$, $F = 1539$, $p < .0001$). The two coefficients of the visual and auditory terms in the model are positive and significant, suggesting that the contribution of the predictors in explaining the predicted term is positive, however the auditory estimate is much smaller than the visual (estimate(auditory) = .055 $t = 6.9$, $p < .0001$; estimate(visual) = .477, $t = 53.3$, $p < .0001$). These results show that the bimodal confusion patterns share features, however to a different extent, with both the visual and the auditory confusions.

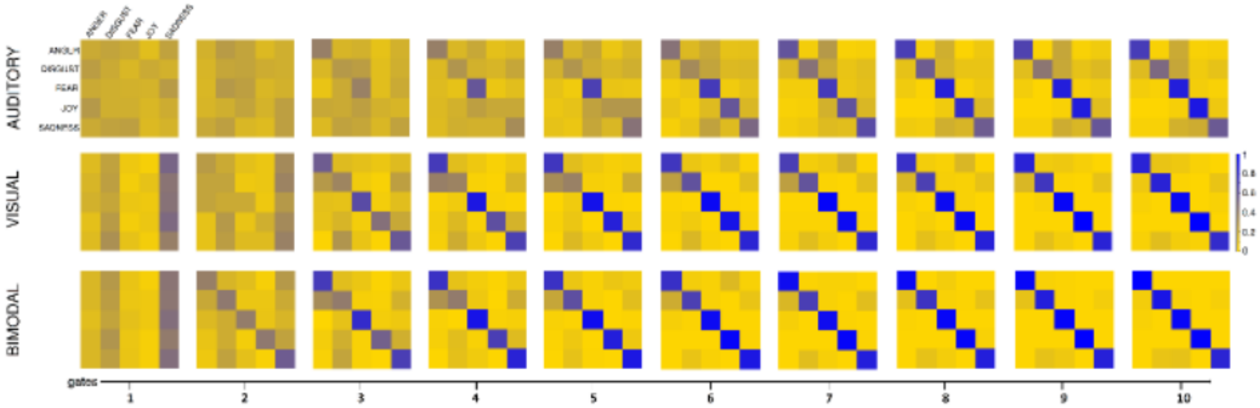

Fig. 5 – Confusion matrices. The emotions labeled on the rows represent the content of the stimuli, while the emotions labeled on the columns represent the subjects' responses. The diagonals therefore represent the correct responses, while the off-diagonal data constitutes the "confusion patterns". The bias to responding Sadness in the shortest neutral facial stimuli can be clearly seen in the visual and bimodal matrices of gate 1.

Discussion

With the present study we aimed at measuring whether discriminatory decisions are achieved at different time points during the unfolding of facial and/or vocal expressions of different basic emotions (anger, disgust, fear, happiness and sadness). Using a gating paradigm (Grosjean 1980), the emotions were presented as stimuli of incremental length visually, auditorily and audio-visually, through dynamic facial expressions and non-linguistic vocalizations.

Our results show that faster discrimination is achieved through dynamic facial expressions than their correspondent vocalization. In other words, when presented visually with a facial emotional expression, we need a shorter exposure to the stimulus compared to the matching vocal expression in order to reach the same efficient discrimination. In the field of word recognition, where the gating task has been more widely exploited, it has been shown that when discriminating a set of words that only differ by one phoneme, the performance can be better either in the visual or in the auditory domain depending on the saliency of the modality for each specific phoneme, and multisensory integration does not necessarily lead to a more successful discrimination (Sánchez-García et al. 2018). Instead, in the context of emotion expressions' discrimination, our results robustly show that a multimodal context is always advantageous, and a discriminatory decision is reached earlier than in either unisensory condition. These findings are consistent across all investigated emotions (see fig. 3) and indicate that the multisensory integration process allows the perceiver to maximally beneficiate from the redundant emotional information carried by the face and the voice.

