2025 IEEE 16th International Conference on Cognitive Infocommunications (CogInfoCom) | 979-8-3503-5692-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/CogInfoCom66819.2025.11200899

The role of speaker gender in vocal emotion recognition

Marialucia Cuciniello, Terry Amorese, Anna Alterio, Daniele Pepe, Anna Esposito
Department of Psychology
Università degli Studi della Campania "Luigi Vanvitelli"
Caserta, Italy

Odette Scharenborg
Multimedia Computing Group
Delft University of Technology
Netherlands

Gennaro Cordasco
Department of Computer Science
Università degli Studi di Salerno
Salerno, Italy

***Abstract*—*This study investigates the impact of stimulus gender on vocal emotion recognition. To that end, 3 groups of listeners were asked to classify joy, neutral state, fear, anger and sadness in Italian meaningful and nonsensical sentences in 2 settings: a controlled laboratory setting to which a group of Dutch listeners was assigned and a naturalistic setting to which 2 groups respectively of Dutch and Italian listeners were assigned. Two levels of background noise were applied to the sentences to increase the level of task difficulty. Results showed an impact of stimulus gender on emotion classification performance, which seemed dependent on the emotion category: anger was better recognized when spoken by male speakers in both experimental settings, while joy was better recognized when spoken by female speakers by Dutch listeners in the laboratory and Italian listeners in the naturalistic setting, and fear was better recognized when spoken by female speakers in the naturalistic setting.*

Keywords*— *Speech recognition; speaker' gender; vocal emotion recognition; background noise

I. INTRODUCTION

Understanding how humans perceive and interpret vocal emotions—especially in the presence of variables such as speaker gender, background noise, and linguistic familiarity—offers valuable insights into the broader dynamics of human cognition and communication. Studies that investigate these perceptual mechanisms contribute to a growing body of knowledge at the intersection of cognitive science and technological systems. In particular, they highlight how cognitive processes adapt to complex auditory environments and how these adaptations can inform the design of systems capable of interpreting human affective states. This convergence of disciplines opens new avenues for enhancing the interaction between natural and artificial agents, fostering more intuitive and emotionally aware communicative technologies. Emotion recognition has attracted great interest amongst researchers as it regulates behaviors by modulating, establishing, and maintaining relationships [1]. Part of this attention has been directed to understanding the role of speaker gender. In this context, most of studies have mostly been focused on the visual channel, e.g., recognizing facial expressions depicting dynamic stimuli of varying emotional intensity [2] or assessing different types of stimuli such as human faces, icons or virtual agents [3,4]. These studies highlighted women greater ability than men in decoding emotions and that female stimuli are recognized more accurately than male ones [4]. The effect of speaker gender on the recognition of vocal emotional expressions has received less attention. The existing research is not conclusive since some studies report no gender differences [5], while others report gender' effects only for specific emotional categories [6]. For instance, it has been shown that anger and fear were better recognized when spoken by a male speaker [7], and neutral [8] and happy expressions [7] were better recognized when conveyed by female voices.

The urgent need for further studies in order to fill these gaps appears tangible. However, well known is the fact that through prosody, semantics or non-linguistic sounds, it is possible to glean emotional information that promotes the interaction and communication process [9,10]. The decoding of speech signals has interested several researchers who performed studies testing different speech processing tasks such as speech and speaker recognition [11,12], as well as speech emotion recognition (SER). Simultaneously, it is possible to assist a whole line of research in constant and fermenting growth. There are studies focused on the identification of sequence-sequence models for automatic speech recognition (ASR) in which speech was considered as the main input mode on mobile technologies and assume a key role in the design of intelligent personal assistants [13]. In addition, no studies have considered speaker gender effects on the recognition of vocal emotional expressions with varying a) background noise (clear, +2dB, and + 5dB), b) experimental setting (laboratory vs. naturalistic settings) and c) knowledge of the language (speakers and non-speakers of the language). These are the effects investigated in the current study.

