

HaessigDB: A Database of Irritable Speech with Intensity Grading

Niklas Weller¹, Marc Grau¹, Ivo Blohm¹

¹ University of St.Gallen, Switzerland

niklas.weller@unisg.ch, marcchristopher.grau@unisg.ch, ivo.blohm@unisg.ch

Abstract

We introduce *HaessigDB*, a curated collection of human-annotated emotional speech snippets with intensity grading. The database contains acted call-center style speech snippets with human ratings on three irritability dimensions, namely annoyance, frustration, and aggression. Each snippet is rated on an ordinal intensity scale, which supports fine-grained analysis beyond binary emotion detection. To improve annotation quality, we construct high-agreement subsets per dimension through iterative processing steps. The resulting corpus preserves temporal ordering, enabling research on irritability development in conversations and early escalation signals. *HaessigDB* addresses a data gap in speech-based emotion research for service settings, where intervention decisions depend on both intensity and temporal development. The database can support supervised learning, trajectory modeling, and design of escalation-aware conversational systems.

Index Terms: speech emotion recognition, irritability in speech, computational paralinguistics, emotion intensity, speech corpus

1. Introduction

The detection of irritability in speech has many important applications. In particular in customer service, automatic fine-grained detection could improve service quality: Irritable customers could be transferred to human agents in call routing systems, and voicebot conversations can automatically be escalated to experts when irritability increases. Such routing decisions matter because customers' negative emotions can lead to negative experiences in bot interactions [1, 2]. When a request is not resolved or when misunderstandings persist, conversations can deteriorate [3]. Consequently, for systems that automate customer service interactions with voice agents, it is essential to identify breakdown risk early enough to trigger corrective action [4].

However, automatic irritability detection in speech remains challenging. Existing speech emotion recognition datasets often do not provide labels that match escalation management needs. Widely used corpora were primarily designed for general emotion recognition. Classic resources such as Emo-DB [5, 6], IEMOCAP [7], and RAVDESS [8] are highly valuable for benchmarking broad emotion categories, but were not designed to capture irritability in service calls and typically do not provide intensity ratings. In practice, binary or coarse emotion labels are often insufficient for corrective action because intervention decisions depend on intensity and temporal development, not only on the presence or absence of a category [9, 10]. For escalation management, systems need an ordinal signal of how irritability evolves over interaction time.

Prior dialog research has shown that problematic interactions can be predicted from early turns and that such predictions can inform recovery policies, including transfer decisions [11, 9]. More recent work reaches a similar conclusion in deployed systems, where frustration detection is linked to retention and user satisfaction outcomes [12, 10]. Together, this line of work indicates that escalation handling is a core operational problem in conversational systems in service environments. Emotion-aware modeling provides two concrete opportunities. First, affective speech cues can reveal early irritability even when lexical content remains formally polite or ambiguous, which can enable adaptive response strategies before failure becomes explicit. Second, if adaptive strategies do not stabilize the interaction, continuous affect tracking can enable timely handoff to a human agent before full conversational breakdown.

Recent multimodal and foundational model research makes this direction increasingly feasible. In conversational AI, multimodal architectures that integrate audio and text improve breakdown detection relative to text-only approaches [10]. In affective computing, current foundation-model discussions emphasize theory-grounded emotion modeling and multimodal integration as key directions for robust real-world systems [13]. However, progress is constrained by data availability. While text resources for sentiment and dialogue-state modeling are abundant, speech resources with intensity labels allowing for more fine-grained prediction of irritability remain comparatively scarce in practical service settings [14, 12].

To address this gap, we introduce a database of human-annotated emotional speech snippets with intensity grading (*HaessigDB*). The corpus focuses on irritability as the target construct and operationalizes it through annoyance, frustration, and aggression ratings on ordinal scales across temporally ordered dialogue snippets. This design supports models that estimate both current irritability level and short-horizon escalation risk. We also provide curated high-agreement subsets for each irritability dimension to support reliable supervised learning. By aligning dataset design with intervention needs in service operations, *HaessigDB* provides a resource for emotion-aware adaptation and handoff policies in conversational systems.

2. Related Work

This section establishes emotion-aware conversational systems as one prominent application field for fine-grained detection of irritability in speech. Additionally, we introduce the construct of irritability and highlight the importance of intensity and temporality in the given application context.