We hypothesize that reaching the IP when discriminating emotions in the three investigated modality conditions (auditory, visual, bimodal) is the product of a fast and transient competition among a range of discrete features of the emotional categories of interest (in our case five emotion expressions) that activate upon the situation in which emotional signal has to be extracted from faces and voices around us. Specifically, the competition seems to be faster-resolving in the visual domain, where the discrete features that identify emotional expressions, e.g. facial AUs, are likely to be diagnostic for discrimination earlier than in the auditory domain. In the multimodal context however, stored discrete features of both visual and auditory emotions can be activated at the beginning of the bimodal perception to form the "initial cohort". This "multimodal initial cohort" will supposedly count a lower number of potential candidates for emotional discrimination, because they will need to match a plausible combination of facial and vocal features instead of the auditory or visual features in isolation. This will likely allow for a faster selection process and it seems to constitutes the optimal situation to generate an advantage in terms of reaching an IP earlier when compared to even the best unimodal condition.

An alternative explanation to the advantage observed in the multimodal context compared to the unimodal ones can also be provided in the frame of a race model (Miller 1982), namely a model in which the information from the two modalities, visual and auditory, is each accumulated in a separate decisional unit; each unit will produce a decision, in this case regarding the emotional

value of the signal, and one of the two decisions will then prevail. This instance is opposite to the one in which the two signals are first integrated in order to lead to one decision. Both situations however, predict an advantage in a multimodal context compare to its unimodal components. The race model inequality test (Miller 1982) has been applied exclusively to reaction times' data, and heavily employed to show whether redundant signal effects (RSE) were due to mere statistical facilitation (no integration), or coactivation of the redundant signals (integration), see e.g. Miller 1986 and Otto and Mamassian 2016. It was previously demonstrated that reaction times to facial and vocal emotional stimuli violate the race model inequality (Collignon et al. 2008) therefore suggesting that emotions concomitantly delivered through faces and voices interact at the processing stage, possibly through the involvement of posterior superior temporal regions showing a functional preference for multisensory emotional signals delivered through faces and voices (Davies-Thompson et al. 2018). The current study has however been designed with optimal features for integration: the employed bimodal stimuli are always spatially and temporally congruent and the displayed face and voice were always coherent in gender, age and emotional content. Nevertheless, we acknowledge that the collected data, which is centered around performance, is not suited to disentangle between facilitation or coactivation.

Our results also show that the isolation point varies across emotions. In particular, among the considered expressions, "Fear" is the one for which discrimination is achieved earliest, regardless of the modality of presentation. Fearful faces have been shown to be detected better compared to neutral or other emotional facial expressions (Milders et al. 2006). They have also been found to reach conscious perception faster than other emotional categories when presented with continuous flash suppression (Yang et al., 2007) and to have higher probability of being detected when presented in binocular rivalry (Amting et al, 2010). In general, humans seem to have a preferential way to process threat-related emotional stimuli compared to other stimuli that, though emotional, do not require a potential fast action towards survival (Öhman et al. 2001; Anderson et al. 2013; Stein et al. 2014). Our results support and extend the view that *Fear* is somehow special among a variety of emotions by showing that people are able to discriminate *Fear* after a very short exposure, regardless of the sensory channel through which the fearful information is conveyed. In other words, when a fearful stimulus is present, the "initial cohort" narrows down to a single option faster than for the other investigated emotions. Furthermore, the finding that the IVC of the eyes in fearful visual stimuli shows a correlation with the profile of the performance in the visual task (see Supplemental Material) may be interpreted as a sign of the evolutionary pressure towards an

efficient displaying of this threat-signaling stimulus. This might have resulted in an identification of this specific emotional instance as being more, compared to the other emotions, a combination of top-down features-matching process (during the selection from the “emotional initial cohort”) and bottom-up features such as the amount of visual change in our perceptual field from each moment to the next, in this case, particularly salient is probably the increasingly enlarging area occupied by the eyes. However, in the light of an evolutionary interpretation, it remains surprising that the same correlational effect is not found for Fear in the vocal domain too.