II. MATERIALS AND METHODS

A. Participants

Group 1, assigned to the laboratory setting consisting of 51 Dutch listeners (26 males; mean age= 22.16; SD=± 3.14) was recruited from the department of Electrical Engineering, Mathematic & Computer Science of Delft University of Technology (NL). Group 2, assigned to the realistic setting, composed of 37 Dutch listeners (22 males, mean age= 21.32; SD=± 2.89) was recruited through personal acquaintances and university libraries in the city of Den Haag (NL). Group 3 consisting of 56 Italian participants (25 males, mean age= 22.21; SD=± 2.85) assigned to the realistic setting was recruited both at the University of Naples "Federico II" and "Vanvitelli" University (UVA), South of Italy. Inclusion criteria for Dutch were no proficiency of the Italian language, whereas for both Dutch and Italian participants, normal hearing, no diagnosis of speech or language disorders were required aspects. All participants declared their free will to join the experiment by signing an informed consent form. The study was approved by the UVA ethics committee with the protocol number 25/2017.

B. Stimuli

The dataset of stimuli exploited in this study derives from the Italian EMOVO corpus [14,15]. A total of 100 recordings was selected consisting of ten semantically different sentences (respectively, 5 meaningful and 5 meaningless sentences), which conveying the emotions of joy, fear, anger and sadness, and a neutral state. The sentences were presented in a clear and two background noise conditions. To realize the noise conditions, after normalizing the intensity, the chatter noise of 8 Italian speakers (4 males and 4 females) was applied to the sentences at two different signal to noise rations (SNR), i.e. SNR -5 dB and SNR +2 dB using a Praat script [16] purposely created. Specifically, this background noise was obtained by randomly selecting 8 Italian sentences from a subset of a different Italian corpus [17] and mixing them together. Details concerning the experimental protocol adopted are described in [18]. Once the stimuli were implemented, 12 different combinations of them were created to obtain lists, to which each participant was randomly assigned. Each list contained 120 sentences; 40 sentences for each condition (i.e., clear, +2dB, and -5dB). Each subset of 40 sentences consisted of 8 sentences per emotion (joy, neutral state, fear, anger, and sadness) expressed respectively by 4 actors and 4 actresses. For each speaker, two expressions had semantically normal content and two were nonsense expressions.

C. Procedure

The listeners assigned to the laboratory setting were seated inside a soundproof booth while those assigned to the realistic setting conducted the experiment in libraries or cafeterias. Each participant was asked to sit in front of a laptop wearing headphones provided by the experimenter which guaranteed good audio quality. The stimuli were randomly presented to each listener which had to categorize the perceived emotion as joy, neutral state, fear, anger, sadness and "I don't know" via a labelled key, on a keyboard. Participants were instructed not to pay attention to the sentences' content but to focus on the emotion with which the sentence was uttered

III. DATA ANALYSIS AND RESULTS

Separate repeated measure's ANOVA were performed for each group of listeners separately using the SPSS 21 IBM software. Five repeated measure's ANOVA were conducted for each emotional category (joy, neutral state, fear, anger, and sadness). Participants' gender was set as between subject, and speaker gender (female and male speakers), and conditions (clear vs +2dB and -5dB levels of background noise) were set as within subject variables. The significance level was set at $\alpha < .05$ and differences among means were assessed by Bonferroni's post hoc tests.

A. Results of Dutch listeners in laboratory setting

Joy recognition

No significant effects of participants' gender [$F(1,49)=.018$, $p=.893$] emerged. A significant effect of stimuli's noise level emerged [$F(2,98)=95.702$, $p<<.01$]. Bonferroni's post hoc tests revealed that in absence of background noise, participants are able to recognize stimuli more accurately (clear mean=2.675; +2dB mean= 2.038; -5dB mean=.983, $p<<.01$). Concerning speaker gender, a significant difference was found [$F(1,49)=18.274$, $p<<.01$]. Bonferroni's post hoc tests revealed that females (mean=2.176) conveying joy, were better recognized than male speakers (mean=1.621, $p<<.01$) (see Figure 1).

