Unrequited Emotions: Investigating the Gaps in Motivation and Practice in Speech Emotion Recognition Research

Taryn Wong[1], Zeerak Talat[2], Hanan Aldarmaki[3], Anjalie Field[1]
[1]Johns Hopkins University, [2]University of Edinburgh, [3]MBZUAI
z@zeerak.org, hanan.aldarmaki@mbzuai.ac.ae, anjalief@jhu.edu

Abstract

Critical analyses of emotion recognition technology have raised ethical concerns around task validity and potential downstream impacts, urging researchers to ensure alignment between their stated motivations and practice. However, these discussions have not adequately influenced or drawn from research on speech emotion recognition (SER). We address this gap by conducting a systematic survey of SER research to uncover what stated motivations drive this work and if they align with the datasets and emotions studied. We find that while SER research identifies appealing goals—such as well-situated voice-activated systems or healthcare applications—commonly-used datasets do not reflect these proposed deployment contexts, thus presenting a gap between motivations and research practices. We argue that such gaps engender ethical concerns, and that SER research should reassert itself with concrete use-cases to prevent misinterpretations, misuse, and downstream harms.

Keywords: emotion recognition, critical survey, ethics

1. Introduction

Emotion AI, including the detection of human emotions from video, image, audio, or text data using machine learning (ML) methods, has become a popular research area with increasing commercialization in a broad range of applications (McStay, 2018). The growing adoption of this technology has raised concerns around its development and use. Researchers have questioned the overall task validity, e.g., the premise of using someone's outer appearance to infer inner character or state, particularly given the lack of consensus on how to define emotions among psychologists (McStay and Pavliscak, 2019; Stark and Hoey, 2021). Empirical qualitative studies have demonstrated that people view emotion AI as invasive and privacy-violating (Andalibi and Buss, 2020; Pyle et al., 2024), and its use in contexts like workplaces or hiring processes create emotional labor with little perceived benefit for those subjected to it (Roemmich et al., 2023; Pyle et al., 2024). Importantly, concerns raised are often specific to emotion AI, rather than technology more generally (Pyle et al., 2024).

To date, this work has had limited engagement with the abundance of research on Speech Emotion Recognition (SER). Research on task validity (Jacobs and Wallach, 2021) and ethics have focused on facial recognition or broadly construed AI systems where underlying data and models are unspecified (Domnich and Anbarjafari, 2021; Pyle et al., 2024; Roemmich et al., 2023; Karizat et al., 2024). While some studies have questioned the ethics and validity of SER (Katirai, 2023; Hartmann et al., 2024; Morozov, 2014; Testa et al., 2023; Juslin et al., 2005), no prior work has conducted a systematic investigation of SER research practices.

In this work, we follow the suggestion proposed by Stark and Hoey (2021) that "[d]esigners and developers should think twice before embarking on emotion AI projects" and we investigate to what extent SER research meets their proposed "necessary though not sufficient condition" that projects have "clear alignment between conceptual models, data, norms, and aims". We formalize our study into three research questions: (1) What do SER researchers state as their motivations? (2) What purposes do popular datasets support? (3) Do datasets align with stated motivations?

RQ1 focuses on *why* researchers work on SER, while RQ2 focuses on *how*. RQ3 assesses the alignment between the *why* and *how*. We investigate these questions by conducting a systematic survey of 88 SER research papers, which we manually coded for stated motivations, specific emotions studied, and datasets used. We ultimately find a substantial gap between stated motivations and experimental setups, and we propose ways to reduce this gap by seeking suitable datasets or pursuing different motivations. Our work encourages future researchers to "think twice" about SER projects, and to seek alignment between objectives, research practices, and potential implications, rather than viewing SER as an isolated research task.

2. Methodology

Our work uses similar systemic survey methodology as previous studies that reflect on practices in speech processing research and related disciplines (Blodgett et al., 2020; Field et al., 2021; Birhane et al., 2022; Raff et al., 2023). First, we queried Se-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 57–73
16 May 2026.

mantic Scholar for papers containing search terms "speech" and "emotion" and {"recognition", "classification", "data", or "database"} in their title or abstract, collecting a total of 7,486 after filtering out papers without publication venue information. Of the initial sample, we kept only those published in popular archival speech, natural language processing, or ML venues, which we determined by manually reviewing venues with the most papers in the initial sample,[1] resulting in 959 papers.

We then randomly sampled 108 (stratified by publication year and venue) for in-depth manual analysis.

We used a standard inductive coding procedure to label the sampled 108 papers. First, three authors independently coded a set of 10 papers each (30 papers total), where each author first determined if the paper consisted of SER. Then, each author labeled the paper's stated motivations, specific emotions investigated, and data sets used in experiments or analysis.[2] In labeling stated motivations, annotators also quoted the exact text from the paper describing these motivations. The annotators then discussed the labeling scheme and finalized a set of frequently-stated motivations for which to code. Annotators then re-coded the initial 30 documents under the revised scheme, and then each independently coded an additional set of 10 documents to ensure mutual understanding of the coding scheme. The remainder of the data was then singly coded. Finally, the annotators reconvened and discussed further revisions to the scheme, specifically adding categories of motivations that were found in the larger data sample and re-coded the data under the revised scheme. Throughout this process, we removed 20 papers whose primary focus was not SER (e.g., papers that focused on depression detection, speech generation, speech representation, etc.), leaving a final set of 88 papers. The final set primarily consists of papers published in speech venues (Interspeech: 50 papers, ICASSP: 18, ASRU: 1, SLT: 2, Speech Communication: 4), as well as 9 papers from IEEE Transactions on Affective Computing. The final coding for all 88 papers is in Table 5 in the Appendix.

3. Results

[1] ICASSP; ASRU; SLT; Interspeech; NeurIPS; ICLR; ICML; AAAI; IEEE/ACM Transactions on Audio, Speech, and Language Processing; IEEE Open Journal of Signal Processing; IEEE Journal of Selected Topics in Signal Processing; Computer Speech & Language; Speech Communication; IEEE Transactions on Affective Computing; JMLR; ACL

[2] We also initially coded acknowledged funding sources, but found they were too sparse to analyze in aggregate.

Responsive bots: other HCI systems	42.05%
Responsive bots: voice assistants	12.50%
Responsive bots: car voice assistants	6.82%
Healthcare (mental health)	18.18%
Call screening	17.05%
Video games, toys, entertainment	13.64%
Education	9.09%
Paralinguistics / behavioral studies	6.82%
Social companion bots	4.55%
Lie detection	3.41%
Prior work	27.27%
Other	14.77%

Table 1: Percent of manually coded papers that reference each motivation. Motivations are grouped thematically. Percents do not add to 100 as papers often reference multiple motivations.

