CHALMERS
UNIVERSITY OF TECHNOLOGY

Neutral by Default? Replicating User Vocal Responses to Negative Affective Cues in Conversational Agents

Downloaded from: https://research.chalmers.se, 2026-10-12 12:04 UTC

Citation for the original published paper (version of record):
Ma, Y., Zhang, Y., Fu, D. et al (2026). Neutral by Default? Replicating User Vocal Responses to Negative Affective Cues in
Conversational Agents. Hri 2026 Proceedings of the 21st ACM IEEE International Conference on Human Robot Interaction: 1268-1272. http://dx.doi.org/10.1145/3757279.3788802

N.B. When citing this work, cite the original published paper.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

(article starts on next page)

Neutral by Default? Replicating User Vocal Responses to Negative Affective Cues in Conversational Agents

Yong Ma
University of Bergen
Bergen Norway
yong.ma@uib.no

Yuchong Zhang*
KTH Royal Institute of Technology
Stockholm Sweden
yuchongz@kth.se

Di Fu
University of Surrey
Guildford United Kingdom
d.fu@surrey.ac.uk

Stephanie Zubicueta Portales
Norwegian University of Science and Technology
Trondheim/Gjovik Norway
stephanieportales@yahoo.com

Morten Fjeld
University of Bergen
Bergen Norway
Chalmers University of Technology
Gothenburg Sweden
morten.fjeld@uib.no,fjeld@chalmers.se

Abstract

Conversational agents (CAs) increasingly detect users' emotions, yet deciding how to respond, especially to negative affect, remains a central design challenge. We conducted a role-switching study in which participants reply as the CAs to simulated users expressing anger, sadness, or fear. Results reveal systematic, gender-linked patterns: most male participants favored a neutral, affect-balanced stance and prioritized clarification or task progress, whereas most female participants produced a wider range of non-neutral responses, more often using explicit empathy, reassurance, and reflective listening. We also observe differences in de-escalation phrasing, validation timing, and follow-up questioning across scenarios. These findings indicate that strategies for handling negative emotions vary with user characteristics and context. Based on these findings, we argue for adaptive CA response policies that calibrate first-turn acknowledgment and information-gathering, tailoring prosody and wording to emotional context in order to support de-escalation, perceived understanding, and user trust.

CCS Concepts

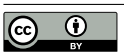
• **Human-centered computing** → **Empirical studies in HCI**; *Interface design prototyping*; **User studies**; • **Social and professional topics** → **User characteristics**.

Keywords

Speech emotion responses, negative affect, conversational agents, user characteristics

ACM Reference Format:
Yong Ma, Yuchong Zhang, Di Fu, Stephanie Zubicueta Portales, and Morten Fjeld. 2026. Neutral by Default? Replicating User Vocal Responses to Negative Affective Cues in Conversational Agents. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3757279.3788802

*Corresponding author.

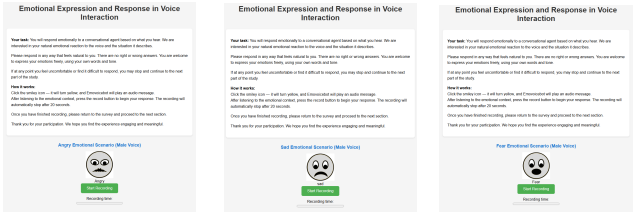


This work is licensed under a Creative Commons Attribution 4.0 International License.
HRI '26, Edinburgh, Scotland, UK
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788802

1 Introduction

Conversational agents (CAs) have become integral to everyday life, supporting tasks ranging from information retrieval to smart-home control [12, 26]. As voice user interface (VUI) technologies mature, user expectations increasingly extend beyond utility toward socially attuned interaction. In particular, the ability to recognize, understand, and appropriately respond to affect has emerged as a key requirement for trustworthy and satisfying experiences. Recent advances in speech processing and affective computing enable CAs to infer users' emotional states from vocal cues. Speech emotion recognition (SER) leverages acoustic features such as pitch, prosody, and temporal patterns [5, 21, 29], while contemporary AI, including large language models and multimodal systems, has improved the robustness of these inferences [17, 18, 24]. In parallel, emotion synthesis supports expressive speech generation aligned with conversational intent [30]. Together, these developments position CAs for more emotionally aware interaction [16, 22, 34]. Yet recognition alone is insufficient; designing appropriate responses to negative affect, such as anger, sadness, or fear, remains a persistent challenge [15]. Emotional expression varies across cultures, genders, and contexts [7, 8, 27], and real-world conditions such as accents or background noise further complicate interpretation [3]. Even when emotions are detected correctly, empathetic responses that are mistimed, intrusive, or stylistically artificial can undermine trust and user satisfaction [1, 2]. Despite increasing experimentation with supportive behaviors in commercial systems (e.g., Alexa) [4, 25], the field lacks fine-grained empirical guidance on what to say next and how to say it in situ.