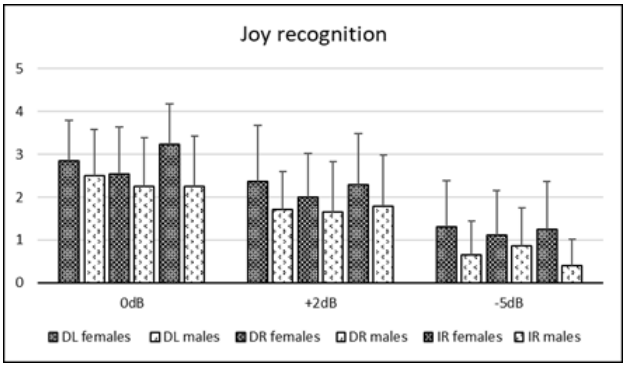

Fig. 1. *DL stands for "Dutch listeners in laboratory", DR for "Dutch listeners in realistic environment", and IR for "Italian listeners in realistic environment". The bars illustrate Joy recognition accuracies split by speaker gender (females and males), for the three noise levels separately (clear, +2 dB and -5 dB, respectively).*

Neutral state recognition

No significant effects of participants' gender [$F(1,49)=.038$, $p=.847$] emerged. A significant effect of stimuli's noise level emerged [$F(2,98)=94.161$, $p<<.01$]. Bonferroni's post hoc tests revealed that in the absence of background noise,

participants were more able to recognize stimuli (clear mean=3.600; +2dB mean=3.463; -5dB mean=2.056, p<<.01). No significant effect of speaker gender [F(1,49)=1.609, p=.211] was found (see Figure 2).

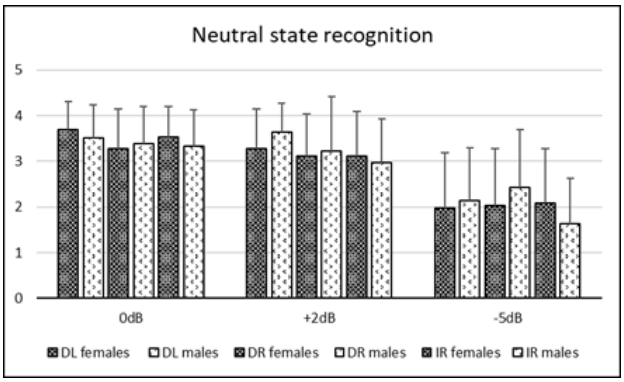

Fig. 2. *DL stands for "Dutch listeners in laboratory", DR for "Dutch listeners in realistic environment", and IR for "Italian listeners in realistic environment". The bars illustrate Neutral recognition accuracies split by speaker gender (females and males), for the three noise levels separately (clear, +2 dB and -5 dB, respectively).*

Fear recognition

No significant effects of participants' gender [F(1,49)=.003, p=.960] emerged. A significant effect of stimuli's noise level emerged [F(2,98)=158.633, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were more accurate in recognizing stimuli (clear mean=2.607, +2dB mean=1.815, -5dB mean=.588, p<<.01). No significant effect of speaker gender [F(1,49)=2.516, p=.119] was found, (see Figure 3).

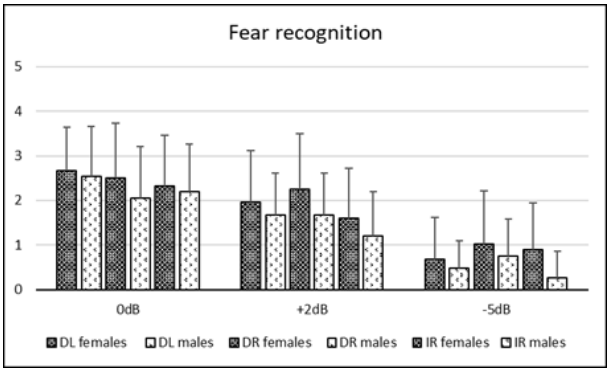

Fig. 3. *DL stands for "Dutch listeners in laboratory", DR for "Dutch listeners in realistic environment", and IR for "Italian listeners in realistic environment". The bars illustrate Fear recognition accuracies split by speaker gender (females and males), for the three noise levels separately (clear, +2 dB and -5 dB, respectively).*