No reliable correlations between the confusion matrices generated by the facial and the vocal signals were observed. This leads to the speculation that, despite an evident temporal co-evolution within the congruent visual and auditory signals, the features of the segmented stimuli that are eventually driving the behavioral confusions among emotions, are substantially different between the two modalities and therefore give rise to very distinct patterns of confusion for facial compared to vocal stimuli. When looking at the multimodal confusion patterns, however, we find that they can be modestly estimated by the visual confusion patterns, and to a lower extent by the auditory confusions. This result points to the notion that we regard vision as our most reliable sensory channel, and that, despite a non negligible effect of the auditory information, in case of uncertainty in a multimodal context, our response will mostly be driven by our visual perception. A compelling approach to the study of how the reliability of each sensory cue influences multisensory integration is the use of a Bayesian framework to evaluate whether human observers are optimal integrators, namely whether humans apply Bayesian laws when perceiving and integrating different sensory signals (Ernst and Banks 2002). However, testing this hypothesis requires specific experimental designs that rely on a manipulation of the discrepancy between the signals from the different sensory modalities and require to compare the precision and variance of multisensory decisions to those taken in unisensory context. Our study however was not suitable for this approach, as the aim of the experiment was the calculation of time-resolved psychophysical thresholds and we therefore never had any discrepancy between sensory channels. Future studies should aim to test whether the integration of emotional faces and voices follow Bayesian principles.

Finally, the absence of consistent positive correlation between the evolution of the participants’ performance across gates and the amount of energy displayed over time in the auditory and visual stimuli (except for the previosuly discussed correlation of the seleceted eye part in the fearful expressions) suggests that the profile of the measured behavior cannot be due merely to a low-level change in the amount of rudimentary energy displayed gate by gate in the stimuli.

For example, the same amount of IVC between two frames might subtend different facial AUs with different degrees of diagnosticity when performing the task of emotion recognition. These results support even further the idea that discriminating emotions while attending the dynamic unfolding of an expression seem to be a complex process that involves more than a mere feed-forward processing of the incoming information.

In conclusion, our results demonstrate how the discrimination of emotion expressions unfold over the time of stimulus presentation across auditory and visual modalities. We interpret our results in the frame of a transient competition between emotion expression representations that get activated in parallel and are progressively narrowed down to fewer, and eventually one representation, as perceptual and cognitive evidences accumulate (see e.g. Marslen-Wilson 1987 for similar reasoning on word recognition). Such competition lasts less in vision, where an isolation point is reached earlier than in audition. Crucially, it lasts shortest in a multisensory context, where information is redundantly portrayed in the face and in the voice. Moreover, our results support the idea of the specific salience of *fearful* expression. Given the sensitivity of our paradigm in disclosing variations across emotions and modalities, it might represent a promising avenue for research involving populations for which a problem in discriminating emotion expressions has been suggested (e.g. autism, Charbonneau et al. 2013; schizophrenia, Kohler et al. 2003; etc).

References

Amting JM, **Greening SG**, **Mitchell DG V.** Multiple Mechanisms of Consciousness: The Neural Correlates of Emotional Awareness. *J Neurosci* 30: 10039–10047, 2010.

Anderson AK, **Christoff K**, **Panitz D**, **De Rosa E**, **Gabrieli JDE**. Neural correlates of the automatic processing of threat facial signals. *Soc Neurosci Key Readings* 23: 185–198, 2013.

Belin P, **Fillion-Bilodeau S**, **Gosselin F**. The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behav Res Methods* 40: 531–539, 2008.

Bochorowski J-A, **Owren MJ**. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J Acoust Soc Am* 106: 1054–1063, 1999.

Brookshire G, **Lu J**, **Nusbaum HC**, **Goldin-Meadow S**, **Casasanto D**. Visual cortex entrains to sign language. *Proc Natl Acad Sci* 114: 6352–6357, 2017.

Charbonneau G, **Bertone A**, **Lepore F**, **Nassim M**, **Lassonde M**, **Mottron L**, **Collignon O**. Multilevel alterations in the processing of audio-visual emotion expressions in autism spectrum disorders. *Neuropsychologia* 51: 1002–1010, 2013.