We argue that progress requires shifting emphasis from detection accuracy to interactional competence: the stance, sequencing, and speech activities that de-escalate, validate, and support users during emotionally charged moments. To investigate this, we adopt a role-switching paradigm [28] in which participants respond as the CA to users' negative emotional expressions [14, 20]. Our study explicitly replicates the methodology of Ma et al. [20] and extends it through gender-balanced recruitment, updated scenarios, the inclusion of female agent voices, and advanced speech analysis. We implement



(a) Anger scenario (b) Sad scenario (c) Fear scenario

Figure 1: Web pages used to elicit participant responses across three negative-emotion scenarios.

this paradigm in a web-based study (Figure 1) in which participants produce spontaneous vocal responses to emotionally expressive CAs. The system records speech for analysis of acoustic features and sentiment, enabling us to observe how response strategies—such as neutrality, explicit empathy, reassurance, information gathering, and action proposals—vary across individuals and contexts.

Our findings reveal a consistent preference for neutral or gently positive stances when responding to negative affect, alongside systematic variation in speech activities and timing. These patterns align with theories of interpersonal emotion regulation, including cognitive reappraisal and expressive suppression [11, 31]. We also observe context- and gender-linked differences, underscoring the need for adaptive response policies rather than one-size-fits-all empathy. This paper contributes: (1) a role-switching method for eliciting concrete human response strategies to negative emotions in CA-style interactions; (2) empirical analysis of stance and speech activities relevant to de-escalation and support; and (3) design implications for adaptive response models emphasizing calibrated first-turn acknowledgment, context-sensitive validation, purposeful information gathering, and safe action proposals. By grounding recommendations in observed human practice, we move beyond emotion recognition toward the practical question of what to say next when users are distressed.

2 Study Design

We conducted an online study to examine how people naturally respond when a CA expresses negative emotions, aiming to derive human-inspired strategies for more adaptive, empathic CAs.

2.1 Negative Emotional Scenarios

We designed three distinct negative-emotion scenario — **Angry** (Fig. 1a), **Sad** (Fig. 1b), and **Fear** (Fig. 1c) —representing common, high-impact situations users might plausibly face. Scenarios were selected through a structured process of brainstorming and group consensus voting to ensure they were potent and relatable. The final three scenarios were:

- **Angry emotional scenario:** "I am betrayed by a close friend or relative."
- **Sad emotional scenario:** "I see children suffering from disease, sickness, or war."
- **Fear emotional scenario:** "I am walking alone in the woods at night when I stumble upon a dead body; the blood seems fresh, and I hear a branch snap behind me."

