The 15th International Conference on Auditory-Visual Speech Processing
10-11 August 2019, Melbourne, Australia

Unbalanced visuo-auditory interactions for gender and emotions processing

Jimmy Debladis[1], *Kuzma Strelnikov*[2], *Shally Marc*[2], *Maïthé Tauber*[3] *and Pascal Barone*[1,2]

[1]Brain & Cognition Research Center (CerCo), CNRS, Faculty of medicine, Toulouse, France
[2]Brain & Cognition Research Center (CerCo), University Paul Sabatier, Toulouse, France
[3]Toulouse Purpan, Physiopathology center, INSERM, France

pascal.barone@cnrs.fr, tauber.mt@chu-toulouse.fr

Abstract

During everyday communication, voice and facial cues are combined. A preference for the auditory or visual channel is chosen automatically whereas, in most of the previous studies, guided attention was used. In our study, we performed a comparison of the visual influence on vocal non-verbal emotions and gender using the same paradigm in the same subjects without instructions on attention direction. The voice for emotions and gender was modeled as a continuum with 11 steps. The validated non-ambiguous images of gender and emotions were presented in the congruent and incongruent with the voice way. Audiovisual performance was assessed with respect to the auditory performance. We observed a small improvement of performance in the congruent audiovisual stimulation both for gender and emotions with a smaller effect for emotions. In the incongruent conditions, face cues strongly dominated the performance with a significantly larger effect for gender. The proportion of the subjects who made their decision on the visual basis was significant only for the gender voice continuum. The strength of facial dominance is significantly different between the identity voice information and emotional prosody. We suggest that face-voice interaction in human may not be the same for linguistic, para-linguistic and identity properties.

Index Terms: emotions, gender, voice, audiovisual, prosody

1. Introduction

Voice carries verbal and non-verbal speech information and is especially important for prosody and paralinguistic contextual cues. But, speech is a multisensory processing and in a face-to-face conversation, visual information provided by the lip movements and the articulatory gestures are complementarily involved in speech comprehension [1]. Indeed, during speech processing when there is a mismatch between the visual and the auditory signal, such as in the McGurk protocol [2], normal hearing subjects often fuse both types of information creating a percept, which is different both from auditory and facial linguistic cues. In many social communicative situations, the voice provides similar information to those carried by face. Voice was suggested to be an auditory face for identity information about gender, age, physical factors and emotions [3], [4]. It is important that in usual social communication, facial and vocal information about speaker's state of mind are highly complementary for linguistic, paralinguistic and affective information [5], [6]. This complementarity requires brain mechanisms of coupling between two different sources of relevant information to extract the important features in the presence of redundancies from each modality as well as, in some cases, to decide on incongruent features.

Based on the evidence for strong neural convergence of complementarity cues from face and voice, a model of face-voice interactions in the brain has been proposed involving an internal supra-modal representation of the person [7]. This model implies multimodal influence on unimodal processing stages, suggesting that the balance between multimodal and unimodal networks depends on the exact nature of the task and stimuli that could underlie the perceptual interpretation of simultaneous signals from multiple sensory modalities. There is now some evidence of the existence of sensory dominance when multimodal interactions are engaged. Indeed, in the spatial domain, the dominance of the visual modality participates in the ventriloquism illusion [8], while the prevalence of the auditory modality in the temporal domains has been also demonstrated to modulate visual perception [9]. The cross-modal bias has been explored in several studies based on face-voice interactions especially through protocols that demonstrate a modulation of multisensory interactions through the engagement of attentional processes. Such studies were conducted independently on different features carried by the voice and the face naming the gender or the emotional content [10]–[12]. In this present study, we have chosen to focus on these two important features for social communication (emotions and gender) and to compare the cross-modal bias that can spontaneously occur in a face-voice interaction. We varied the voice using a continuum between female and male identity and between sad and happy intonations. We assessed the performance of subjects using a paradigm, which was adapted for the subjective perceptual preferences as no instructions as to the attention to the auditory or visual information was given. Each participant was asked for general decision on the person's gender identity or emotions with respect to any spontaneously preferred criteria.