What do SER researchers state as their motivations? Table 1 presents the 12 stated motivations identified in our coding scheme with examples for each label in Appendix A. The most common motivation was enabling human-computer interaction systems to respond more naturally and effectively ("Responsive bots"), which we subdivide into "voice assistants", "car voice assistants", and "other HCI systems" (typically unspecified systems). Other common motivations focused on specific deployment contexts, including healthcare (almost always mental health), 911 or customer service calls, education, and entertainment. Less common motivations included social companion bots and studies of human behavior or paralinguistics. While most studies include motivations drawn from prior SER work, in data coding, we use the label "Prior work" to indicate that the paper drew motivation *exclusively* from prior SER research without without specifying how they expect this technology to be used. We also identified some papers with "Other" motivations. In practice, this label typically indicates that the paper includes a long list of applications, without depth of focus on any one. In Figure 1 we

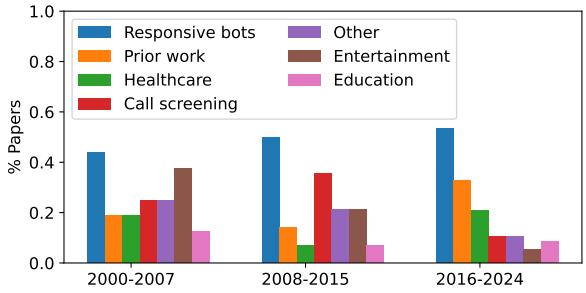

Figure 1: Percent of papers in each time window that reference each motivation. The number of papers in each time window are [16, 14, 58] respectively. We drop infrequent motivations for readability.

additionally report shifts over time. While "Respon-

Dataset	% Papers	Acted/Spont.	Emotions
IEMOCAP	40.91%	Acted (naturalistic)	Angry, Neutral, Sad, Happy, Fearful, Surprised, Frustrated, Excited, Valence/Sentiment, Arousal/Activation, Dominance
EMO-DB	17.05%	Acted	Angry, Neutral, Happy, Fearful, Disgusted, Anxious, Bored
RAVDESS	9.09%	Acted	Produced at both normal & strong intensities: Angry, Neutral,Sad, Happy, Fearful, Surprised, Disgusted, Calm
SUSAS	6.82%	Both	Stressed
MSP-Improv	6.82%	Acted (naturalistic)	Angry, Neutral, Sad, Happy
RECOLA	6.82%	Spontaneous	Valence/Sentiment, Arousal/Activation, Social behaviors rated as either positive or negative: Agreement, Dominance, Engagement, Performance, Rapport
Exclusive	35.23%		Paper exclusively used data in the top 6
Custom	22.73%		Paper custom-collected data
Other	45.45%		Paper used data not in the top 6

Table 2: Datasets used in our analysis corpus ≥ 5 times. We provide mode of collection and emotions labels for each dataset for reference. Categorical labels from Basic Emotion Theory are in violet. Axes from dimensional theories are in teal.

Angry	76.14%	Surprised	19.32%
Neutral	72.73%	Boredom	14.77%
Sad	67.05%	Joy	9.09%
Happy	65.91%	Stressed	7.95%
Fearful	30.68%	Calm	6.82%
Disgusted	27.27%	Dominance	5.68%
Valence/Sentiment	22.73%	Frustrated	3.41%
Arousal/Activation	20.45%	Excited	2.27%
Other/Unspecified	21.59%		

Table 3: Percent of papers that use each emotion label, ordered by frequency. Color-coding is as Table 2.

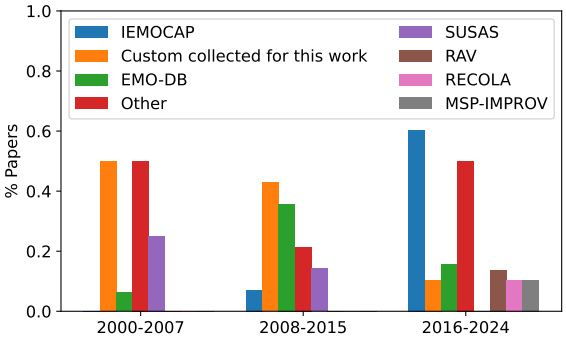

Figure 2: Percent of papers in each time window that use the specified dataset in their experiments.

sive bots" has consistently been a motivating application, "Call screening" and "Entertainment" have declined from 2000-2015 papers to 2016-2024 papers in our sample.

What purposes do popular datasets support? In Table 2, we identify "popular" datasets as ones used by at least five papers: IEMOCAP (Busso et al., 2008), EMO-DB (Burkhardt et al., 2005), RAVDESS (Livingstone and Russo, 2018), SUSAS (Hansen et al., 1997), MSP-Improv (Busso et al., 2017), and RECOLA (Ringeval et al., 2013). These six datasets support the majority of research in our corpus: 63.64% of papers use at least one of them, and 35.23% exclusively use them. Of the six popular datasets, four (IEMOCAP, EMO-DB, RAVDESS, MSP-Improv) consist of acted emotions and two (SUSAS, RECOLA) consist of emotions captured in spontaneous speech, which encourage vastly different use-cases. For example, SUSAS was created to study speech under different stressful situations, such as operating a helicopter or being subject to psychiatric analysis. In contrast, MSP-Improv was recorded in a laboratory environment by student actors. Furthermore, the datasets are often labeled or validated by annotators who are distinct from the speakers. While SER is often framed as identifying speakers' true emotions, as would be necessary to support use-cases like call screening or mental health support, a more accurate characterization of datasets annotated this way would be identifying how third parties perceive speakers' emotions.

Further, our analysis of the specific emotions studied in each paper reveals that papers use these datasets selectively (Table 3). Papers study angry, neutral, sad, and happy twice as often as other emotions, even though datasets support other labels, and there is no clear indication that these emotions would better suit the motivations in Table 1. IEMOCAP exemplifies this selective use. Almost all ($> 90\%$) of papers that use IEMOCAP use categorical labels, but far fewer (2-20%) use dimensional labels. Even within these categories usage is not consistent. While 19.44% of IEMOCAP papers use Valence/Sentiment labels, only 2.78% use Dominance.

Curiously, we find that dataset usage varies over time (Figure 2). One notable trend is the decline in data that was custom-collected for the specific paper, which are sometimes, but not always, more directly constructed to support the motivating task. Although 22.73% of the papers in our sample use custom-built datasets overall (Table 2), they are far more common in pre-2016 papers than 2016-2024 papers. Use of SUSAS has also declined, in contrast to IEMOCAP which accounts for almost 60% all data used in 2016-2024. The shifts in dataset usage are generally inconsistent with the shifts in motivations shown in Figure 1. For instance, there is no corresponding rise in interest for a use case

particularly suited to IEMOCAP that would justify its increased popularity in 2016-2024 as compared to 2008-2015 (IEMOCAP was released in 2008).