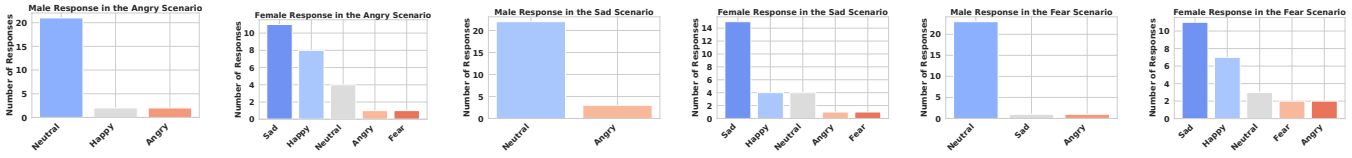
2.2 Participants and Apparatus

An a priori power analysis was conducted using G*Power to determine the required sample size. Assuming a desired power of .80, an alpha level of .05, and a medium effect size ($d = 0.5$) informed by prior role-switching research, the analysis indicated a minimum of 44 participants. Our final gender-balanced sample ($N = 50$; 25 female, 25 male) exceeds this requirement, providing additional robustness to variability in online audio recordings and increased statistical sensitivity through repeated measures, yielding 150 total speech responses (three emotional scenarios per participant). The study materials are available online [1]. Participants were recruited via the Prolific platform[2] and screened for native English proficiency, ages 20–50, and no self-reported hearing or vision impairments. From an initial pool of 60 respondents, 10 were excluded due to missing, incorrect, or unintelligible audio data. The final sample had a mean age of 30.19 years for women and 28.51 years for men. The study was accessed via a web link[3] and hosted on Qualtrics[4]. Spoken responses were analyzed using a multi-tool pipeline. Affective states were inferred using Vokaturi[5], and acoustic features were extracted using Librosa[6] and openSMILE[7]. Participants used their computers' built-in microphones, and recordings were collected as 48 kHz mono WAV files and uploaded for offline analysis. All participants provided informed consent and received £4.50 compensation. The study was conducted anonymously, with data collection limited to voice recordings, questionnaire responses, and basic demographics (age and gender), as described in the study's privacy notice. Questionnaire data are analyzed and reported separately in related work [23, 34].

2.3 Experimental Procedure

Participants responded to three negative-emotion scenarios presented in randomized order. Upon landing on the study page (Fig. 1), participants were first shown an overview of the study and step-by-step instructions. The application guided them through a microphone check, requiring browser permission and verification that their audio hardware was functioning correctly. The core task involved responding as if they were CAs. For each scenario, participants listened to a short, contextualized audio clip spoken by either a male or female agent voice, counterbalanced across participants. After each clip, participants were instructed to produce a spontaneous spoken response aimed at de-escalating or comforting the distressed agent. Recording was initiated by clicking a *Start Recording button*, and participants had up to 20 seconds to speak. Responses were unguided, and participants were free to say anything they felt was appropriate. Following each scenario, participants completed a brief questionnaire about their interaction experience. Demographic information (age and gender) was collected once at the end of the study. Upon completion, participants were shown a debriefing screen indicating that the study had concluded.

[1] https://github.com/WAM-YOMAR/Emotion_Response_Demo
[2] https://www.prolific.com/
[3] https://emo-voice.free.nf/
[4] https://www.qualtrics.com/
[5] https://developers.vokaturi.com/getting-started/overview
[6] https://librosa.org/doc/latest/feature.html
[7] https://audeering.github.io/opensmile-python/



(a) Anger — male (b) Anger — female (c) Sadness — male (d) Sadness — female (e) Fear — male (f) Fear — female

Figure 2: Dominant emotions in participants' spoken responses across scenarios and gender. From left to right: anger (male, female), sadness (male, female), and fear (male, female).

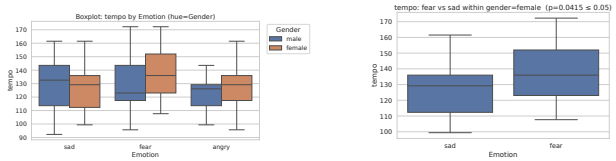
3 Results

3.1 Emotion Expressed in Participants' Responses

Automatic analysis of participants' spoken replies revealed a tendency toward *neutral* or *mildly positive* affect across all three scenarios. Across responses, Vokaturi outputs showed that neutral responses were most frequent overall, followed by calm or positive affect, while explicitly negative counter-affect (e.g., responding to anger with anger or fear) was rare. This pattern held consistently across *Angry*, *Sad*, and *Fear* scenarios, replicating the dominant finding reported in prior role-swapping work [20]. The *Sad* condition elicited comparatively more calm or soothing responses than *Angry* or *Fear*, suggesting greater affective alignment in sadness-related contexts [20]. Beyond this global trend, the distributions in Fig. 2 reveal several scenario-specific and gender-linked nuances:

- **Male responses concentrate on neutrality.** In both *Sad* and *Fear* contexts (top-left and top-right panels), male participants predominantly produced neutral replies, with only small tails of other affects. In *Angry*, males again favored neutrality, with occasional brief positive turns (top-center).
- **Female responses spread across non-neutral categories.** For *Angry* and *Fear* contexts (bottom-center and bottom-right), female participants displayed a broader mix: alongside "Neutral," we observe substantial "Sad" and "Happy" categories. In *Sad*, many female participants matched the scenario with "Sad" responses but also added "Happy" and "Neutral" turns (bottom-left), suggesting supportive alignment followed by uplift.
- **Cross-valence regulation is common.** Especially among female participants, responses to *Angry* often included *sad* or *happy* affect (bottom-center), a pattern consistent with comfort-oriented reframing (acknowledging harm, then offering gentle positivity). For *Fear*, we similarly see a mixture of steadying ("Neutral") and encouraging ("Happy") responses (bottom-right).
- **Scenario effects.** *Sad* scenes drew the largest share of soothing or matching-affect replies (e.g., calm/sad-to-sad), whereas *Angry* drew more boundary-setting "Neutral" turns, and *Fear* elicited stabilizing neutrality with selective positive lifts.

Overall, these patterns suggest that many participants, especially men, default to an affect-balanced stance as a de-escalation tactic, whereas women more often engage in *affective scaffolding*: briefly matching or acknowledging the negative state (e.g., "Sad" to anger), then shifting toward gentle positivity. The relative rarity of overtly negative counter-affect (e.g., responding to anger with anger) indicates a general preference for regulation over confrontation.



(a) Tempo by emotion and gender (b) Female: fear vs. sadness ($p = 0.0415$)

Figure 3: Speech tempo distributions: (a) overall by gender; (b) significant within-gender comparison (female).

Design implications. For CA response policies, (i) a neutral first turn is a safe default across scenarios; (ii) calibrated affect matching (particularly for sadness) followed by a soft positive move can aid regulation; and (iii) policies should allow adaptive mixing (acknowledge → reframe → propose) rather than enforcing a single empathic style. These observations motivate controllable parameters for stance (neutral vs. expressive), timing (when to shift from validation to guidance), and allowable cross-valence moves.

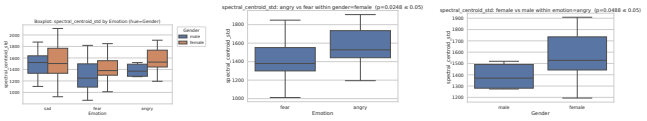
3.2 Speech Features Analysis

We analyzed two prosodic indicators linked to perceived arousal and vocal "brightness": *tempo* (syllabic/speech-rate proxy) and *spectral-centroid variability* (`spectral_centroid_std`). Distributions are shown in Fig. 3 and Fig. 4. Because features were non-normally distributed, Wilcoxon signed-rank tests (within-participant) and Mann–Whitney U tests (between-gender) were used, with exact statistics, effect sizes, and 95% confidence intervals reported.

Tempo. Across emotions, tempo tended to rise with scenario arousal, with *fear* generally faster than *sad* (Fig. 3a). Within-gender analysis revealed a significant increase for *female* participants: *fear* > *sad* ($p = 0.0415$; Fig. 3b). This pattern is consistent with a reassuring-yet-alert stance in high-arousal contexts (fear), whereas *sad* elicited slower, more measured delivery.

Spectral "brightness" variability. `Spectral_centroid_std` was higher for *angry* than for *sad*, and intermediate for *fear* (Fig. 4a), indicating greater fluctuation of high–frequency energy when responding to anger. Within females, variability was *greater for angry than fear* ($p = 0.0248$; Fig. 4b), suggesting more dynamic timbral modulation (e.g., sharper onsets, brighter bursts) during anger de-escalation attempts. Within the *angry* condition, *female* responses showed higher variability than *male* responses ($p = 0.0488$; Fig. 4c),

aligning with the broader distribution of non-neutral strategies observed.



(a) By emotion and gender **(b) Female: anger vs. fear ($p = 0.0248$)** **(c) Anger: female vs. male ($p = 0.0488$)**

Figure 4: Spectral-centroid variability in spoken responses. (a) Overall distributions by emotion with gender as hue; (b) within females, *angry* shows higher variability than *fear*; (c) within *angry*, females show higher variability than males.