2.1. Emotion-Aware Conversational Systems

Conversational systems deployed in customer service environments are increasingly equipped with some form of emotion awareness. For instance, a growing body of research investigates empathetic conversational agents, where agents are deliberately designed to elicit or project empathy [15]. While some empathetic designs lead to beneficial consumer responses [16], overall findings are mixed [17, 18]. Designing emotion-aware systems typically implies dynamic recognition of a human user’s feelings and emotional state [19]. More generally, in human-computer interaction, the emotional intelligence of a conversational agent is evaluated based on the agent’s ability to perceive user emotions, use emotions, understand emotions, and manage emotions [20]. Compared to text-based agents, acoustic signals in voice interactions provide richer data for emotion detection [20]. Therefore, multimodal approaches to emotion-aware conversational systems are a promising research direction. However, training data are often scarce, unbalanced, and noisy across modalities [21].

2.2. Irritability in Emotional Speech

In this paper, we use irritability to denote a construct of interaction-relevant negative affective states that are plausibly expressed in speech and may vary in degree. We operationalize irritability through three closely related dimensions, namely annoyance, frustration, and aggression. This is in line with prior conceptual work that characterizes irritability as mild anger and as a blend of being easily annoyed or frustrated, with the potential to progress toward more overt anger [22]. Classic and reformulated accounts of the frustration–aggression relationship further motivate treating these states as connected rather than independent: Goal blockage may induce negative affect that can facilitate aggressive inclinations under appropriate cues [23, 24]. More recent empirical work similarly associates frustration episodes with irritation-related emotional responses, reinforcing the relevance of an annoyance–frustration–aggression framing for escalation-sensitive settings [25].

A second premise is that irritability is more usefully treated as graded rather than purely categorical. By intensity, we mean an ordinal notion of “how irritated” a speaker is perceived to be along a given dimension, rather than a binary indicator of whether that dimension is present. This choice is consistent with previous findings showing that many affective judgments are ordinal in structure and that ordinal labeling can better align with human perception than categorical annotation [26]. In speech, early work demonstrated that annoyance and frustration can be inferred from prosodic cues in dialogue, suggesting that these emotions have vocal correlates that are noticeable to listeners [27]. More recent work in complaint-service call analysis has explicitly modeled annoyance as a degree, including segment-level annotation schemes intended to capture shifts in annoyance intensity over a conversation. [28, 29]

Finally, we treat irritability as temporal within an interaction. By temporality, we mean that perceived intensities may increase, plateau, or decrease as a conversation unfolds over time. This evolution can be consequential for downstream decisions such as escalation management. Analyses of customer-facing calls have emphasized tracking annoyance as it develops within a call, rather than assuming a stationary emotional condition [29]. More broadly, work on the temporality of frustration in learning contexts underscores that frustration is not a static state and can transform as conditions change, which motivates studying trajectories rather than isolated snapshots [30]. Thus,

our database contains both a temporal and an ordinal intensity component.

3. Gathering and Curating the Data

We created the database in three phases: First, we generated the typical textual call transcripts and hired voice actors to record them. Second, we crowd-sourced the annotation of the three irritability dimensions (annoyance, frustration, aggression) on ordinal scales. Third, we iteratively removed snippets with low inter-rater agreement until we had a curated high-agreement labeled dataset for each irritability dimension.

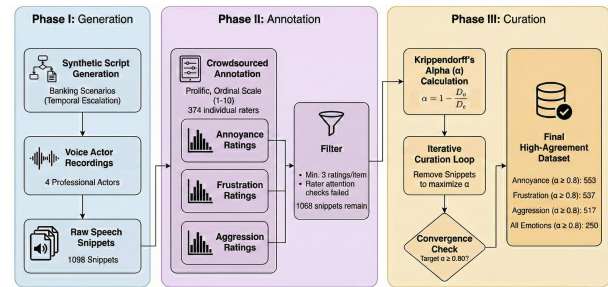


Figure 1: Methodology of dataset creation and curation.

3.1. Text Material

The text material was generated synthetically to resemble customer service center conversations in a banking context. The synthetic generation pipeline ensured that a variety of scenarios are covered. Additionally, it was configured such that the roleplayed customer would become increasingly irritable as their request could not be resolved to their satisfaction. This design choice was intended to convey the temporality and graded affective states to actors, rather than a single static emotional condition. In total, 45 distinct call scripts were selected for recording.