Do datasets align with stated motivations? In Figure 3, we display the mapping between the most common stated motivations and the six popular datasets, showing a notable lack of pattern. Despite the widely divergent downstream applications, at least one paper with every stated motivation used IEMOCAP. Similarly, although responsive bots or better human-computer interaction has been a persistent motivation for SER (Figure 1), none of the popular data sets were constructed for this application (Table 2), yet they continue to be used for it (Figure 3). Datasets with acted emotions are often used for applications intended for spontaneous speech, yet research has shown that spontaneous emotional expressions differ in various ways from acted expressions (Juslin et al., 2018). Furthermore, most datasets are collected from spontaneous or acted human-human interactions but are frequently used to advance human-computer interaction, even though people express themselves differently and more subtly when interacting with chatbots (Kovacevic et al., 2024), resulting in a significant domain mismatch.

Quantitative metrics of specific datasets and motivations yield similar mismatches. For example, 16 papers exclusively use IEMOCAP, but only 2 of them state Entertainment as a motivation, even though IEMOCAP's data collection protocol involving scripted and improvised acted speech is most similar to entertainment applications like movies and video games where speech is similarly acted. On the other hand, 16 papers listed Healthcare as a motivation, an application which requires uncovering a speakers true emotional state. While 3 of those papers exclusively use custom-collected data, and 2 papers evaluate over SUSAS (which was designed to capture simulated and actual stress), the remaining 11 papers all evaluate over almost entirely acted speech (e.g., 1 paper uses ASVP-ESD (Landry et al., 2020), which is more spontaneous).

4. Discussion

While these results identify a mismatch between stated motivations and underlying datasets, research and deployment are not necessarily expected to be identical. Strong popularity of a small number of datasets, such as frequent use of IEMOCAP in recent years, potentially reflects increasing standardization in evaluation setups, e.g., using established benchmark datasets facilitates publication by enabling direct comparison with prior work. Standardization of tasks has typically been constructive for accelerating progress. However, several characteristics of emotion prediction make the mismatch between stated motivations and empirical data problematic. We first discuss why this mismatch is potentially harmful and then offer avenues for reducing it.

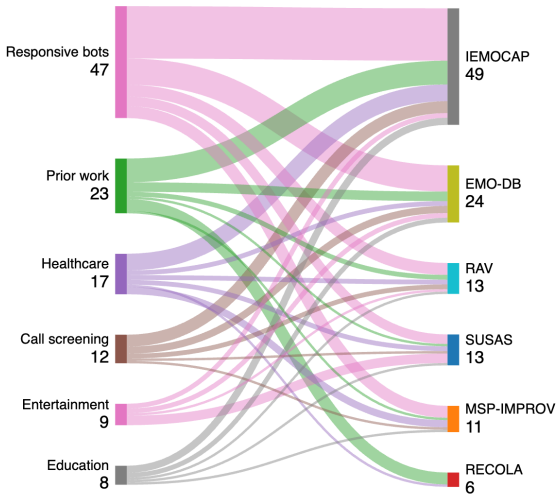

Figure 3: Mapping between most common stated motivations and datasets.

Does the mismatch between data and motivations cause harm? First, there is substantial interest in deploying emotion recognition technology in decision-making contexts where misclassification can have harmful consequences. In their analysis of patent applications for uses of emotion recognition technology in workplace settings, Boyd and Andalibi (2023) find that many patent applications treat outputs of their technology as ground-truth representations of internal states that can inform consequential decisions like hiring, firing, and calling law enforcement. SER systems have been shown to generalize poorly across datasets, even ones collected in similar ways (e.g., acted speech in laboratory settings) (Rí et al., 2023). There is also a fundamental difference between a person's true internal state and the emotions that an independent observer attributes to them, yet independent annotation to label or validate collected data is standard practice. The implication that a system designed and evaluated over a particular dataset could be useful in a mismatched downstream application can lead to harmful errors.

Additionally, people consider emotions private (McStay, 2020; Pyle et al., 2024), and even well-intentioned motivations are not necessarily well-received by people subject to this technology. In a survey of the attitudes of help seekers towards

predictive tools for mental health risks, only about half of the respondents (49%) were in favor of using risk prediction (Mantell et al., 2021), and similar concerns about the use of AI emotion in education exist (Stark and Hoey, 2021). Because of the sensitive nature of this topic, researchers should set high standards for SER research and ensure there are clear downstream benefits rather than treating SER as a generic task to optimize for and assuming future adoption will be done responsibly.

What data are needed to support these motivations? We next discuss how researchers could use task-appropriate datasets that reflect the intended downstream use cases, focusing for brevity on the most commonly referenced motivation: Responsive bots. Several studies collected datasets that are more aligned with this setup than the popular ones in Table 2. We identify a few examples out of the six papers that list responsive bots as a motivation and custom collect datasets. Wang and Tashev (2017) work with utterances from interactions with a spoken dialogue system, though emotion labels are separately crowd-sourced. Aubergé et al. (2003) propose a wizard-of-oz system for collecting this type of data, where a user believes they are interacting with a real computer system, but in reality a human is manipulating the system to create particular stimuli. Conducting research with real human-system interactions from deployed systems can introduce privacy concerns, but collecting similar data is feasible in simulated laboratory settings where users can give informed consent.

What use-cases can we support with this data? Alternatively, researchers could better align motivations and datasets by pursuing motivations that popular SER datasets do support. For example, because of its annotation setup, IEMOCAP offers a unique opportunity to investigate the alignment and discrepancies between the perception of emotions—i.e., the annotator's labels of emotions—and the intended portrayal of emotion—i.e., self-assessed emotions of the actors. Detection of emotions in acted speech could be used to investigate questions around emotional portrayals in acted media such as television shows (Zhou and Bamman, 2024), which has been of interest in gender studies (Weissmann, 2016) and media theory (Gorton, 2009; García, 2016) literature.

Conclusions Our analysis uncovers substantial gaps in how SER research is motivated and conducted. This mismatch exacerbates the risks associated with SER deployment, especially as the technology becomes increasingly commercialized. We recommend ways to reduce this gap, by shifting datasets to match motivations or motivations to match datasets. Given the sensitivity of emotion recognition technology, we urge researchers to critically evaluate the motivations behind SER projects, ensuring they are well-justified and that experiments meaningfully support them.

5. Ethical Considerations and Limitations

The primary limitation of our work is its reliance on a specific data sample. Although we carefully chose a range of search terms to identify SER papers and we stratify our data sample by year and publication venue, it is possible that analysis of a broader set of papers could yield different findings.

6. Bibliographical References

Noam Amir, Ori Kerret, and Dimitry Karlinski. 2001. Classifying emotions in speech: a comparison of methods. In *Interspeech*.

Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proc. of CHI*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Véronique Aubergé, Nicolas Audibert, and Albert Rilliard. 2003. Why and how to control the authentic emotional speech corpora. In *Interspeech*.

Lokesh Kumar Bansal, S. Pavankumar Dubagunta, Malolan Chetlur, Pushpak Jagtap, and Aravind Ganapathiraju. 2023. On the efficacy and noise-robustness of jointly learned speech emotion and automatic speech recognition. In *Interspeech*.

Fang Bao, Michael Neumann, and Ngoc Thang Vu. 2019. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *Interspeech*.

Dario Bertero and Pascale Fung. 2017. A first look into a convolutional neural network for speech emotion detection. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proc. of FAccT*, pages 173–184.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proc. of ACL*, pages 5454–5476.

