Inferring Emotional Information from Vocal and Visual Cues: a Cross-Cultural Comparison

Maria Teresa Riviello[1], Anna Esposito[1], Mohamed Chetouani[2], David Cohen[2]
[1]Seconda Università di Napoli, Department of Psychology, and IIASS, Italy
[2]University Pierre and Marie Curie (UPMC), Paris, France

Emails: mariateresa.riviello@unina2.it; iiass.annaesp@tin.it; mohamed.chetouani@upmc.fr; david.cohen@psl.ap-hop-paris.fr

Abstract—The present work reports results of perceptual experiments aimed to explore the human ability to recognize emotional expressions through the visual and auditory channel, investigating if one channel is more effective than the other to infer emotional information and if this effectiveness is affected by the cultural context and in particular by the language.
To this aim American, French, and Italian subjects were involved in a comparative analysis of subjective perceptions of emotional states dynamically portrayed by visual and vocal cues, exploiting a cross-modal database of verbal and non-verbal American video clips. What should make the difference is that while one group of participants is native speaker of the language and belongs to the same cultural context of the video-clips used as stimuli (the American English), the other two are not. Results showed that emotional information is affected by the communication mode and that language plays a role.

I. INTRODUCTION

Emotion is one of the topics that has received much attention in the last few years in the context of Human Computer Interaction (HCI) and recent studies in this field aimed to investigate the perceptual and cognitive role of visual and auditory channels in conveying emotional information, identifying methods and procedures capable of automatically detect them exploiting the multimodal nature of emotional expressions [1-2]. A considerable part of the research on emotions and related perceptual cues to infer them has focused separately on three expressive domains: facial expressions, voice and body movements. On these research lines, some studies sustain that facial expressions are more informative than gestures and vocal expressions [3-4] whereas others underline the faithfulness of vocal expressions in portraying emotional states since physiological processes, such as respiration and muscle tension, are naturally influenced by emotional responses [5-7]. In this debate, the data reported in literature seem to favor facial expressions [8-9]. However, most of the studies investigating the perceptual and automatic recognition of emotional facial expressions exploited static images as the ones proposed in the FACS [10] or the Japanese Female Facial Expression (JAFFE) database [11] and the ORL Database of Faces [12]. There, emotional facial expressions were evaluated in terms of static positions in still photographs that usually capture the apex of the expression, i.e. the instant at which the indicators of emotion are most marked, while vocal expressions were evaluated along the time dimension as speech is intrinsically a dynamic process.
The present work attempt to investigate the power of visual and auditory channels in conveying emotional information exploiting dynamicity also in facial expressions. Is dynamic visual information still emotionally richer than auditory information? In the attempt to answer this question, a cross-modal database comprising dynamic verbal and non-verbal emotional stimuli based on video-clips extracted from American movies was created, and a series of perceptual experiments aimed to portray the underlying meta-structure of the affective communication were defined. The collected stimuli allowed to characterize emotional dynamic cues of some basic emotions transmitted dynamically by the visual and auditory channels, and identified some preferential channels exploited by humans for perceiving an emotional state, in particular for Italian subjects [13-16]. In a cross cultural perspective, it would be worth to investigate if the ability to recognize emotional expressions as a function of the channel is also affected by the cultural context and in particular by the subject's native language. To this aim, perceptual experiments devoted to assess the subjective perception of emotional states, exploiting emotional audio and video stimuli were conducted on separate groups of American, French and Italian subjects.

II. THE CROSS-MODAL EMOTIONAL DATABASE

The collected data are based on extracts from American movies whose protagonists were carefully chosen among actors and actresses that are largely acknowledged by the critique and considered capable of giving some very real and careful interpretations. The use of audio and video stimuli extracted from movies provided a set of realistic emotional expressions [14]. Differently from the other existing emotional databases proposed in literature, in this case the actors/actresses had not been asked to produce an emotional expression, rather, they were acting according to a movie script and their performance had been judged as appropriate to the required emotional context by the movie director (supposed to be an expert). Moreover, even though the emotions expressed in such video-clips were simulations under studio conditions (and may not have reproduced a genuine emotion but an idealization of it) they were able to catch up and engage the emotional feeling of the spectators (the addressers) and therefore, provided more confidence

on the value of their perceptual emotional content.