3.2. Selection of Actors

We hired four voice actors through the online platform Upwork. All of them were recruited from different countries where English is an official language, had experience in voice acting, and a professional recording setup. Actors were compensated at an average wage of 64 USD per recording hour.

3.3. Online Rating via Prolific

To obtain ratings of the intended irritability dimensions, we segmented full call recordings into snippets roughly corresponding to individual sentences. The average snippet recording is 3.91 seconds long with a standard deviation of 2.13 seconds.

Each voice snippet was rated by multiple annotators on an ordinal scale from 1–10 for aggression, frustration, and annoyance. In total, we recruited 374 annotators on Prolific, that rated a maximum of 20 snippets each. We included attention checks and removed all ratings from the nine annotators who failed them. For the reliability analysis, we retained only snippets with at least three ratings per irritability dimension (28 snippets were removed). This filtering step was used to mitigate issues arising from sparsely rated items caused by failed attention checks. After filtering, we thus ended up with 587 snippets rated three times and 482 snippets rated four times.

3.4. Curating the Data

We quantified inter-rater reliability using Krippendorff’s alpha α , computed at the ordinal level of measurement. In this setting, Krippendorff’s α is used as a chance-corrected agreement measure that accommodates varying numbers of ratings per item [31]. We computed α separately for each irritability dimension. This per-dimension computation reflects that agreement may differ across emotion constructs, rather than constituting a single aggregate property of the dataset.

Because α is defined over a set of items rather than per item, we constructed a curated subset of snippets whose annotations achieve a target dataset-level reliability. Starting from all eligible snippets, we iteratively removed snippets that most increased dataset-level α . Iteration stopped when $\alpha \geq 0.80$ or when no further removals increased α . We refer to the resulting curated subset as the high-agreement subset.

4. Evaluating the Data

In this section, we provide a description and preliminary evaluation of our database.

4.1. Overview of Datasets

After curation, we arrived at three separate datasets for each target dimension. While the raw dataset after filtering contained 1068 valid snippets, the high-agreement subsets for annoyance, frustration, and aggression contained 553, 537, and 516 snippets respectively. Additionally, we provide two variants of a merged dataset containing 250 snippets when using an inner join and 832 snippets for the outer join method. Table 1 gives an overview of the different subsets. All datasets are made publicly available on Zenodo and Hugging Face.¹

Table 1: Overview of Raw and Curated (High-agreement) Datasets

Name	Description	# Snippets
All ratings	Full collected voice data and ratings	1068
Annoyance ($\alpha > 0.8$)	High-agreement subset for annoyance	553
Frustration ($\alpha > 0.8$)	High-agreement subset for frustration	537
Aggression ($\alpha > 0.8$)	High-agreement subset for aggression	516
Inner join ($\alpha > 0.8$)	High-agreement subset satisfying $\alpha > 0.8$ across all emotions	250
Outer join ($\alpha > 0.8$)	High-agreement subset satisfying $\alpha > 0.8$ in at least one emotion	832

Figure 2 shows the distribution of ratings for each irritability dimension before and after the curation of the dataset. While the raw ratings for frustration and annoyance approximately follow a normal distribution, the rating distribution for aggression is skewed right. After curation, the most frequent rating values are attenuated, resulting in a flatter distribution. For a labeled data set, this is generally beneficial. For example, when the dataset is utilized as training data for a machine learning model, a more even distribution among labels will prevent issues arising from class imbalance.

¹<https://huggingface.co/datasets/nwllr/haessigDB>

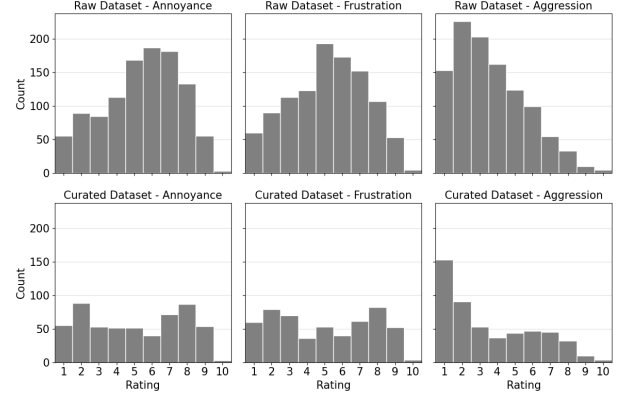


Figure 2: Label distributions for each emotion before and after the data curation.