2. Materials and methods

2.1. Participants

To explore audiovisual integration of gender and emotion information, we tested a group of 30 normally hearing subjects, native French speakers (13 Males, 17 Females, and aged 26±6(SD) years) with no self-reported history of auditory, neurological or psychiatric disorders. They were asked to perform in separate tasks gender and emotion two forced-choice categorization on the basis of their own criteria.

10.21437/AVSP.2019-8

2.2. Stimuli and procedure

All voice stimuli were developed at the Voice Neurocognition Laboratory of the University of Glasgow (http://vnl.psy.gla.ac.uk). The task requires participants to categorize by gender or emotions either in voice or face stimuli from a morphing-generated voice continuum between a male and a female voice speaking the same syllable or between the sad and happy voice. The syllables were "ha" (1-1.4 sec) for emotions and "pa" (0.4 sec) for gender. The two-extreme voices each corresponded to an average voice from 16 voices of the same gender. Morphing was performed using STRAIGHT toolbox (Hideki Kawahara, University of Wakayama) in Matlab 6.5. STRAIGHT performs instantaneous pitch-adaptive spectral smoothing to separate the contributions of the glottal source (including F0) from the supra-laryngeal filtering (distribution of spectral peaks, including the first formant F1) to the voice signal. Voice stimuli were decomposed by STRAIGHT into five parameters: fundamental frequency (F0), formant frequencies, duration, spectro-temporal density, and aperiodicity; each parameter can be independently manipulated. Anchor points, that is, time-frequency landmarks, were determined in both extreme voices based on easily recognizable features of the spectrograms. The temporal landmarks were defined as the onset, the offset, and the initial burst of the sound. Spectro-temporal anchors were the first and second formant at onset of phonation, onset of formant transition, and end of phonation. Morphed stimuli were then generated by re-synthesis based on a logarithmic interpolation of extremes' anchor templates and spectrograms in steps of 10%. We thus obtained a continuum of 11 voices ranging from 100% female to 100% male as well as from 100% sad to 100% happy with 9 gender-interpolated voices in 10% steps.

Auditory gender stimuli were paired to a male, a female or a sad and a happy static face in audiovisual (AV) conditions in order to obtain the same number of congruent and incongruent AV stimulations. Visual stimuli corresponded to two colored photographs of a male and a female that we chose as gender representative. They were improved using Adobe Photoshop © and were light and contrast normalized using a Matlab ® algorithm. Before using them, we checked their validity by asking 10 subjects to categorize the faces as male or female, sad or happy and 100% scores were obtained for each type faces. The duration of image presentation was the same as the duration of voice presentation.

Subjects were tested in a sound-attenuated chamber with volume adjusted to 72 dB SPL. Auditory stimuli (16-bits, stereo, 22 050 Hz sampling rate) were presented binaurally via Sennheiser Eh 250 headphones.

Subjects performed the two tasks in two blocks independently, (first the gender, then the emotion task or vice versa). In each block, stimuli were presented randomly. Participants heard three repetitions of the sounds in the continuum and three repetition of the congruent and incongruent condition for a total of 99 stimuli presentation (33 AV congruent, 33 AV incongruent and 33 auditory). The participant had to categorize the gender or the emotion of the person, and not to focus on a specific modality.

2.3. Analysis of the data

We calculated the rate of responses "female" or "sad" for each of the 11 voices in each continuum. To analyse the effect of simultaneous presentation of a visual face on voice categorisation performance, we computed for each subject a Visuo-auditory interaction index (VIx), in the A and AV conditions. Firstly, we calculated the area under the raw data curve (AUC) separately for each side (male or female, sad or happy) of the continuum. This computation was made in A, AV congruent (AVc,) or AV incongruent (AVic) conditions. The values were standardised through a mirror image with respect to the response rate to allow for comparisons between each side of the continuum (Figure 1). VIx corresponds to the ratio of the surface area obtained in A and AV conditions normalised with respect to the A condition (VIx = (AV-A)/A), where the AV conditions (female and male, happy and sad face) are averaged. Values close to zero indicate an absence of the dominance of face presentation on auditory categorisation.