The database consisted of audio and video stimuli representing 6 emotional states: *happiness*, *sarcasm/irony*, *fear*, *anger*, *surprise*, and *sadness* (except for sarcasm/irony, the remaining emotions are considered by many theories as basic and therefore universally shared [3-6]). For each of the above emotional states, 10 stimuli were identified, 5 expressed by an actor and 5 expressed by an actress, for a total of 60 video-clips, each acted by a different actor and actress to avoid bias in their ability to portray emotional states. The stimuli were short in duration (the average length was 3s, SD = ± 1s) to avoid the overlapping of emotional states that could confuse the subject's perception. Care was taken in selecting video clips where the protagonist's face and the upper part of the body were clearly visible. In addition, the semantic meaning of the produced utterances was not clearly expressing the portrayed emotional state and its intensity level was moderate. For example, stimuli of sadness where the actress/actor was clearly crying or stimuli of happiness where the protagonist was strongly laughing are not included in the data. This was an attempt to let the participants to exploit less obvious emotional cues that were generally employed in every natural and not extreme emotional interaction.

The emotional labels assigned to the stimuli were first given by two experts and then by three naïve judges independently. The expert judges made a decision on the stimuli carefully exploiting emotional information on facial and vocal expressions such as frame by frame analysis of changes in facial muscles, and F0 contour, rising and falling of intonation contour, etc, as reported by several authors in literature [13, 18-19], and the contextual situation the protagonist is interpreting. The naïve judges made their decision after watching the stimuli several times. There were no opinion exchanges between the experts and naïve judges and the final agreement on the labeling between the two groups was 100%.

The collected stimuli, being extracted from movie scenes contain environmental noise and therefore are also useful for testing realistic computer applications. The database is available in the context of the COST Action 2102 and could be required mailing the second author of this paper.

The audio and the video alone were extracted from each complete stimulus (video-clip) coming up with a total of 120 stimuli: 60 mute video and 60 audio stimuli.

The stimuli in each set were randomized and proposed to separate group of American, French and Italian participants.

A. Participants

A total of 180 subjects (60 Americans, 60 French and 60 Italians), participated at the perceptual experiments. For each nationality, 30 subjects (equally distributed for gender) were involved in the evaluation of the audio and 30 in the evaluation of the mute video stimuli. The participants' age was similar among countries, ranging from 18 to 35 years. The knowledge of English by Italian and French subjects was comparable, since all of them used it as second language.

The subjects were randomly assigned to the task and were required to carefully listen and/or watch the stimuli via headphones in a quiet room. They were instructed to pay attention to each presentation and decide which of the 6 emotional states was expressed in it. Responses were recorded on matrix paper form (60x8) where rows listed the stimuli numbers and columns the 6 selected emotional states plus the "*others*" indicating any other emotion not listed and the option of "*no emotion"* that was suggested when according to the subject's feeling the protagonist did not show emotions. For each emotional stimulus the percentage of correct emotion recognition was computed.

III. RESULTS

The data obtained from each group of subjects (60 American, 60 French and 60 Italian) were first analyzed separately. This analysis provided the percentages of label's agreement (recognition accuracy) provided by the representatives of each country, split in subgroup of 30 and tested separately on the audio and mute video stimuli. For each emotional state, the percentages were computed considering the number of correct answers provided by participants over the total number of expected correct answers. Since American subjects share both the language and the cultural background of the selected emotional expressions with the encoders of the stimulus material, their performance can be considered as a reference for an optimal identification of the emotional states under examination. Figures I, II, and II display the percentage of the recognition accuracy of the American, French, and Italian subjects, respectively.