Implications for CA design. (1) A modest *tempo increase* can be an appropriate default for fear-like contexts, while *tempo reduction* supports soothing responses in sadness. (2) Allowing controlled *brightness variability* (via prosody shaping or synthesis parameters) may help de-escalate anger without sounding flat. (3) Because prosodic preferences vary by user and context, response policies should adapt these parameters (tempo, pitch/energy dynamics) rather than applying a single "empathic" profile.

4 Discussion

4.1 Summary and Design Implications

Viewed as a theory confirming partial replication, our findings converge with prior role swapping research on emotion aware CAs [20]. Consistent with earlier studies, when responding as a CA, participants predominantly adopted neutral or mildly positive first turn strategies rather than mirroring negative affect. This pattern held across emotional scenarios, agent voice gender, and acoustic analyses, reinforcing neutrality as a reliable baseline for de escalation. Meanwhile, our results refine existing insights. Whereas prior systems oriented work has emphasized explicit empathic responses driven by emotion recognition pipelines [13], our findings show that human responders often regulate affect through neutrality and controlled prosodic modulation. Gender balanced sampling revealed systematic variation: female participants more frequently used explicit empathy and affective variation, while male participants tended toward neutral, task oriented replies. These differences were reflected in both stance choices and prosodic features, including increased tempo in fear contexts and greater spectral centroid variability in responses to anger.

Overall, these patterns indicate that effective responses to negative affect are strategic and context sensitive rather than purely reactive. Neutrality serves as a safe default, but successful interactions often involve calibrated transitions from acknowledgment to validation and information gathering, shaped by emotional context and user characteristics. This aligns with HCI research showing that users value appropriateness, restraint, and perceived intent in emotionally expressive technologies [10]. For CA design, these findings argue for prioritizing interactional competence over emotion recognition alone. Adaptive response strategies should combine linguistic choices with prosodic control to support de escalation and human aligned interaction.

4.2 Limitations and Future Directions

This study has several limitations. The sample size is modest and drawn from an online population, and the scenarios are necessarily scripted, which may limit ecological validity. In addition, affect labeling relies on automated tools and should be interpreted as indicative rather than definitive at the individual level. Gender is treated descriptively and should not be essentialized or interpreted as a fixed determinant of behavior.

Future work will extend this role-swapping paradigm to richer, multi-turn interactions and a broader range of emotional states (e.g., frustration or guilt). Evaluations with deployed CAs would enable assessment of behavioral transfer, while longitudinal studies could examine how adaptive response strategies influence trust, perceived understanding, and successful de-escalation over time. Incorporating cultural and contextual diversity will further strengthen the generalizability of these findings.

4.3 Beyond Virtual CAs: Implications for Embodied Interaction in HRI

Although this study focuses on virtual CAs, the findings have direct relevance for HRI and embodied AI systems [6, 9, 32], where social presence, multimodality, and safety constraints play a central role in shaping interaction [19, 33, 35]. In embodied settings, a neutral first turn remains a safe default, but physical embodiment introduces additional channels, including gaze, posture, and motion, to signal stance and regulate interaction timing. For sadness, slower speech paired with softened gaze and reduced movement may support affect regulation; for fear, slightly increased tempo combined with orienting movements can convey stabilizing alertness; and for anger, vocal brightness modulation should be balanced with non-threatening posture and interpersonal distance. Together, these considerations highlight the importance of coordinating verbal and nonverbal behaviors in HRI and suggest that personalization of emotion-aware response strategies should extend across modalities when translating emotion-aware response strategies from virtual agents to physically embodied systems.

5 Conclusion

This paper demonstrates that effective human responses to emotional distress are not about perfect recognition, but about competent response—orchestrating stance, sequencing, and prosody. Our role-swapping study revealed a consistent preference for a neutral, calming baseline, with systematic variations in expressivity linked to both emotion type and user gender. These human strategies provide a concrete blueprint for moving beyond one-size-fits-all empathy in CAs. The critical design shift is towards building CAs that adapt their language and, just as importantly, their vocal delivery, to the specific context and user, ultimately fostering interactions that are not just accurate but practically trustworthy.

Acknowledgments

This work was partially funded by the Research Council of Norway (326907), the Swedish Foundation for Strategic Research (FUS21-0067), the S-FACTOR project from NordForsk, and the HORIZON-CL4-2021-HUMAN-01 ELSA project.