Direct comparisons of the performances (VIx, slope values) between groups were performed using the bootstrap method. The data for each group were re-sampled 10,000 times. As a result, we obtained a distribution of 10,000 stimulated observations for subjects in each condition, from which we obtained the means of the samples. We used bias-corrected and accelerated confidence intervals [13] and estimated the significance of the mean correct response at the level of $p<0.05$.

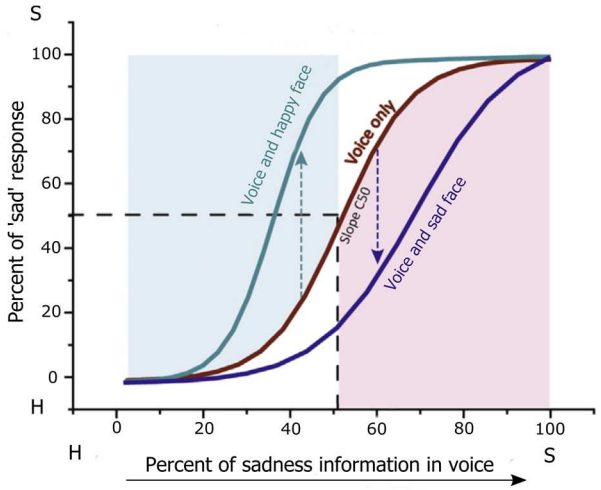

Figure 1: *Theoretical sigmoid curves for A and the two AV conditions (A and congruent or incongruent face). This example illustrates the emotion categorization task. For the gender task, the same paradigm was used. The calculated surfaces with respect to the curves are shown.*

3. Results

We revealed a facilitator effect on reaction time in the congruent condition to the same extent between gender and emotion task.

Figure 2 represents multisensory gains ((A-AV)/A) for reaction times obtained for the congruent and incongruent presentations. In the incongruent condition, no gain was obtained. Moreover, a significant difference is observed between gender and emotion tasks in the incongruent situation ($p<0.05$, paired bootstrap, Cohen's $d=0.26$ [0.20, 0.72]), meaning that incongruent face stimuli disrupt the speed of auditory categorization.

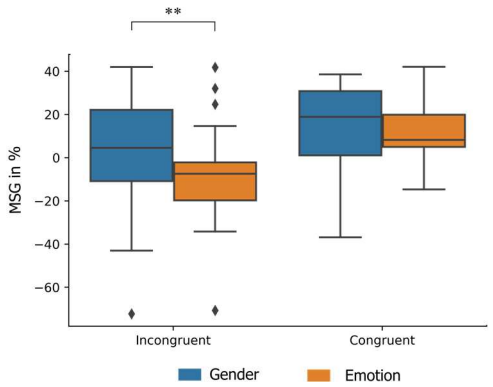

Figure 2 : *Multisensory gain (MSG) for reaction times expressed in percentages for the congruent and incongruent conditions.*

Figure 3 illustrates the psychometric function of the subjects during the auditory-only categorisation tasks as well as during the audio-visual tasks, in which voices were combined with faces. One can see that subjects categorised correctly the unambiguous voices at the extremities. When stimuli were closer to 50% on the continuum, subjects categorised the voice as female or male, happy or sad with approximately the same probability. Globally, as shown in Figure 3, the psychometric curves of the participants can be fitted with a sigmoid function.