FIGURE I

Percentage of agreement obtained under the two experimental conditions (only audio –black bars- and only mute video –gray bars) by the American subjects

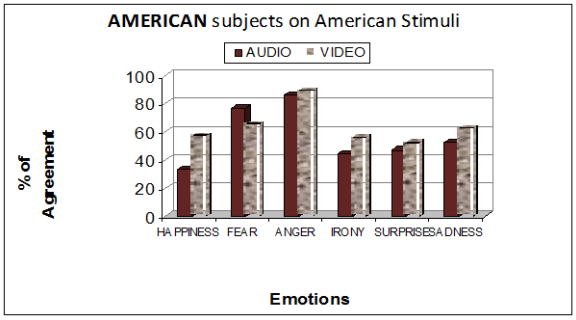


FIGURE II

Percentage of agreement obtained under the two experimental conditions (only audio –black bars- and only mute video –gray bars) by the French subjects

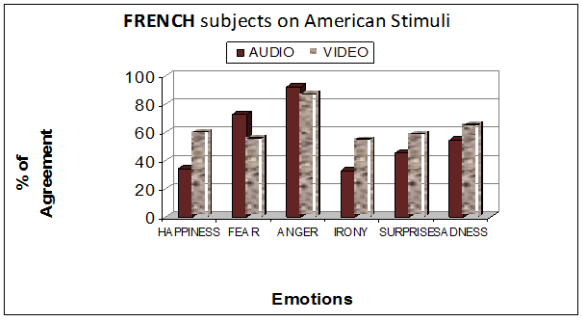
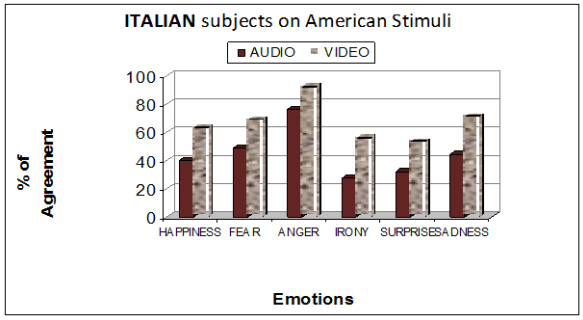


FIGURE III
Percentage of agreement obtained under the two experimental conditions (only audio –black bars- and only mute video –gray bars) by the Italian subjects



An ANOVA analysis was performed on the data obtained from the American, French and Italian subjects separately, considering the *Perceptual mode* as a between subjects variable and the *Emotions and Actor's Gender* as within subjects variables. Significance was established for α=.05. The ANOVA shows that for American (F (1, 8) = 1.696, ρ= .22) and French (F (1, 8) = 3.427, ρ= .1013) subjects, mute video and audio alone convey the same amount of emotional information. This is not the case for Italian subjects (F (1, 8) = 33.74, ρ=.0004), where the perceptual mode plays a significant role. It is worth to note that for Americans the identification of stimuli is not affected by the specific emotion portrayed (F (5,40)= 1.401, ρ= .24), while this is not true for French (F (5,40) = 2.520, ρ= .04), and Italian (F (5,40)= 4.050, ρ = .0046) subjects that showed a preference in inferring emotional information of fear (especially from vocal cues for French, and from visual cues for Italians), happiness and sadness (both French and Italian subjects obtained high percentage of correct recognition on mute video stimuli), and anger (that is very well recognized both from the audio and video modes). In addition, the gender of the protagonist do not affects the recognition accuracy of the American (F (1, 8) = 3.254, ρ = .10), French (F (1, 8) = 6.443, ρ = .03), and Italian (F (1, 8) = 4.370, ρ = .07) subjects. However, an interaction between the category of emotion and actor's gender was found for American (F (5, 40) = 9.721, ρ= .0001), French (F (5, 40) = 7.156 ρ= .0001), and Italian (F (5, 40) = 8.532, ρ= .0001) subjects. This interaction is significant for all the emotional categories under examination except for sadness and anger. Anger is also the emotional category with the highest percentage of recognition accuracy independently from the nationality and the perceptual mode.