Karen L Boyd and Nazanin Andalibi. 2023. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. *Proc. of CSCW*, 7(CSCW1):1–37.

F Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, and B Weiss. 2005. A database of german emotional speech. In *Proc. of INTERSPEECH*, pages 805–808.

Felix Burkhardt, Jitendra Ajmera, Roman Englert, Joachim Stegmann, and Winslow Burleson. 2006. Detecting anger in automated voice portal dialogs. In *Interspeech*.

C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Samuel Cahyawijaya, Holy Lovenia, Willy Chung, Rita Frieske, Zihan Liu, and Pascale Fung. 2023. Cross-lingual cross-age adaptation for low-resource elderly speech emotion recognition. In *Interspeech*.

CasaleSalvatore, RussoAlessandra, and SerranoSalvatore. 2007. Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*.

Ming Chen and Xudong Zhao. 2020. A multi-scale fusion framework for bimodal speech emotion recognition. In *Interspeech*.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023. DST: Deformable speech transformer for emotion recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Woan-Shiuan Chien and Chi-Chun Lee. 2023. Achieving fair speech emotion recognition via perceptual fairness. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Danilo de Oliveira, Navin Raj Prabhu, and Timo Gerkmann. 2023. Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models. In *Interspeech*.

Artem Domnich and Gholamreza Anbarjafari. 2021. Responsible AI: Gender bias assessment in emotion recognition. *ArXiv*, abs/2103.11436.

Tiantian Feng and Shrikanth S. Narayanan. 2022. Semi-fedser: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling. In *Interspeech*.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proc. of ACL*, pages 1905–1925.

Yuan Gao, Chenhui Chu, and Tatsuya Kawahara. 2023. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining. In *Interspeech*.

Alberto N. García, editor. 2016. *Emotions in Contemporary TV Series*. Palgrave Macmillan UK, London.

Jakub Gałka, Joanna Grzybowska, Magdalena Igras, Pawel Jaciów, Kamil Wajda, Marcin Witkowski, and Mariusz Ziólko. 2015. System supporting speaker identification in emergency call center. In *Interspeech*.

Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2009. A dimensional approach to emotion recognition of speech from movies. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68.

Lucas Goncalves and Carlos Busso. 2022. Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks. In *Interspeech*.

Erik Goron, Lena Asai, Elias Rut, and Martin Dinov. 2024. Improving domain generalization in speech emotion recognition with whisper. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11631–11635.

Kristyn Gorton. 2009. *Media Audiences: Television, Meaning and Emotion*. Edinburgh University Press.

Michael Grimm, Kristian Kroschel, and Shrikanth S. Narayanan. 2007. Support vector regression for automatic recognition of spontaneous emotions in speech. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4:IV–1085–IV–1088.

Purnima Gupta and Nitendra Rajput. 2007. Two-stream emotion recognition for call center monitoring. In *Interspeech*.

Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, and Björn Schuller. 2018. Towards conditional adversarial training for predicting emotions from speech. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6822–6826.

Wenjing Han, Haifeng Li, Huabin Ruan, Lin Ma, Jiayin Sun, and Björn Schuller. 2013. Active learning for dimensional speech emotion recognition. In *Interspeech*.

John HL Hansen, Sahar E Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. 1997. Getting started with SUSAS: a speech under simulated and actual stress database. In *Eurospeech*, pages 1743–46.

Kris Vera Hartmann, Giovanni Rubeis, and Nadia Primc. 2024. Healthy and happy? an ethical investigation of emotion recognition and regulation technologies (err) within ambient assisted living (aal). *Science and Engineering Ethics*, 30(1):2.

Mixiao Hou, Zheng Zhang, Qi Cao, David Zhang, and Guangming Lu. 2022. Multi-view speech emotion recognition via collective relation construction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:218–229.

Desheng Hu, Xinhui Hu, and Xinkang Xu. 2022. Multiple enhancements to lstm for learning emotion-salient features in speech emotion recognition. In *Interspeech*.

Jian Huang, Ya Li, Jianhua Tao, and Zheng Lian. 2018. Speech emotion recognition from variable-length inputs with triplet loss function. In *Interspeech*.

Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su, and Yu-Ting Kuo. 2020. Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. *IEEE Transactions on Affective Computing*, 11:393–404.

Yu-Lin Huang, Bo-Hao Su, Y.-W. Peter Hong, and Chi-Chun Lee. 2021. An attribute-aligned strategy for learning speech representation. *ArXiv*, abs/2106.02810.

Richard Huber, Anton Batliner, Jan Buckow, Elmar Nöth, Volker Warnke, and Heinrich Niemann. 2000. Recognition of emotion in a realistic dialogue scenario. In *Interspeech*.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada. ACM.

Christian Martyn Jones and Andrew Deeming. 2008. Speech interaction with an emotional robotic dog. In *Interspeech*.

Patrik N Juslin, Petri Laukka, and Tanja Bänziger. 2018. The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, 42:1–40.

Patrik N Juslin, Klaus R Scherer, J Harrigan, and R Rosenthal. 2005. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135.

Zuheng Kang, Junqing Peng, Jianzong Wang, and Jing Xiao. 2022. SpeechEQ: Speech emotion recognition based on multi-scale unified datasets and multitask learning. In *Interspeech*.

Nadia Karizat, Alexandra H Vinson, Shobita Parthasarathy, and Nazanin Andalibi. 2024. Patent applications as glimpses into the sociotechnical imaginary: Ethical speculation on the imagined futures of emotion ai for mental health monitoring and detection. *Proc. of CSCW*, 8(CSCW1):1–43.

Amelia Katirai. 2023. Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics*, pages 1–22.

Jonghwa Kim, Elisabeth André, Matthias Rehm, Thurid Vogt, and Johannes Wagner. 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Interspeech*.

Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. On multimodal emotion recognition for human-chatbot interaction in the wild. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 12–21.

O. W. Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Interspeech*.

Dejoli Landry, Qianhua He, Haikang Yan, and Yanxiong Li. 2020. Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*, 8:1793–1798.

Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, and Philip C. Woodland. 2024. Parameter efficient finetuning for speech emotion recognition and domain adaptation. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10986–10990.

Siddique Latif, Rajib Kumar Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn Schuller. 2020a. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*, 13:992–1004.

Siddique Latif, Rajib Kumar Rana, Sara Khalifa, Raja Jurdak, and Björn Schuller. 2020b. Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition. In *Interspeech*.

Baojie Li, Keikichi Hirose, and Nobuaki Minematsu. 2002. Robust speech recognition using inter-speaker and intra-speaker adaptation. In *Interspeech*.

Chia-Yu Li, Daniel Ortega, Dirk Vath, Florian Lux, Lindsey Vanderlyn, Maximilian Schmidt, Michael Neumann, Moritz Volkel, Pavel Denisov, Sabrina Jenne, Zorica Kacarevic, and Ngoc Thang Vu. 2020. Adviser: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents. In *Annual Meeting of the Association for Computational Linguistics*.

Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, and John D. Newman. 2007. Stress and emotion classification using jitter and shimmer features. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4:IV–1081–IV–1084.

Xingfeng Li, Xiaohan Shi, Desheng Hu, Yongwei Li, Qingcheng Zhang, Zhengxia Wang, Masashi Unoki, and Masato Akagi. 2023a. Music theory-inspired acoustic representation for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2534–2547.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2022. Exploration of a self-supervised speech model: A study on emotional corpora. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 868–875.

Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. 2023b. Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. In *Interspeech*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ignacio López-Moreno, Carlos Ortego-Resa, Joaquín González-Rodríguez, and Daniel Ramos. 2009. Speaker dependent emotion recognition using prosodic supervectors. In *Interspeech*.

Veronika Makarova and Valery A. Petrushin. 2002. Ruslana: a database of russian emotional utterances. In *Interspeech*.

Pauline Katharina Mantell, Annika Baumeister, Stephan Ruhrmann, Anna Janhsen, and Christiane Woopen. 2021. Attitudes towards risk prediction in a help seeking population of early detection centers for mental disorders—a qualitative approach. *International journal of environmental research and public health*, 18(3):1036.

Shuiyang Mao, Pak-Chung Ching, and Tan Lee. 2019. Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition. In *Interspeech*.

Sho Matsumiya, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Data-driven generation of text balloons based on linguistic and acoustic features of a comics-anime corpus. In *Interspeech*.

A. McStay. 2018. *Emotional AI: The Rise of Empathic Media*. SAGE Publications Limited.

Andrew McStay. 2020. Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society*, 7(1):2053951720904386.

Andrew McStay and Pamela Pavliscak. 2019. Emotional artificial intelligence: Guidelines for ethical use. *COMEST/UNESCO*.

Hanyu Meng, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2023. What is learnt by the learnable front-end (leaf)? adapting per-channel energy normalisation (pcen) to noisy conditions. In *Interspeech*.

Patrick Meyer and Tim Fingscheidt. 2014. A new evaluation methodology for speech emotion recognition with confidence output. In *ITG Symposium on Speech Communication*.

Rosanna Milner, Md. Asif Jalal, Raymond W. M. Ng, and Thomas Hain. 2019. A cross-corpus study on speech emotion recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 304–311.

Vikramjit Mitra, Jingping Nie, and Erdrin Azemi. 2023. Investigating salient representations and label variance in dimensional speech emotion

analysis. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11111–11115.

EVGENY Morozov. 2014. To save everything, click here: the folly of technological solutionism. *J. Inf. Policy*, 4(2014):173–175.

Anish Nediyanchath, Periyasamy Paramasivam, and Promod Yenigalla. 2020. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7179–7183.

Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. Speech emotion recognition using hidden markov models. In *Interspeech*.

Christopher Oates, Andreas Triantafyllopoulos, Ingmar Steiner, and Björn Schuller. 2019. Robust speech emotion recognition under different encoding conditions. In *Interspeech*.

Georgios Paraskevopoulos, Efthymios Tzinis, Nikolaos Ellinas, Theodoros Giannakopoulos, and Alexandros Potamianos. 2019. Unsupervised low-rank representations for speech emotion recognition. In *Interspeech*.

Zi Jun Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu. 2021. Efficient speech emotion recognition using multi-scale cnn and attention. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3020–3024.

Hai X. Pham, Yuting Wang, and Vladimir Pavlovic. 2022. Learning continuous facial actions from speech for real-time animation. *IEEE Transactions on Affective Computing*, 13:1567–1580.

Navin Raj Prabhu, Nale Lehmann-Willenbrock, and Timo Gerkmann. 2022. End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning. *IEEE Transactions on Affective Computing*, 15:579–592.

Emily Mower Provost, Maja J. Matarić, and Shrikanth S. Narayanan. 2009. Evaluating evaluators: a case study in understanding the benefits and pitfalls of multi-evaluator modeling. In *Interspeech*.

Cassidy Pyle, Kat Roemmich, and Nazanin Andalibi. 2024. Us job-seekers' organizational justice perceptions of emotion ai-enabled interviews. *Proc. of CSCW*, 8(CSCW2):1–42.

Edward Raff, Michel Benaroch, Sagar Samtani, and Andrew L Farris. 2023. What do machine learning researchers mean by" reproducible"? In *Proceedings of INTERSPEECH*.

Mandar Rahurkar and John H. L. Hansen. 2003. Frequency distribution based weighted sub-band approach for classification of emotional/stressful content in speech. In *Interspeech*.

Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. 2018. State of mind: Classification through self-reported affect and word use in speech. In *Interspeech*.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion ai at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In *Proc. of CHI*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Francesco Ardan Dal Rí, Fabio Cifariello Ciardi, and Nicola Conci. 2023. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. *IEEE Access*, 11:116638–116649.

Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Y. Espy-Wilson. 2019. Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription. In *Interspeech*.

Michelle Hewlett Sanchez, Gökhan Tür, Luciana Ferrer, and Dilek Z. Hakkani-Tür. 2010. Domain adaptation and compensation for emotion detection. In *Interspeech*.

Izhak Shafran and Mehryar Mohri. 2005. A comparison of classifiers for detecting emotion from speech. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 1:I/341–I/344 Vol. 1.

Abhinav Shukla, Stavros Petridis, and Maja Pantic. 2020. Does visual self-supervision improve learning of speech representations for emotion recognition? *IEEE Transactions on Affective Computing*, 14:406–420.

Julia Sidorova, Simon Karlsson, Oliver Rosander, Marcelo L. Berthier, and Ignacio Moreno-Torres.

2021. Towards disorder-independent automatic assessment of emotional competence in neurological patients with a classical emotion recognition system: Application in foreign accent syndrome. *IEEE Transactions on Affective Computing*, 12:962–973.

Peng Song and Wenming Zheng. 2020. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 11:373–382.

Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proc. of FAccT*, FAccT '21, page 782–793, New York, NY, USA. Association for Computing Machinery.

Bo-Hao Su and Chi-Chun Lee. 2021. A conditional cycle emotion gan for cross corpus speech emotion recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 351–357.

Rui Sun and Elliot Moore II. 2012. A preliminary study on cross-databases emotion recognition using the glottal features in speech. In *INTERSPEECH*, pages 1628–1631.

Brian Testa, Yi Xiao, Harshit Sharma, Avery Gump, and Asif Salekin. 2023. Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–30.

Andreas Triantafyllopoulos, Gil Keren, Johannes Wagner, Ingmar Steiner, and Björn Schuller. 2019. Towards robust speech emotion recognition using deep residual networks for speech enhancement. In *Interspeech*.

Panagiotis Tzirakis, Anh-Tuan Nguyen, Stefanos Zafeiriou, and Björn W. Schuller. 2021. Speech emotion recognition using semantic information. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283.

Bogdan Vlasenko, Björn Schuller, Kinfe Tadesse Mengistu, Gerhard Rigoll, and Andreas Wendemuth. 2008. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In *Interspeech*.

Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. 2007. Combining frame and turn-level information for robust recognition of emotions within speech. In *Interspeech*.

Bogdan Vlasenko and Andreas Wendemuth. 2009. Processing affected speech within human machine interaction. In *Interspeech*.

Hongxuan Wang and Prahlad Vadakkepat. 2024. Gradient-based dimensionality reduction for speech emotion recognition using deep networks. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11496–11500.

Zhong-Qiu Wang and Ivan Tashev. 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5154.

Elke Weissmann. 2016. Women, Television and Feelings: Theorising Emotional Difference of Gender in SouthLAnd and Mad Men. In Alberto N. García, editor, *Emotions in Contemporary TV Series*, pages 87–101. Palgrave Macmillan UK, London.

Chung-Hsien Wu and Wei-Bin Liang. 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2:10–21.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2021. A novel markovian framework for integrating absolute and relative ordinal emotion information. *IEEE Transactions on Affective Computing*, 14:2089–2101.

Wen Wu, C. Zhang, and Philip C. Woodland. 2023. Integrating emotion recognition with speech recognition and speaker diarisation for conversations. *ArXiv*, abs/2308.07145.

Jie Xie, Mingying Zhu, and Kai Hu. 2023. Fusion-based speech emotion classification using two-stage feature selection. *Speech Commun.*, 152:102955.

Zixiaofan Yang and Julia Hirschberg. 2018. Predicting arousal and valence from waveforms and spectrograms using deep neural networks. In *Interspeech*.

Jiaxin Ye, Xin-Cheng Wen, X. Wang, Yan Luo, Chang-Li Wu, Liyan Chen, and Kunhong Liu. 2022. Gm-tcnet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition. *ArXiv*, abs/2210.15834.

Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. 2018. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826.

Sungrack Yun and Chang Dong Yoo. 2009. Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4169–4172.

Noor Aina Zaidan and Sah Hj Salam. 2016. Mfcc global features selection in improving speech emotion recognition rate. In *International Conference on Machine Learning*.

Ruoxi Zhang, Atsushi Ando, Satoshi Kobashikawa, and Yushi Aono. 2017. Interaction and transition model for speech emotion recognition in dialogue. In *Interspeech*.

Xiaoheng Zhang and Y. Li. 2023. A dual attention-based modality-collaborative fusion network for emotion recognition. In *Interspeech*.

Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2019. Attention-enhanced connectionist temporal classification for discrete speech emotion recognition. In *Interspeech*.

Ziping Zhao, Yu Zheng, Zixing Zhang, Haishuai Wang, Yiqin Zhao, and Chao Li. 2018. Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition. In *Interspeech*.

Naitian Zhou and David Bamman. 2024. Once more, with feeling: Measuring emotion of acting performances in contemporary american film. In *Computational Humanities Research (CHR)*.

A. Examples of coded motivations

Responsive bots: other HCI systems	“These models could lead to computer agents and robots that more naturally and functionally blend into human society” (Provost et al., 2009)
Responsive bots: voice assistants; Responsive bots: other HCI systems	“Speech emotion recognition is becoming more and more important for many applications related to human computer interactions, especially for spoken dialogue systems. With emotion recognition from users’ speech, a better user experience can be achieved” (Wang and Tashev, 2017)
Healthcare (mental health); Responsive bots: car voice assistants; Education; Other	“In affective computing field, emotion recognition from speech plays a very important role, and has received much attention over the past few decades... It has been proven useful in many applications which require human-machine interaction, e.g., in-car board system, diagnostic tool for therapists, automatic translation systems, computer tutorial applications” (Song and Zheng, 2020)
Call screening	"In a call-center, such a system should be able to determine in a critical phase of the dialogue if the call should be passed over to a human operator." (Huber et al., 2000)
Paralinguistics / behavioral studies	“Speech signals carry rich information on an individual’s emotional states, expressed through both paralinguistic and semantic cues.” (de Oliveira et al., 2023)
Social companion bots; Video games, toys, entertainment; Education; Other	“Research on emotion recognition can advance many applications, like distance education, social robots, video games, affective mirrors and many others.” (Li et al., 2023a)
Lie detection; Call screening; Healthcare (mental health	“Especially in the field of human-machine interaction (HCI), growing interest can be observed in recent years. In addition, the detection of lies, monitoring of call centers and medical diagnoses are often claimed as promising application scenarios for speech emotion recognition.” (Mao et al., 2019)
Prior work	“While the majority of traditional research in emotional speech recognition has utilized a single database for analysis, it is becoming clear that the lack of sufficiently large databases with varied emotion types is a significant hindrance to the design of a robust emotion classification system.” (Sun and Moore II, 2012)

Table 4: Examples of human-selected text snippets and coded motivations, with at least one example per motivation. We remove citations from within quotations for readability.

Table 5: All 88 manually coded papers

	Venue	Motivations	Datasets	Emotions
Huber et al. (2000)	Interspeech	Call screening	Custom	Angry, Other or Unspecified
Amir et al. (2001)	Interspeech	Prior work	Universidad Politecnica de Madrid	Angry, Sad, Happy, Neutral
Nogueiras et al. (2001)	Interspeech	Healthcare, Entertainment	Interface Database	Surprised, Joy, Angry, Fearful, Disgusted, Sad, Neutral
Li et al. (2002)	Interspeech	Prior work	Custom	Angry, Other or Unspecified, Sad, Neutral
Makarova and Petrushin (2002)	Interspeech	Entertainment, Other	RUSLANA	Neutral, Surprised, Happy, Angry, Fearful, Sad
Rahurkar and Hansen (2003)	Interspeech	Paralingusitic / behavioral studies, RB: voice assistants, Entertainment, RB: other HCI systems	SOQ	Neutral, Stressed
Aubergé et al. (2003)	Interspeech	RB: other HCI systems, RB: voice assistants	Custom	Other or Unspecified
Kwon et al. (2003)	Interspeech	Entertainment, RB: other HCI systems	SUSAS, FAU Aibo Emotion Corpus	Stressed, Angry, Boredom, Happy, Neutral, Sad
Kim et al. (2005)	Interspeech	RB: other HCI systems	Custom	Arousal/Activation, Valence or sentiment
Shafran and Mohri (2005)	ICASSP	RB: other HCI systems, Other, Call screening, Education	Custom	Valence or sentiment
Burkhardt et al. (2006)	Interspeech	Call screening	Custom	Angry
Gupta and Rajput (2007)	Interspeech	Call screening	EPST, Custom	Angry, Happy, Neutral
CasaleSalvatore et al. (2007)	Speech Communication	RB: other HCI systems, Healthcare, Other, Entertainment, Education	SUSAS	Angry, Stressed, Neutral
Li et al. (2007)	ICASSP	Healthcare, Other, Lie detection, Entertainment	SUSAS, African Elephant Emotional Arousal Dataset, Custom	Other or Unspecified, Angry, Arousal/Activation,Neutral
Vlasenko et al. (2007)	Interspeech	Prior work	EMO-DB, SUSAS	Happy, Angry, Fearful, Boredom, Disgusted, Stressed, Neutral
Grimm et al. (2007)	ICASSP	RB: other HCI systems	VAM	Valence or sentiment, Dominance, Other or Unspecified
Jones and Deeming (2008)	Interspeech	Call screening, Other, RB: car voice assistants, Entertainment, Lie detection	Custom	Happy, Boredom, Sad, Surprised, Angry, Other or Unspecified, Excited, Fearful, Frustrated
Vlasenko et al. (2008)	Interspeech	RB: other HCI systems	EMO-DB, SUSAS	Angry, Boredom, Disgusted, Fearful, Joy, Neutral, Sad, Stressed
Vlasenko and Wendemuth (2009)	Interspeech	RB: other HCI systems	Kiel Corpus of Read Speech , EMO-DB, FAU Aibo Emotion Corpus	Angry, Other or Unspecified, Neutral, Joy