References

[1] Mohamed Hussein Ramadan Atta, Mervat Mostafa El-Gueneidy, and Ola Ahmed Rashad Lachine. 2024. The Influence of an Emotion Regulation Intervention on Challenges in Emotion Regulation and Cognitive Strategies in Patients with Depression. *BMC Psychology* 12, 1 (2024), 496. doi:10.1186/s40359-024-01949-6

[2] Matthias Berking and Peggilee Wupperman. 2012. Emotion Regulation and Mental Health: Recent Findings, Current Challenges, and Future Directions. *Current Opinion in Psychiatry* 25, 2 (2012), 128–134. doi:10.1097/YCO.0b013e3283503669

[3] Yekta Said Can, Bhargavi Mahesh, and Elisabeth André. 2023. Approaches, Applications, and Challenges in Physiological Emotion Recognition—A Tutorial Overview. *Proc. IEEE* (2023). doi:10.1109/JPROC.2023.3286445

[4] Astrid Carolus, Carolin Wienrich, Anna Törke, Tobias Friedel, Christian Schwietering, and Mareike Sperzel. 2021. 'Alexa, I Feel for You!' Observers' Empathetic Reactions Towards a Conversational Agent. *Frontiers in Computer Science* 3 (2021), 682982. doi:10.3389/fcomp.2021.682982

[5] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition* 44, 3 (2011), 572–587. doi:10.1016/j.patcog.2010.09.020

[6] Zhaohan Feng, Ruiqi Xue, Lei Yuan, Yang Yu, Ning Ding, Meiqin Liu, Bingzhao Gao, Jian Sun, Xinhu Zheng, and Gang Wang. 2025. Multi-agent embodied ai: Advances and future directions. *arXiv preprint arXiv:2505.05108* (2025). doi:10.48550/arXiv.2505.05108

[7] Agneta H Fischer and Antony S. R. Manstead. 2000. The Relation Between Gender and Emotions in Different Cultures. *Gender and Emotion: Social Psychological Perspectives* 1 (2000), 71–94. doi:10.1037/1528-3542.4.1.87

[8] Agneta H Fischer, Patricia M Rodriguez Mosquera, Annelies E. M. Van Vianen, and Antony S. R. Manstead. 2004. Gender and Culture Differences in Emotion. *Emotion* 4, 1 (2004), 87. doi:10.1037/1528-3542.4.1.87

[9] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355* (2025). doi:10.48550/arXiv.2506.22355

[10] Esther Görnemann and Sarah Spiekermann. 2024. Emotional responses to human values in technology: The case of conversational agents. *Human–Computer Interaction* 39, 5-6 (2024), 310–337. doi:10.1080/07370024.2022.2136094

[11] James J Gross. 2015. Emotion Regulation: Current Status and Future Prospects. *Psychological Inquiry* 26, 1 (2015), 1–26. doi:10.1080/1047840X.2014.940781

[12] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly* 37, 1 (2018), 81–88. doi:10.1080/02763869.2018.1404391

[13] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2022. The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses. *IEEE Transactions on Affective Computing* 14, 1 (2022), 17–30. doi:10.1109/TAFFC.2022.3205919

[14] Ziming Huang, Shulin Chen, and Hang Chen. 2024. Relationship Between Emotional Awareness and Self-Acceptance: The Mediating Role of Emotion Regulation Strategies. *Current Psychology* (2024), 1–9. doi:10.1007/s12144-024-05945-2

[15] Carroll E Izard. 2002. Translating emotion theory and research into preventive interventions. *Psychological bulletin* 128, 5 (2002), 796. doi:10.1037/0033-2909.128.5.796

[16] Philip Kossack and Herwig Unger. 2023. Emotion-Aware Chatbots: Understanding, Reacting, and Adapting to Human Emotions in Text Conversations. In *International Conference on Autonomous Systems*. Springer, 158–175. doi:10.1007/978-3-031-61418-7_8

[17] C. U. Om Kumar, N. Gowtham, Mohammed Zakariah, and Absulaziz Almazyad. 2024. Multimodal Emotion Recognition Using Feature Fusion: An LLM-Based Approach. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3425953

[18] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 10, 10 (2021), 1163. doi:10.3390/electronics10101163