Anger recognition

No significant effects of participants' gender [F(1,49)=.026, p=.873] emerged. A significant effect of stimuli's noise level emerged [F(2,98)=27.634, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were able to recognize stimuli more accurately (clear mean=3.687; +2dB mean=3.393; -5dB mean=2.970, p<.01). Concerning speaker gender, a significant difference was found [F(1,49)=103.527, p<<.01]. Bonferroni's post hoc tests revealed that anger produced by male speakers (mean=3.745) was better recognized than anger produced by female speakers (mean=2.954, p<<.01), (see Figure 4).

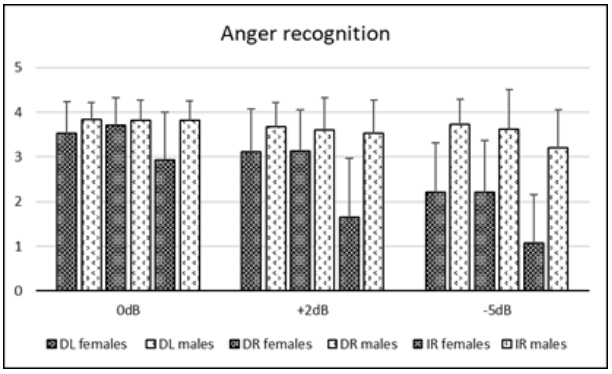

Fig. 4. *DL stands for "Dutch listeners in laboratory", DR for "Dutch listeners in realistic environment", and IR for "Italian listeners in realistic environment". The bars illustrate Anger recognition accuracies split by speaker gender (females and males), for the three noise levels separately (clear, +2 dB and -5 dB, respectively).*

Sadness recognition

No significant effects of participants' gender [F(1,49)=1.935, p=.171] emerged. A significant effect of stimuli's noise level emerged [F(2,98)=44.746, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were able to recognize stimuli more accurately (clear mean=2.424, +2dB mean=1.798, -5dB mean=1.208, p<<.01). No significant effect of speaker gender [F(1,49)=1.072, p=.306] was found, (see Figure 5).

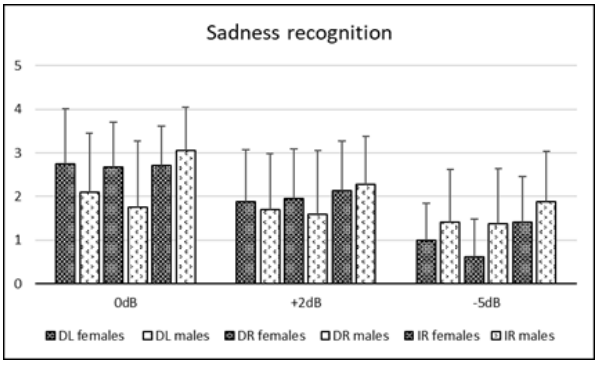

Fig. 5. *DL stands for "Dutch listeners in laboratory", DR for "Dutch listeners in realistic environment", and IR for "Italian listeners in realistic environment". The bars illustrate Sadness recognition accuracies split by speaker gender (females and males), for the three noise levels separately (clear, +2 dB and -5 dB, respectively).*

To sum up, Dutch listeners' performances in laboratory setting were significantly better in clear condition. Only joy and anger were significantly affected by speaker gender, with

joy better recognized when portrayed by female speakers and anger better recognized when portrayed by male speakers.

B. Results of Dutch listeners in realistic setting

Joy recognition

No significant effects of participants' gender [F(1,35)=.100, p=.763] emerged. A significant effect of stimuli's noise level emerged [F(2,70)=59.177, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were able to recognize stimuli more accurately (clear mean=2.409, +2dB mean=1.805, -5dB mean=.967, p<<.01). No significant effect of speaker gender [F(1,35)=3.091, p=.087] was found (see Figure 1).