Collignon O, **Girard S**, **Gosselin F**, **Roy S**, **Saint-Amour D**, **Lassonde M**, **Lepore F**. Audio-visual integration of emotion expression. *Brain Res* 1242: 126–135, 2008.

Collignon O, **Girard S**, **Gosselin F**, **Saint-Amour D**, **Lepore F**, **Lassonde M**. Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia* 48: 220–225, 2010.

Darwin C. *The expressions of emotions in man and animals*. 1872.

Davies-Thompson J, **Elli G V**, **Rezk M**, **Benetti S**, **van Ackeren MJ**, **Collignon O**. Hierarchical brain network for face and voice integration of emotion expression. *Cereb. Cortex*. .

Davis MH, **Marslen-Wilson WD**, **Gaskell MG**. Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *J Exp Psychol Hum Percept Perform* 28: 218–244, 2002.

Dolan RJ, **Morris JS**, **de Gelder B**. Crossmodal binding of fear in voice and face. *Proc Natl Acad Sci* 98: 10006–10010, 2001.

Ekman P. Universals and cultural differences in facial expressions of emotion. In: *Nebraska Symposium on Motivation, 19*. 1971, p. 207–283.

Ekman P, **Friesen W V.** *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

Ekman P, **Friesen W V.** A New Pan-Cultural Facial Expression of Emotion. *Motiv Emot* 10: 159–168, 1986.

Ernst MO, **Banks MS**. Humans integrate visual and haptic information in a statistically optimal fashion. 415: 429–433, 2002.

Fridlund AJ, **Ekman P**, **Oster H**. Facial expressions of emotion. *Nonverbal Behav. Commun.* .

Gelder B De, **Vroomen J**, **Jong SJ De**, **Masthoff ED**, **Trompenaars FJ**, **Hodiamont P**. Multisensory integration of emotional faces and voices in schizophrenics. *Schizofr Res* 72: 195–203, 2005.

Grèzes J, **Pichon S**, **de Gelder B**. Perceiving fear in dynamic body expressions. *Neuroimage* 35: 959–967, 2007.

Grosjean F. Spoken word recognition processes and the gating paradigm. *Percept Psychophys* 28: 267–283, 1980.

Jack RE, **Blais C**, **Scheepers C**, **Schyns PG**, **Caldara R**. Cultural Confusions Show that Facial Expressions Are Not Universal. *Curr Biol* 19: 1543–1548, 2009.

Jack RE, **Garrod OGB**, **Schyns PG**. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr Biol* 24: 187–192, 2014.

Jack RE, **Garrod OGB**, **Yu H**, **Caldara R**, **Schyns PG**. Facial expressions of emotion are not culturally universal. *Proc Natl Acad Sci* 109: 7241–7244, 2012.

Jesse A, **Massaro DW**. The temporal distribution of information in audiovisual spoken-word identification. *Atten Percept Psychophys* 72: 209–225, 2010.

Kohler CG, **Turner TH**, **Bilker WB**, **Brensinger CM**, **Siegel SJ**, **Kanes SJ**, **Gur RE**, **Gur RC**. Facial emotion recognition in schizophrenia: Intensity effects and error pattern. *Am J Psychiatry* 160: 1768–1774, 2003.

Labarbera JD, **Izard C**, **Vietze P**, **Parisi SA**. Four- and Six-Month-Old Infants ' Visual Responses to Joy , Anger , and Neutral Expressions Author (s): J . D . LaBarbera , C . E . Izard , P . Vietze and S . A . Parisi Published by : Wiley on behalf of the Society for Research in Child Development Sta. *Child Dev* 47: 535–538, 1976.

Linares D, **López-Moliner J**. quickpsy: An R Package to Fit Psychometric Functions for Multiple Groups [Online]. *R J* 8: 122–131, 2016. https://journal.r-project.org/archive/2016-1/linares-na.pdf.