[19] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 633–644. doi:10.1145/3322276.3322340

[20] Yong Ma, Heiko Drewes, and Andreas Butz. 2022. How Should Voice Assistants Deal With Users' Emotions? *arXiv preprint arXiv:2204.02212* (2022). doi:10.48550/arXiv.2204.02212

[21] Yong Ma, Oda Elise Nordberg, Yuchong Zhang, Arvid Rongve, Miroslav Bachinski, and Morten Fjeld. 2024. Understanding Dementia Speech: Towards an Adaptive Voice Assistant for Enhanced Communication. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 15–21. doi:10.1145/3660515.3661326

[22] Yong Ma, Yuchong Zhang, Miroslav Bachinski, and Morten Fjeld. 2023. Emotion-Aware Voice Assistants: Design, Implementation, and Preliminary Insights. In *Proceedings of the Eleventh International Symposium of Chinese CHI*. 527–532. doi:10.1145/3629606.3629665

[23] Yong Ma, Yuchong Zhang, Di Fu, Stephanie Zubicueta Portales, Danica Kragic, and Morten Fjeld. 2025. Advancing User-Voice Interaction: Exploring Emotion-Aware Voice Assistants Through a Role-Swapping Approach. In *International Conference on Human-Computer Interaction*. Springer, 303–320. doi:10.1007/978-3-031-92977-9_19

[24] Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen. 2024. Leveraging Speech PTM, Text LLM, and Emotional TTS for Speech Emotion Recognition. In *The 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11146–11150. doi:10.1109/ICASSP48485.2024.10445906

[25] Alex Mari, Andreina Mandelli, and René Algesheimer. 2024. Empathic Voice Assistants: Enhancing Consumer Responses in Voice Commerce. *Journal of Business Research* 175 (2024), 114566. doi:10.1016/j.jbusres.2024.114566

[26] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa… Examine the Variables Influencing the Use of Artificial Intelligent In-Home Voice Assistants. *Computers in Human Behavior* 99 (2019), 28–37. doi:10.1016/j.chb.2019.05.009

[27] Batja Mesquita and Nico H Frijda. 1992. Cultural Variations in Emotions: A Review. *Psychological Bulletin* 112, 2 (1992), 179. doi:10.1037/0033-2909.112.2.179

[28] Natawut Monaikul, Bahareh Abbasi, Zhanibek Rysbek, Barbara Di Eugenio, and Miloš Žefran. 2020. Role switching in task-oriented multimodal human-robot collaboration. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1150–1156. doi:10.1109/RO-MAN47096.2020.9223461

[29] Tarun Rathi and Manoj Tripathy. 2024. Analyzing the Influence of Different Speech Data Corpora and Speech Features on Speech Emotion Recognition: A Review. *Speech Communication* (2024), 103102. doi:10.1016/j.specom.2024.103102

[30] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. 2023. An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era. *Proc. IEEE* 111, 10 (2023), 1355–1381. doi:10.1109/JPROC.2023.3250266

[31] Jamil Zaki and W. Craig Williams. 2013. Interpersonal Emotion Regulation. *Emotion* 13, 5 (2013), 803. doi:10.1037/a0033839

[32] Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongsatorn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25. doi:10.1145/3706598.3713110

[33] Yuchong Zhang, Khaled Kassem, Zhengya Gong, Fan Mo, Yong Ma, Emma Kirjavainen, and Jonna Häkkilä. 2024. Human-centered AI technologies in human-robot interaction for social settings. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 501–505. doi:10.1145/3701571.3701610

[34] Yuchong Zhang, Yong Ma, Di Fu, Stephanie Zubicueta Portales, Morten Fjeld, and Danica Kragic. 2025. Personalizing Emotion-aware Conversational Agents? Exploring User Traits-driven Conversational Strategies for Enhanced Interaction. *arXiv preprint arXiv:2511.06954* (2025). doi:10.48550/arXiv.2511.06954

[35] Yuchong Zhang, Yong Ma, and Danica Kragic. 2024. Vision beyond boundaries: An initial design space of domain-specific large vision models in human-robot interaction. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*. 1–8. doi:10.1145/3640471.3680244

Received 2025-10-07; accepted 2025-12-09