Neutral state recognition

No significant effects of participants' gender [F(1,35)=2.281, p=.140] emerged. A significant effect of stimuli's noise level emerged [F(2,70)=27.654, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were able to recognize stimuli more accurately (clear mean=3.389, +2dB mean=3.179, -5dB mean=2.230, p<<.01). No significant effect of speaker gender [F(1,35)=1.862, p=.181] was found (see Figure 2).

Fear recognition

No significant effects of participants' gender [F(1,35)=.007, p=.934] emerged. A significant effect of stimuli's noise level emerged [F(2,70)=37.023, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were more accurate in stimuli recognizing (clear mean=2.260, +2dB mean=1.923, -5dB mean=.957, p<.05). A significant effect of speaker gender [F(1,35)=8.453, p=.006] was found. Bonferroni's post hoc tests revealed that female speakers (mean=1.918) portraying fear were more accurately recognized than male speakers (mean=1.509, p=.006), (see Figure 3).

Anger recognition

No significant effects of participants' gender [F(1,35)=.335, p=.567] emerged. A significant effect of stimuli's noise level emerged [F(2,70)=26.970, p<<.01]. Bonferroni's post hoc tests revealed that participants were more accurate in stimuli recognition in absence of background noise (clear mean=3.769, +2dB mean=3.365, -5dB mean=2.932, p<.01). A significant effect of speaker gender [F(1,35)=32.759, p<<.01] was found. Bonferroni's post hoc tests revealed that male speakers portraying anger (mean=3.671) were better recognized than female speakers (mean=3.040, p<<.01), (see Figure 4).

Sadness recognition

No significant effects of participants' gender [F(1,35)=1.362, p=.251] emerged. A significant effect of stimuli's noise level emerged [F(2,70)=32.387, p<<.01]. Bonferroni's post hoc tests revealed that participants were able to recognize stimuli more accurately in absence of background noise (clear mean=2.267, +2dB mean=1.807, -5dB mean=1.005, p<.05). No significant effect of speaker gender [F(1,35)=.999, p=.324] was found, (see Figure 5). To sum up, Dutch listeners' performances in realistic setting were significantly better in clear condition. Only fear and anger were significantly affected by speaker gender, with fear better recognized when portrayed by female speakers and anger better recognized when portrayed by male speakers.

C. Results of Italian listeners in realistic setting

Joy recognition

No significant effects of participants' gender [F(1,54)=2.427, p=.125] emerged. A significant effect of stimuli's noise level emerged [F(2,108)=87.266, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants were able to recognize stimuli more accurately (clear mean=2.754; +2dB mean=2.005; -5dB mean=.812, p<<.01). Concerning speaker gender, a significant difference was found [F(1,54)=64.377, p<<.01]. Bonferroni's post hoc tests revealed that female speakers (mean=2.244) producing joy were better recognized than male speakers (mean=1.464, p<<.01), (see Figure 1).

Neutral state recognition

No significant effects of participants' gender [F(1,54)=.003, p=.960] emerged. A significant effect of stimuli's noise level emerged [F(2,108)=66.077, p<<.01]. Bonferroni's post hoc tests revealed that participants were able to recognize stimuli more accurately in the absence of background noise (clear mean=3.447; +2dB mean=3.021; -5dB mean=1.861, p<.01). Concerning speaker gender, a significant difference was found [F(1,54)=5.980, p=.018]. Bonferroni's post hoc tests revealed that female speakers (mean=2.905) portraying a neutral state were better recognized than male speakers (mean=2.648, p=.018), (see Figure 2).

Fear recognition

No significant effects of participants' gender [F(1,54)=2.055, p=.157] emerged. A significant effect of stimuli's noise level emerged [F(2,108)=95.605, p<<.01]. Bonferroni's post hoc tests revealed that stimuli were recognized more accurately in the absence of background noise (clear mean=2.240; +2dB mean=1.400; -5dB mean=.571, p<<.01). Concerning speaker gender, a significant difference was found [F(1,54)=12.332, p=.001]. Bonferroni's post hoc tests revealed that female speakers (mean=1.594), portraying fear were better recognized than male speakers (mean=1.213, p=.001), (see Figure 3).