Continued on next page

	Venue	Motivations	Datasets	Emotions
Provost et al. (2009)	Interspeech	RB: other HCI systems	IEMOCAP	Angry, Happy, Sad, Neutral, Valence or sentiment, Arousal/Activation
Yun and Yoo (2009)	ICASSP	Call screening, Entertainment, RB: other HCI systems	EMO-DB	Angry, Fearful, Disgusted, Sad, Boredom, Neutral, Happy
López-Moreno et al. (2009)	Interspeech	RB: car voice assistants, Entertainment, Call screening	SUSAS	Neutral, Stressed
Giannakopoulos et al. (2009)	ICASSP	Other	Custom	Valence or sentiment, Arousal
Sanchez et al. (2010)	Interspeech	Call screening	Custom	Fearful, Sad, Neutral
Wu and Liang (2011)	IEEE Transactions on Affective Computing	Healthcare, Education	Custom	Angry, Happy, Neutral, Sad
Sun and Moore II (2012)	Interspeech	Prior work	EMO-DB, EPST, EMA	Neutral, Angry, Sad, Happy
Han et al. (2013)	Interspeech	Prior work	SEMAINE	Arousal/Activation, Dominance, Valence or sentiment, Other or Unspecified
Matsumiya et al. (2014)	Interspeech	Other	Custom	Angry, Sad, Happy, Fearful, Neutral
Meyer and Fingscheidt (2014)	Speech Communication	RB: other HCI systems	EMO-DB	Fearful, Disgusted, Happy, Boredom, Neutral, Sad, Angry
Gałka et al. (2015)	Interspeech	Call screening	Custom	Angry, Stressed, Neutral
Zaidan and Salam (2016)	Advances in Machine Learning and Signal Processing	RB: other HCI systems	EMO-DB	Happy, Sad, Fearful, Angry, Disgusted, Boredom, Neutral
Zhang et al. (2017)	Interspeech	RB: other HCI systems, Other, Entertainment, Education	IEMOCAP	Angry, Happy, Sad, Neutral, Other or Unspecified
Wang and Tashev (2017)	ICASSP	RB: other HCI systems, RB: voice assistants	Custom	Neutral, Happy, Sad, Angry
Bertero and Fung (2017)	ICASSP	Prior work	Custom	Angry, Happy, Sad
Rathner et al. (2018)	Interspeech	Healthcare	Custom	Arousal/Activation, Valence or sentiment
Yang and Hirschberg (2018)	Interspeech	Prior work	RECOLA, SEMAINE	Valence or sentiment, Arousal
Han et al. (2018)	ICASSP	Prior work	RECOLA	Arousal/Activation, Valence or sentiment
Zhao et al. (2018)	Interspeech	Prior work	CHEAVD, IEMOCAP	Angry, Disgusted, Happy, Sad, Surprised, Neutral, Other or Unspecified
Huang et al. (2018)	Interspeech	Paralingusitic / behavioral studies	IEMOCAP	Angry, Happy: Happy + Excited, Neutral, Sad
Yenigalla et al. (2018)	Interspeech	Prior work	IEMOCAP	Neutral, Happy, Sad, Angry

Continued on next page

	Venue	Motivations	Datasets	Emotions
Milner et al. (2019)	ASRU	Prior work	eNTERFACE, RAVDESS (RAV), IEMOCAP, MOSEI	Angry, Happy, Sad, Surprised, Disgusted, Fearful, Neutral, Frustrated, Excited, Calm, Other or Unspecified
Bao et al. (2019)	Interspeech	RB: other HCI systems	IEMOCAP, MSP-IMPROV, TEDLIUM (release 2)	Angry, Happy: Happy + Excited, Sad, Neutral
Mao et al. (2019)	Interspeech	Healthcare, Call screening, Lie detection	IEMOCAP, CASIA	Sad, Neutral, Angry, Fearful, Happy, Surprised
Oates et al. (2019)	Interspeech	Prior work	EMO-DB, RECOLA, eNTERFACE, Polish Emotional Speech Database	Angry, Boredom, Fearful, Joy, Neutral, Sad, Disgusted, Happy, Surprised, Arousal/Activation, Valence or sentiment
Latif et al. (2020a)	IEEE Transactions on Affective Computing	Healthcare, Call screening, RB: other HCI systems	IEMOCAP, MSP-IMPROV	Angry, Happy, Neutral, Sad
Yoon et al. (2019)	ICASSP	RB: other HCI systems, Paralingusitic / behavioral studies	IEMOCAP	Happy: Happy + Excited, Angry, Neutral, Sad
Zhao et al. (2019)	Interspeech	Prior work	IEMOCAP, FAU Aibo Emotion Corpus	Happy: Happy + Excited, Angry, Sad, Neutral, Emphatic, Positive, Rest
Triantafyllopoulos et al. (2019)	Interspeech	Prior work	RECOLA, EMO-DB, eNTERFACE, Mozilla Common Voice Audio Set	Arousal/Activation, Other or Unspecified
Sahu et al. (2019)	Interspeech	Healthcare, Paralingusitic / behavioral studies, RB: voice assistants	IEMOCAP, MSP-IMPROV	Angry, Happy, Neutral, Sad
Paraskevopoulos et al. (2019)	Interspeech	RB: other HCI systems	EMO-DB, IEMOCAP	Angry, Disgusted, Fearful, Joy, Boredom, Sad, Neutral, Happy
Huang et al. (2020)	IEEE Transactions on Affective Computing	Healthcare	Custom	Happy, Fearful, Angry, Surprised, Sad, Disgusted
Song and Zheng (2020)	IEEE Transactions on Affective Computing	Healthcare, Other, RB: voice assistants, Education	EMO-DB, eNTERFACE, FAU Aibo Emotion Corpus	Angry, Boredom, Disgusted, Fearful, Happy, Sad, Surprised
Shukla et al. (2020)	IEEE Transactions on Affective Computing	Healthcare, Other	CREMA-D, RECOLA, RAVDESS (RAV), SEWA, IEMOCAP	Angry, Fearful, Disgusted, Happy, Sad, Neutral, Calm, Surprised, Arousal/Activation,Valence or sentiment
Li et al. (2020)	ACL Demos	RB: voice assistants, Social Companion bots	IEMOCAP	Angry, Happy, Sad, Neutral, Arousal/Activation, Valence or sentiment
Latif et al. (2020b)	Interspeech	Prior work	IEMOCAP, MSP-IMPROV, DEMAND	Angry, Neutral, Sad, Happy: Happy + Excited
Chen and Zhao (2020)	Interspeech	RB: other HCI systems	IEMOCAP	Happy: Happy + Excited, Neutral, Angry, Sad