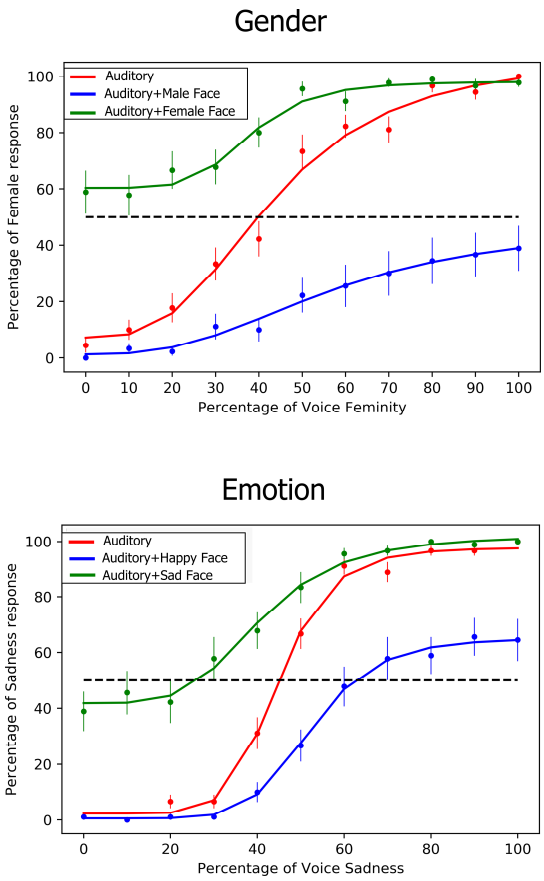

Figure 3: *Sigmoidal fitting curves during gender and emotion categorization task.*

When a face is presented with a voice, categorization is based on the gender carried by the face and not the voice (more than 50% of female response for a male sound and conversely). For the emotion categorization, the visual effect is weaker than that for the voice (40% of happy responses for a sad voice).

To determine the dominant modality, we calculated an average score based on the responses made for the incongruent condition. If a participant obtained an average score above 55%, we considered this participant as visually dominant and below 45%, as auditory dominant. For those comprised between 45 and 55 %, they have been considered as a mixed profile.

Table 1: *Proportion of visual and auditory profiles during gender and emotion categorization tasks. The p-values represent the statistical difference between the two profiles.*

	Visual	Auditory	Chi^2	p value
Gender	63.3%	30%	5.42	0.02
Emotion	40%	43.3%	$1.44e^{-30}$	1

We revealed (Table1) that in the gender categorization the proportion of visual profile is higher compared to auditory profile (p=0.02) whereas in the emotion categorization, the proportions were in the same range (40% visual against 43.3% auditory).

To quantify the obtained psychometric curves, we calculated the surfaces under the curves and the ratio of the surface area obtained in A and AV conditions normalised with respect to the A condition (VIx = (AV-A)/A), where the AV conditions (female and male face or happy and sad face) are averaged.

If the presentation of a face interacts with the voice gender/emotion categorisation, we expect an increase of the surface area representing a facilitator effect in case of congruency. This would increase the VIx values. Inversely, in case of incongruence between the voice and face stimuli, the AUC would be reduced as well as the VIx. This is exactly what is observed in Figure 4. However, the decrease in performance for Gender condition is significantly higher than for Emotion condition in the incongruent situation ($p<0.05$, paired bootstrap, Cohen's d=0.32 [0.06, 0.59]). Besides, the increase of performance in the congruent situation is significantly higher for Gender than for Emotions ($p<0.05$, paired bootstrap, Cohen's d=0.30 [0.04, 0.56]). Thus, emotional faces have weaker dominance for emotional information than gender-specific faces for gender information.

If one considers the slopes of the psychometric curves at 50% of the continuum, one can see that for both Gender and Emotion conditions slopes are higher for the auditory modality than for the audiovisual ones in our paradigm ($p<0.05$, paired bootstrap, for Gender, Cohen's d=0.65 [0.26, 1.06], for Emotions, Cohen's d=0.63 [0.23, 1.03]),), meaning that the categorization is efficient in the auditory-only condition (Figure 5).

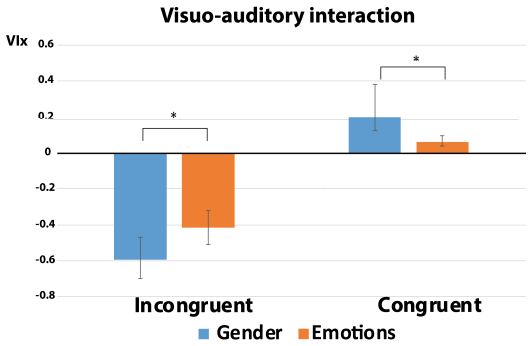

Figure 4: *Audiovisual indices based on the areas below (above) the psychometric curves. Visuo-auditory interaction index (VIx) is calculated according to the formula VIx = (AV-A)/A).*

Besides, already in the auditory-only modality the slope for Emotion condition is higher than for Gender condition ($p<0.05$, paired bootstrap).