Anger recognition

Significant effects of participants' gender [F(1,54)=5.133, p=.028] emerged. Bonferroni's post hoc tests revealed that female participants (mean=2.844) better recognized stimuli expressing anger than male participants (mean=2.520, p=.028). A significant effect of stimuli's noise level emerged [F(2,108)=61.734, p<<.01]. Bonferroni's post hoc tests revealed that in the absence of background noise, participants

were more able to recognize stimuli (clear mean=3.358; +2dB mean=2.571; -5dB mean=2.117, p<<.01). Concerning speaker gender, a significant difference was found [F(1,54)=199.278, p<<.01]. Bonferroni's post hoc tests revealed that male speakers (mean=3.501), portraying anger were better recognized than female speakers (mean=1.863, p<<.01), (see Figure 4).

Sadness recognition

Significant effects of participants' gender [F(1,54)=7.069, p=.010] emerged. Bonferroni's post hoc tests revealed that female participants (mean=2.425) better recognized sad vocal stimuli than male participants (mean=2.027, p=.010). A significant effect of stimuli's noise level emerged [F(2,108)=49.339, p<<.01]. Bonferroni's post hoc tests revealed that stimuli were recognized more accurately in the absence of background noise (clear mean=2.866; +2dB mean=2.201; -5dB mean=1.610, p<<.01). Concerning speaker gender, a significant difference was found [F(1,54)=9.964, p=.003]. Bonferroni's post hoc tests revealed that male speakers (mean=2.395) portraying sadness were better recognized than female speakers (mean=2.056, p=.003), (see Figure 5).

To sum up, Italian listeners' performances in realistic setting were significantly better in clear condition and the speaker gender depending on the emotional categories. Joy, neutral state, and fear were better recognized when portrayed by female speakers and anger and sadness were better recognized when portrayed by male speakers. Additionally, female Italian listeners were more accurate than male listeners in the recognition of anger and sadness.

IV. DISCUSSION

The present work investigated potential differences related to speaker gender in vocal emotions recognition. The results showed that Dutch listeners, were more accurate in recognizing anger portrayed by male speakers, while in the laboratory setting, they better recognized joy when conveyed by female speakers, in line with [7, 22] and in the realistic setting, they better discriminated fear when conveyed by female speakers, as observed by [18]. The effect of speakers' gender was more effective for the Italian group acting in realistic setting. Italian listeners were able to better recognize joy, neutral state and fear when portrayed by female speakers and anger and sadness when portrayed by male speakers. Interestingly, anger was better recognized when produced by male speakers than by female speakers by all participants involved. This result is in line with [7], who suggested that vocal expressions of anger are more tied with the gender than any other emotional category investigated. Concerning listeners' gender, outcomes for Dutch in both realistic and laboratory setting showed no effect of listeners' gender, while female Italian listeners were more accurate than male listeners in the recognition of anger and sadness.

The Dutch results are in line previous findings showing that male and female listeners are equally accurate in decoding anger and neutral expressions [18,19,20,21,22]. More consistent with the line of research supporting that female listeners are better at vocal emotion recognition [23,24,25,26,27] are the results obtained partly by the Italian group, even though the outcome concern only the recognition of two emotions (sadness and anger) out of five. As regards the difficulty of the task implemented by increasing the levels of background noise, all listeners were more accurate in the clear condition, worsen with increasing SNRs which is in line with [28,29]. Taken together, these results seem to demonstrate that lack of knowledge of the language did not hinder the emotional recognition of Dutch listeners, which is in line with existing evidence assuming that vocal expressions of anger, fear, sadness and joy are accurately categorized when listening to a foreign language [30,31]. Seemingly, these emotions possess acoustic- perceptive properties that are not affected by the language with which they are conveyed [32]. A further possible explanation could be linked to the duration of the emotional expressions administered in line with results presented in [33]. In this work focusing on the time course underlying the conscious recognition of basic emotions from vocal expressions, it was shown that anger, sadness, fear and neutral expressions are recognized more accurately at short intervals than happiness [33]. The possible influence of the experimental setting cannot be completely excluded since the lack of involvement of a group of Italians acting in a laboratory setting represents a limitation of the present research. Future research could include audio performances related to disgust and surprise to broaden knowledge and further investigate possible differences in the recognition of vocally expressed emotion. In conclusion, observing the outcomes, the present study seems to fall within the line of research which maintains that effects of speaker gender occur only for specific emotional categories, that noise is disruptive for speech and more for the recognition of emotional vocal expressions, and that experimental setting differently affects listeners' decoding ability.