Macmillan NA, **Creelman CD**. *Signal Detection Theory: a User's Guide*. Psychology Press, 2004.

Marslen-Wilson W. Functional parallelism in word recognition [Online]. *Cognition* 25: 71–102, 1987. http://www.sciencedirect.com/science/article/pii/0010027787900059%0Apapers3://publication/uuid/EC313344-6F08-4FF3-A828-5C41D9DD9DFB.

Marslen-Wilson WD, **Welsh A**. Processing interactions and lexical access during word recognition in continuous speech. *Cogn Psychol* 10: 29–63, 1978.

Milders M, **Sahraie A**, **Logan S**, **Donnellon N**. Awareness of faces is modulated by their emotional meaning. *Emotion* 6: 10–17, 2006.

Miller J. Divided attention: Evidence for coactivation with redundant signals. *Cogn Psychol* 14: 247–279, 1982.

Miller J. Timecourse of coactivation in bimodal divided attention. 40: 331–343, 1986.

Öhman A, **Lundqvist D**, **Esteves F**. The face in the crowd revisited: A threat advantage with schematic stimuli. *J Pers Soc Psychol* 80: 381–396, 2001.

Otto TU, **Mamassian P**. Multisensory Decisions : the Test of a Race Model , Its Logic , and Power. (2016). doi: 10.1163/22134808-00002541.

R_Core_Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,2017.

Ritchie JB, **Bracci S**, **Op de Beeck H**. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *Neuroimage* 148: 197–200, 2017.

Sánchez-García C, **Kandel S**, **Savariaux C**, **Soto-Faraco S**. The Time Course of Audio-Visual Phoneme Identification: A High Temporal Resolution Study. *Multisens Res* 31: 57–78, 2018.

Sauter DA, **Eisner F**, **Ekman P**, **Scott SK**. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc Natl Acad Sci* 107: 2408–2412, 2010.

Simon D, **Craig KD**, **Miltner WHR**, **Rainville P**. Brain responses to dynamic facial expressions of pain. *Pain* 126: 309–318, 2006.

Sprengelmeyer R, **Rausch M**, **Eysel UT**, **Przuntek H**. Neural structures associated with recognition of facial expressions of basic emotions. *Proc Biol Sci* 265: 1927–1931, 1998.

Stein T, **Seymour K**, **Hebart MN**, **Sterzer P**. Rapid Fear Detection Relies on High Spatial Frequencies. *Psychol Sci* 25: 566–574, 2014.

Susskind JM, **Lee DH**, **Cusi A**, **Feiman R**, **Grabski W**, **Anderson AK**. Expressing fear enhances sensory acquisition. *Nat Neurosci* 11: 843–850, 2008.

Tanner WP, **Swets JA**. A decision-making theory of visual detection. *Psychol Rev* 61: 401–409,

1954.

Trautmann-Lengsfeld SA, **Domínguez-Borràs J**, **Escera C**, **Herrmann M**, **Fehr T**. The Perception of Dynamic and Static Facial Expressions of Happiness and Disgust Investigated by ERPs and fMRI Constrained Source Analysis. *PLoS One* 8, 2013.

Tyler LK, **Wessels J**. Is gating an on-line task? Evidence from naming latency data. *Percept Psychophys* 38: 217–222, 1985.

Vytal K, **Hamann S**. Neuroimaging Support for Discrete Neural Correlates of Basic Emotions: A Voxel-based Meta-analysis. *J Cogn Neurosci* 22: 2864–2885, 2010.

Yang E, **Zald D**, **Blake R**. Fearful expressions gain preferential access to awareness during continuous flash suppression. *Emotion* 7: 882–886, 2007.

Young-Browne G, **Rosenfeld HM**, **Horowitz FD**. Infant Discrimination of Facial Expressions. *Child Dev* 48: 555, 1977.

Young AW, **Rowland D**, **Calder AJ**, **Etcoff NL**, **Seth A**, **Perrett DI**. Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition* 63: 271–313, 1997.