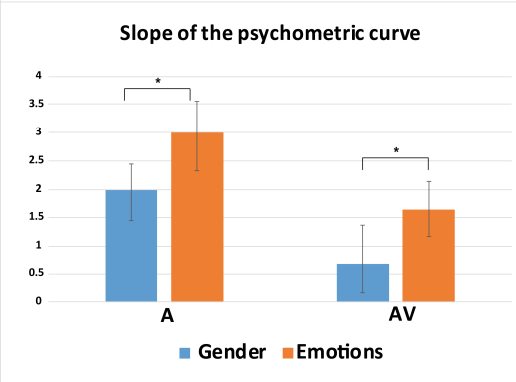

Figure 5: *Slopes of the psychometric curves in the auditory (A) and audiovisual (AV) conditions.*

4. Discussion

Our results provide further evidence that voice carries information on a person's identity and that this identity information is highly dependent on visual modality.

Though the role of visual modality for gender discrimination has already been described in literature [14], it remained unclear whether this role is the same as that in emotional prosody. In other words, whether the use of emotional load for voice in speech can have different interaction with the non-linguistic identity information such as gender.

Most of the previous studies used directed attention towards auditory or visual stimulation [10], [14], however, in ecological conditions a person directs her attention automatically on the basis of her individual communicational preferences. These preferences may also differ between the gender and emotional identifications.

The originality of our study consists in the direct comparison of the visual dominance for voice emotions and gender using the same paradigm in the same subjects without instructions on attention direction.

This comparison permitted us to discover several points, which distinguish the visual and auditory processing for gender and emotions. Though the subjects were already good performers in the auditory condition, we observed a small improvement of their performance in the congruent audiovisual stimulation for both gender and emotions (Figure 4). However, this improvement was smaller for emotions.

The most impressive effects were observed in the incongruent conditions where the incongruent face strongly dominated the performance compared to the auditory-only condition. The shift towards the face-coded information was about 50% for the extremities in the gender continuum and about 30% in the emotional continuum leading to a significant difference between the conditions (Figure 4). Besides, the proportion of the subjects who made their decision on the visual basis was significant for the gender continuum but not significant for the emotions continuum.

To explain this important difference between the visual impact of face on gender and emotions, one can suggest that it probably depends on the different visual or auditory cues that are used to distinguish each component. First, gender distinction in voice is based in a large part on pitch (as well as other properties [15]) with higher pitch for female voices. However, pitch characteristics partly overlap between male and female voices, the same is true for other parameters of gender voice discrimination [15]. EEG studies demonstrated that during gender voice discrimination there is a very early pitch discrimination and a later more accurate determination on gender [12]. Such features can explain partly that gender categorization could be more sensitive to the visual presentation of a face that help to disambiguate the auditory voice. Indeed, the beneficial impact of a visuo-auditory presentation is observed with the more androgynous voices. Concerning emotional prosody in voice, it implies a modulation in the frequential (pitch) domain but also in the temporal domain, e.g. faster or slower speech rate, raising or falling intonation etc. [16]. The same applies to face because emotions can be differently perceived when comparing static versus dynamic faces. Such distinction between gender and emotion processing in the temporal domain could explain why the visuo-auditory interaction for emotions tends toward the auditory channel in our protocol. Static faces in our study may be less effective for emotions than for gender, which does not require dynamic face expressions.

The slope of the psychometric curve (Figure 5) is significantly higher in the auditory condition for emotions than for gender meaning that emotions are more easily categorized than gender. Though a certain overlap may also exist between happy and sad prosody, the linguistic expertise may help subjects disambiguate emotional prosody more efficiently than auditory gender information. Moreover, given that subjects are less certain about voice gender, they are more susceptible to the information from the face both in the congruent and incongruent situations. This is supported by a significant proportion of visually oriented subjects in the gender condition. Conversely, having more certainty about emotional prosody in the auditory modality, subjects rely to a lesser extent on visual cues than in the gender discrimination. This concerns only ambiguous sounds because the differences between emotions and gender in the auditory psychometric curves are observed only for the ambiguous stimuli but are absent for the clearly distinguished stimuli at the extremities of the curves.