The findings of this study offer meaningful contributions to the field of Cognitive Info-Communications, which explores the interaction between cognitive processes and information technologies. By examining how speaker gender, background noise, and linguistic familiarity influence vocal emotion recognition, this research provides insights into how human cognitive mechanisms operate in complex auditory environments. These insights are particularly relevant for the development of affect-aware communication systems, such as intelligent virtual agents and social robots, which must interpret emotional cues accurately to engage in naturalistic and adaptive interactions with users.

The observed variability in emotion recognition based on speaker gender suggests that future systems should incorporate demographic and contextual sensitivity to enhance emotional understanding.

Moreover, the finding that non-native listeners were able to recognize emotions effectively despite language barriers supports the feasibility of cross-linguistic emotion recognition modules, which are essential for globalized human-machine communication. These results align with the CogInfoCom vision of enabling seamless and intuitive communication between natural and artificial cognitive agents by embedding human-like perceptual and interpretive capabilities into technological systems.

ACKNOWLEDGMENT

This research received funding by the EU-H2020 program, grant No. 101182965 (CRYSTAL), EU NextGenerationE PNRR Mission 4 Component 2 Investment 1.1 – D.D 1409 del 14-09-2022 PRIN 2022 – UNDER the IRRESPECTIVE project, code P20222MYKE - CUP: B53D23025980001 and PNRR MUR under AI-PATTERNS FAIR Project CUP:E63C22002150007.

REFERENCES

[1] D. W. Murray, K. D. Rosanbalm, C. Christopoulos, and A. Hamoudi, "Self-regulation and toxic stress: Foundations for understanding self-regulation from an applied developmental perspective.", 2015.

[2] T. S. Wingenbach, C. Ashwin, and M. Brosnan, “Sex differences in facial emotion recognition across varying expression intensity levels from videos”. PLoS one, 13(1), e0190634, 2018.

[3] A. H. Fischer, M. E. Kret, and J. Broekens, “Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis.” PloS one, 13(1), e0190712, 2018.

[4] T. Amorese, M. Cuciniello, A. Vinciarelli, G. Cordasco, and A. Esposito, “Synthetic vs Human Emotional Faces: What Changes in Humans’ Decoding Accuracy”. IEEE Transactions on Human-Machine Systems, 52(3), 390-399, 2021.

[5] M. T. Riviello, and A. Esposito. “On the perception of dynamic emotional expressions: A cross-cultural comparison”. Vol. 6. Springer, 2016.

[6] S. T. Hawk, G. A. Van Kleef, A. H. Fischer, and J. Van Der Schalk, " Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. Emotion, 9(3), 293, 2009.

[7] T. L. Bonebright, J. L. Thompson, and D. W. Leger, “Gender stereotypes in the expression and perception of vocal affect”. Sex Roles, 34, 429-445, 1996.

[8] K. S. Young, C. E. Parsons, R. T. LeBeau, B. A. Tabak, A. R. Sewart, A. Stein, and M. G. Craske, “Sensing emotion in voices: Negativity bias and gender differences in a validation study of the Oxford Vocal (‘OxVoc’) sounds database”. Psychological assessment, 29(8), 967, 2017.

[9] M. W. Kraus, "Voice-only communication enhances empathic accuracy." American Psychologist 72(7), 644., 2017.

[10] J. Fischer, and T. Price. "Meaning, intention, and inference in primate vocal communication." Neuroscience & Biobehavioral Reviews 82, 22-31., 2017.