In order to be able to assess the effectiveness of communication modes (auditory and visual) in conveying emotional information Figures IV and V report the percentage of correct agreement, on the audio and video stimuli, obtained by the American, French and Italian subjects.

FIGURE IV
Percentage of agreement obtained under the audio alone experimental condition by the American (black bars), French (grey bars) and Italian (green bars) subjects

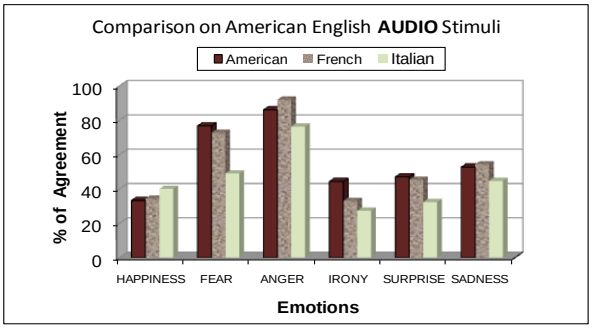
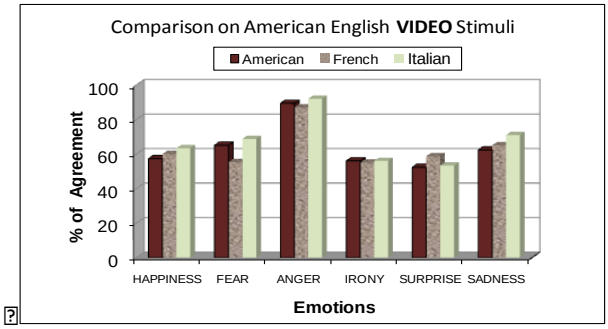


FIGURE V
Percentage of agreement obtained under the mute video experimental condition by the American (black bars), French (grey bars) and Italian (green bars) subjects



An ANOVA analysis was performed considering the subjects' *Nationality* as a between subjects variable and the *Emotions and Actor's Gender* as within subjects variables. Significance was established for α= .05. The analyses show that, when American, French and Italian subjects are tested on Audio stimuli, Nationality plays a significant role (F (2, 12) = 4.288, ρ=.04). Post hoc test revealed that the Italian differ significantly both from French and American subjects for the audio whereas, no significant differences were found for mute video stimuli, since, as expected the visual channel share visual emotional features across cultures.

IV. CONCLUSIONS

In this work, the power of visual and auditory channels in conveying emotional information exploiting dynamicity in facial and vocal emotional expressions has been investigated exploring a cross-modal database comprising dynamic visual and vocal emotional stimuli extracted from American movies. The database allows the cross-modal analysis of audio and video recordings for defining distinctive, multi-modal, and cultural specific emotional features, and identifying emotional states from multi-modal signals as well as for the definition of new methodologies and mathematical models for the automatic modeling and implementation of naturally human-like

communication interfaces.

The present work compare American, French, and Italian subjects on their ability to identify facial and vocal emotional expressions encoded exploiting separately the auditory and visual communication mode by American actors and actress. The results made possible to hypothesize that the ability to recognize emotional expressions as a function of the communication mode is affected by the cultural context and in particular by the language. In fact, American and French subjects are able to perform equally well both on the visual and vocal cues, whereas Italian subjects rely more on visual information. This may suggest that speakers of different languages may exhibit a different sensitivity to vocal emotional information. It is possible that at the base of the emotional encoding there is a more language specific process that is strictly related to the prosodic feature of the subject native language. This hypothesis is supported by previous perceptual data obtained on Italian subjects tested on Italian audio and mute video stimuli showing that vocal information is favored by Italians when tested on emotional stimuli expressed in their native cultural context and native language [13, 16]. Cultural specificity seems do not affect the recognition of emotional visual information: visual channel share visual emotional features across cultures, supporting the data presented in literature.

More data are needed to support the above hypothesis by extending the proposed perceptual experiments to the members of other countries. To this aim, current experiments involve Hungarian and Indian participants, and in the future, we hope to include other European and overseas countries.

.