One should note that our results demonstrate the differences between the visual dominance for the emotions and gender taken as separate conditions but they do not exclude the

interactions between emotional and gender information, which can be quite complex in the auditory and visual modalities and even more complex in case of their multimodal fusion [17].

The obtained results on the role of face cues may have further applications in the studies of pathology where either the auditory signal can be disturbed (deafness, cochlear implantation) or face-related information may be less pertinent (like in autism spectrum disorders). They would provide further understanding of the deficits in social communication in the important human pathologies and would indicate a possible perspective for audiovisual rehabilitation of such deficits.

5. Conclusions

Audiovisual interaction exists both for prosodic and identity cues, however, the strength of this interaction is significantly different. Our results suggest that the dominance of face may be more pronounced for the identity information in voice than for emotional prosody. This further may suggest that face-voice interaction in human is not the same for linguistic, para-linguistic and identity properties.

6. Acknowledgements

This work was supported by grants from the Foundation for Prader-Willi Research (FPWR to JD, BP and MT) and recurrent funding from the CNRS (to BP, JD and KS). All the authors report no potential conflicts of interest with this work.

7. References

[1] J. I. Skipper, V. van Wassenhove, H. C. Nusbaum, and S. L. Small, "Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception," *Cereb Cortex*, vol. 17, no. 10, pp. 2387–99, 2007.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–8, 1976.

[3] P. Belin, P. E. G. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *Br J Psychol*, vol. 102, no. 4, pp. 711–725, 2011.

[4] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: neural correlates of voice perception," *Trends Cogn. Sci. (Regul. Ed.)*, vol. 8, no. 3, pp. 129–135, 2004.

[5] J. M. Foxton, L.-D. Riviere, and P. Barone, "Cross-modal facilitation in speech prosody," *Cognition*, vol. 115, pp. 71–78, 2010.

[6] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: head movement improves auditory speech perception," *Psychol Sci*, vol. 15, no. 2, pp. 133–137, 2004.

[7] S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends Cogn Sci*, vol. 11, pp. 535–43, 2007.

[8] P. Bertelson, J. Vroomen, B. De Gelder, and J. Driver, "The ventriloquist effect does not depend on the direction of deliberate visual attention," *Perception & Psychophysics*, vol. 62, no. 2, pp. 321–332, 2000.

[9] L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear," *Nature*, vol. 408, no. 6814, p. 788, 2000.

[10] J. Vroomen and B. de Gelder, "Sound enhances visual perception: cross-modal effects of auditory organization on vision," *J Exp Psychol Hum Percept Perform*, vol. 26, no. 5, pp. 1583–1590, 2000.

[11] C. S. Watson, W. W. Qiu, M. M. Chamberlain, and X. Li, "Auditory and visual speech perception: confirmation of a modality-independent source of individual differences in speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 2 Pt 1, pp. 1153–62, 1996.

[12] M. Latinus and M. J. Taylor, "Discriminating Male and Female Voices: Differentiating Pitch and Gender," *Brain Topogr*, vol. 25, no. 2, pp. 194–204, 2012.

[13] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," *Statistics in medicine*, vol. 19, no. 9, pp. 1141–64, 2000.

[14] M. Latinus, R. VanRullen, and M. Taylor, "Top-down and bottom-up modulation in processing bimodal face/voice stimuli," *BMC neuroscience*, vol. 11, p. 36, 2010.

[15] M. Latinus, P. McAleer, P. E. G. Bestelmeyer, and P. Belin, "Norm-Based Coding of Voice Identity in Human Auditory Cortex," *Curr Biol*, vol. 23, no. 12, pp. 1075–1080, 2013.

[16] L. Rachman *et al.*, "DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech," *Behav Res Methods*, vol. 50, no. 1, pp. 323–343, 2018.

[17] D. A. Harris and V. M. Ciaramitaro, "Interdependent Mechanisms for Processing Gender and Emotion: The Special Status of Angry Male Faces," *Front Psychol*, vol. 7, p. 1046, 2016.