[11] M. Lee, J. Lee, and J.H. Chang, “Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition”. Digital Signal Processing, 85, 1-9, 2019.

[12] P. Dhakal, P. Damacharla, A.Y. Javaid, and V. Devabhaktuni, “A near real-time automatic speaker recognition architecture for voice-based user interface”. Machine learning and knowledge extraction, 1(1), 504-520, 2019.

[13] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, “Personalized speech recognition on mobile devices,” in Proc. of ICASSP, pp. 5955–5959, IEEE, 2016.I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, “Personalized speech recognition on mobile devices,” in Proc. of ICASSP, pp. 5955–5959, IEEE, 2016.

[14] Corpus downloaded from: http://voice.fub.it/activities/corpora/emovo/index.html

[15] G. Constantini, I. Iadarola, A. Paoloni, and M. Todisco, “EMOVO corpus: an Italian emotional speech database”, Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014.

[16] P. Boersma and D. Weenink, D. “Praat: doing phonetics by computer [Computer program]”, 2013. Retrieved from http://www.praat.org/

[17] Corpus downloaded from: http://www.clips.unina.it/en/index.jsp

[18] O. Scharenborg, S. Kakouros, and J. Koemans, “The effect of noise on emotion perception in an unknown language.” In Proceedings of the International Conference on Speech Prosody (pp. 364-368), 2018.

[19] T. X. Fujisawa, and K Shinohara, (2011). “Sex differences in the recognition of emotional prosody in late childhood and adolescence”. The Journal of Physiological Sciences, 61(5), 429-435, 2011.

[20] L. Lambrecht, B. Kreifelts, and D. Wildgruber, “Gender differences in emotion recognition: Impact of sensory modality and emotional category”. Cognition & emotion, 28(3), 452-469, 2014.

[21] L. R. Demenescu, Y. Kato, and K. Mathiak, “Neural processing of emotional prosody across the adult lifespan”. BioMed Research International, 2015.

[22] B. Zupan, D. Babbage, D. Neumann, and B. Willer, “Sex differences in emotion recognition and emotional inferencing following severe traumatic brain injury”. Brain Impairment, 18(1), 36-48, 2017.

[23] O. Collignon, S. Girard, F. Gosselin, D. Saint-Amour, F. Lepore, and M. Lassonde, “Women process multisensory emotion expressions more efficiently than men”. Neuropsychologia, 48(1), 220-225, 2010.

[24] K. R. Scherer, and U. Scherer, “Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index”. Journal of Nonverbal Behavior, 35, 305-326, 2011.

[25] S. Paulmann, and A. K. Uskul, “Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners”. Cognition & emotion, 28(2), 230-244, 2014.

[26] N. Keshtiari, and M. Kuhlmann, “The effects of culture and gender on the recognition of emotional speech: Evidence from Persian speakers living in a collectivist society”. International Journal of Society, Culture & Language, 4(2), 71, 2016.

[27] A. Lausen, and A. Schacht, “Gender differences in the recognition of vocal emotions”. Frontiers in psychology, 9, 882, 2018.

[28] O. E. Scharenborg, E. Kolkman, S. Kakouros, and B. M. B. Post, “The effect of sentence accent on non-native speech perception in noise”, Interspeech, pp. 863-867, 2016.

[29] O. Scharenborg, J. M. Coumans, R. van Hout, “The effect of background noise on the word activation process in non-native spoken- word recognition”, Journal of Experimental Psychology: Learning, Memory, and Cognition, 44(2), 233, 2018.

[30] W. F. Thompson and L. L. Balkwill, “Decoding speech prosody in five languages”, 407-424, 2006.

[31] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, “Recognizing emotions in a foreign language”. Journal of Nonverbal Behavior, 33, 107-120, 2009.

[32] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, “Factors in the recognition of vocally expressed emotions: A comparison of four languages”. Journal of Phonetics, 37(4), 417-435, 2009.

[33] M. D. Pell and S. A. Kotz, “On the time course of vocal emotion recognition”. PLoS One, 6(11), e27256, 2011.