Continued on next page

	Venue	Motivations	Datasets	Emotions
Nediyanchath et al. (2020)	ICASSP	RB: voice assistants, Social Companion bots, RB: other HCI systems	IEMOCAP	Sad, Angry, Neutral, Happy
Sidorova et al. (2021)	IEEE Transactions on Affective Computing	Healthcare	Custom, Interface Database	Angry, Sad, Happy, Neutral
Huang et al. (2021)	Interspeech	Prior work	MSP-Podcasts	Neutral, Angry, Sad, Happy, Disgusted
Wu et al. (2021)	IEEE Transactions on Affective Computing	Prior work	RECOLA, IEMOCAP	Arousal/Activation, Valence or sentiment
Tzirakis et al. (2021)	ICASSP	RB: other HCI systems	SEWA	Arousal/Activation,Other or Unspecified, Valence or sentiment
Su and Lee (2021)	SLT	RB: other HCI systems, RB: car voice assistants	IEMOCAP, MSP-IMPROV, CIT: USC CreativeIT Corpus	Arousal/Activation, Valence or sentiment
Peng et al. (2021)	ICASSP	Call screening, Healthcare, RB: other HCI systems	IEMOCAP	Angry, Sad, Neutral, Happy: Happy + Excited
Li et al. (2022)	SLT	Prior work	IEMOCAP, RAVDESS (RAV)	Angry, Happy: Happy + Excited, Neutral, Happy, Calm, Sad, Fearful, Surprised, Disgusted
Pham et al. (2022)	IEEE Transactions on Affective Computing	RB: other HCI systems	RAVDESS (RAV), SAVEE, VidTIMIT, GRID, GEMEP	Neutral, Calm, Happy, Sad, Angry, Fearful, Surprised, Disgusted
Prabhu et al. (2022)	IEEE Transactions on Affective Computing	Prior work	AVEC'16, MSP-Conversation	Valence or sentiment, Arousal
Hu et al. (2022)	Interspeech	RB: other HCI systems	IEMOCAP	Neutral, Angry, Happy, Sad
Kang et al. (2022)	Interspeech	RB: other HCI systems	Custom	Other or Unspecified
Feng and Narayanan (2022)	Interspeech	Other, Healthcare, RB: voice assistants, Education	IEMOCAP, MSP-IMPROV	Happy, Neutral, Angry, Other or Unspecified
Ye et al. (2022)	Speech Communication	RB: other HCI systems	CASIA, EMO-DB, RAVDESS (RAV), SAVEE	Angry, Boredom, Calm, Disgusted, Fearful, Happy, Neutral, Sad, Surprised
Hou et al. (2022)	IEEE/ACM Transactions on Audio, Speech, and Language Processing	RB: other HCI systems	IEMOCAP, EMO-DB	Angry, Sad, Neutral, Boredom, Fearful, Disgusted, Happy: Happy + Excited
Goncalves and Busso (2022)	Interspeech	Prior work	CREMA-D, MSP-Face	Happy, Fearful, Disgusted, Angry, Sad, Neutral

Continued on next page

	Venue	Motivations	Datasets	Emotions
Li et al. (2023a)	IEEE/ACM Transactions on Audio, Speech, and Language Processing	Other, Social Companion bots, Entertainment, Education	IEMOCAP	Neutral, Happy: Happy + Excited, Sad, Angry, Valence or sentiment, Arousal/Activation,Dominance
Chien and Lee (2023)	ICASSP	RB: other HCI systems	IEMOCAP	Neutral, Happy, Angry, Frustrated, Sad
Meng et al. (2023)	Interspeech	Prior work	CREMA-D	Other or Unspecified
Cahyawijaya et al. (2023)	Interspeech	Prior work	CREMA-D, ElderReact, ESD: Emotional Speech Database, TESS, YueMotion, CSED, IEMOCAP	Happy, Sad, Neutral, Disgusted, Fearful, Angry
Wu et al. (2023)	Interspeech	Prior work	IEMOCAP	Happy, Sad, Angry, Neutral
Xie et al. (2023)	Speech Communication	RB: other HCI systems, Other, Call screening, Entertainment, RB: car voice assistants, Education	RAVDESS (RAV), EMO-DB, SAVEE, EMOVO	Angry, Boredom, Fearful, Disgusted, Joy, Sad, Neutral, Happy, Surprised
Chen et al. (2023)	ICASSP	Paralingusitic / behavioral studies	IEMOCAP, MELD	Happy: Happy + Excited, Angry, Sad, Neutral, Other or Unspecified
de Oliveira et al. (2023)	Interspeech	Paralingusitic / behavioral studies	MSP-Podcasts	Dominance, Arousal/Activation,Valence or sentiment
Bansal et al. (2023)	Interspeech	RB: other HCI systems, RB: car voice assistants, Call screening	IEMOCAP	Angry, Happy, Sad, Neutral
Mitra et al. (2023)	ICASSP	RB: voice assistants, RB: other HCI systems, Healthcare	MSP-Podcasts	Dominance, Valence or sentiment
Li et al. (2023b)	Interspeech	Prior work, RB: car voice assistants	IEMOCAP, MELD, CMU-MOSI	Angry, Happy: Happy + Excited, Neutral, Valence or sentiment, Disgusted, Fearful, Sad, Surprised, Joy
Zhang and Li (2023)	Interspeech	RB: voice assistants, Social Companion bots	IEMOCAP, MELD	Happy: Happy + Excited, Angry, Neutral, Sad, Joy, Disgusted, Fearful, Surprised
Gao et al. (2023)	Interspeech	RB: other HCI systems	IEMOCAP	Angry, Neutral, Sad, Happy: Happy + Excited
Lashkarashvili et al. (2024)	ICASSP	RB: other HCI systems	IEMOCAP	Happy: Happy + Excited, Sad, Angry, Neutral, Other or Unspecified
Goron et al. (2024)	ICASSP	RB: other HCI systems, Healthcare, Call screening	ASVP-ESD, IEMOCAP, RAVDESS (RAV), CREMA-D, CAFE, EMO-DB	Happy: Happy + Excited, Happy, Sad, Angry, Neutral
Wang and Vadakkepat (2024)	ICASSP	RB: other HCI systems, RB: voice assistants	RAVDESS (RAV), SAVEE, TESS	Calm, Happy, Sad, Angry, Fearful, Surprised, Disgusted, Neutral