ACKNOWLEDGMENTS

This work has been supported by the European projects: COST 2102 “Cross Modal Analysis of Verbal and Nonverbal Communication” (cost2102.cs.stir.ac.uk/) and COST TD0904 “TMELY: Time in MEntal activitY (www.timely-cost.eu/).

Acknowledgment goes to three unknown reviewers for their helpful comments and suggestions and to Tina Marcella Nappi for her editorial help.

REFERENCES

[1] V. Aubergé , M. Cathiard: Can we hear the prosody of smile ? *Speech Communication 40*, 87-97, (2003),

[2] B. De Gelder, P. Bertelson: Multisensory integration, levels of processing and ecological validity. *Trends in Cognitive Science 7 (10)*, 460 -467, (2003)

[3] P. Ekman: The argument and evidence about universals in facial expressions of emotion. In H. Wagner, H., Manstead, A. (eds.). *Handbook of social psychophysiology*, Chichester: Wiley, 143-164, (1989)

[4] C.E. Izard: Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115, 288–299, (1994)

[5] R.. Banse, K. Scherer: Acoustic profiles in vocal emotion expression. *Journal of Personality & Social Psychology* 70(3), 614-636, (1996)

[6] K.R. Scherer: Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227-256, (2003)

[7] J.T. Cacioppo, G.G. Berntson, , J.T. Larsen, K.M. Poehlmann, T.A. Ito: The Psychophysiology of emotion. In J.M. Lewis, M. Haviland-Jones (Eds.), *Handbook of Emotions*, 2nd edition, 173-191, New York: Guilford Press, (2000)

[8] P. Ekman, W.V. Friesen, J.C. Hager: *The facial action coding system*. Second edition. Salt Lake City: Research Nexus eBook.London: Weidenfeld & Nicolson, (2002)

[9] C.E Izard, B.P. Ackerman: Motivational, organizational, and regulatory functions of discrete emotions. In J.M. Lewis, M. Haviland-Jones (Eds.), *Handbook of Emotions*, 2nd edition, 253-264, New York: Guilford Press,(2000).

[10] P. Ekman, W.V. Friesen, J.C. Hager: *The facial action coding system* Second edition. Salt Lake City: Research Nexus eBook.London: Weidenfeld & Nicolson, (2002)

[11] M.. Kamachi, M. Lyons, J. Gyoba: Japanese Female Facial Expression Database, Psychology Department in Kyushu University, http://www.kasrl.org/jaffe.html

[12] F. Samaria, A. Harter: The ORL Database of Faces. Cambridge University Press, Cambridge, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[13] A. Esposito: The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. Cognitive Computation Journal, 1(2), 268-278, (2009)

[14] A. Esposito, M.T. Riviello, G. Di Maio: ”The COST 2102 Italian Audio and Video Emotional Database”. In In B. Apolloni, et al. (Eds) Frontiers in Artificial Intelligence and Applications, vol. 204, 51-61, ISBN 978-1-60750-072-8 (print) ISBN 978-1-60750-515-0 (2009), http://www.booksonline.iospress.nl/Content/View.aspx?piid=14188

[15] A. Esposito, M. T. Riviello, N. Bourbakis: Cultural Specific Effects on the Recognition of Basic Emotions: A Study on Italian Subjects. In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 135–148.(2009)

[16] A. Esposito: The amount of information on emotional states conveyed by the verbal and nonverbal channels: Some perceptual data. In Y. Stilianou et al. (Eds): *Progress in Nonlinear Speech Processing*, Lecture Notes in Computer Science, 4392, 245-264, Springer-Verlag, (2007)

[17] P. Ekman, P.: An argument for basic emotions. Cognition and Emotion, 6, 169-200 (1992)

[18] K.R. Scherer, J. S. Oshinsky: Cue utilization in emotion attribution from auditory stimuli. Motivation and Emotion, 1, 331–346 (1977)

[19] D. Ververidis, C. Kotropoulos: Emotional Speech Recognition: Resources, Features and Methods. Elsevier Speech Communication 48(9), 1162–1181, (2006)