Interestingly, in the curated subset, almost no ratings are filtered out for the lowest level of each emotion. This indicates that there tends to be high agreement in the ratings when the respective emotion is clearly absent.

4.2. Temporal Component

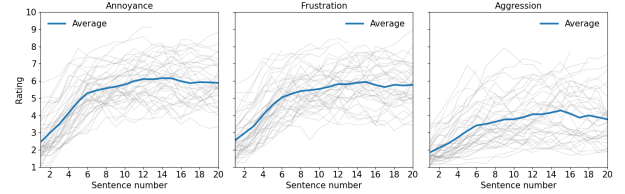


Figure 3: Smoothed emotion trajectories for the first 20 snippets of each call.

Figure 3 shows the trajectories of perceived aggression, frustration, and annoyance for the first 20 snippets of each call. The grey lines show the rolling-mean (window = 5 snippets) ratings for individual calls, and the thick line shows the average trajectory across calls.

We can observe, that, for the first 10 snippets, each irritability dimension monotonically increases on average. Afterwards, they remain at a high level. This is in line with our intention of increasingly irritable customers in the synthetic call scripts. It also illustrates the suitability of our dataset for early stage detection, as - especially in the first 10 snippets - the snippets show relatively low levels of irritability that are followed by higher levels.

In both Figure 2 and Figure 3, it is apparent that aggression levels are perceived lower than frustration and annoyance. This aligns well with the previously mentioned frustration-aggression relationship, which claims that frustration can lead to aggression and usually arises when individuals face obstacles in achieving their goals [23, 24].

4.3. Baseline Evaluation of Emotion Detection Models

There is a lack of off-the-shelf models explicitly targeting the detection of irritability or its subdimensions. Thus, we evaluated the performance of five popular speech emotion recognition models from Hugging Face on our database (search string

”emotion” in the audio classification category). Concretely, we selected the five most liked models for emotional speech recognition that included an irritability proxy (anger/angry) as a label². We related the categorical irritability label (anger/angry) to our ordinal ratings of each irritability subdimension (annoyance, frustration, aggression) by performing a threshold sweep. For a minimum threshold $t \in 0, \dots, 9$, snippets with human rating $> t$ were treated as ground-truth *anger present*, and all remaining snippets as *anger absent*. For reporting, we focus on the results for the aggression subdimension. The results for annoyance and frustration follow a similar pattern.

To evaluate the model performance, we focus on the recall score that the models achieve across each threshold. A desirable model would have a high recall for a high threshold value (only snippets that are rated highly aggressive by the human raters are assigned the *anger present* label) and a medium to low level recall for a low threshold value (most snippets are assigned to *anger present*, even though many of them are not rated at a high level of aggression by humans as apparent in Figure 2). On the thresholds in between, the recall rate would ideally monotonically increase.

The recall scores across the threshold sweep reveal the major failure modes of the existing models. We find that there is only a small difference between the recall score for a low and high threshold across all tested models. They either remain at very low levels or very high levels throughout the whole threshold sweep. This highlights two major limitations of deploying existing categorical speech emotion recognition models for escalation-sensitive settings. First, existing categorical emotion models may not provide stable behavior across operationalizations of irritability, because their classification propensity varies materially across architectures and training regimes. Second, none of these models provides an ordinal notion of *how aggressive* a snippet is. The relatively flat lines resulting from the threshold sweep illustrate that, even when models can recover high-aggression cases, their categorical outputs do not resolve intermediate aggression levels in a stable fine-grained manner. This is precisely what matters for many applications including early intervention and escalation management.

4.4. Fine-tuning of Emotion Detection Models

In a next step, we evaluate whether our curated dataset can be used to fine-tune existing speech models for continuous intensity prediction. We again use the subdimension aggression to evaluate the models. We split the labeled audio data into a training and testing set (80/20 random split, not speaker-disjoint) and fine-tune two of the aforementioned pretrained speech encoders using Low-Rank Adaptation (LoRA) [37]. We add a simple regression head predicting a single scalar score per clip to adjust the models for ordinal outputs. Specifically, we fine-tune the models based on HuBERT and XLSR-wav2vec for five epochs using a batch size of 4 and a learning rate of 1×10^{-4} on a 16GB NVIDIA T4 GPU, which is sufficient for parameter-efficient adaptation of these transformer backbones. After five epochs, both models perform well on the held-out test set. HuBERT achieves a mean average error (MAE) of 1.502 with a Pearson correlation of $r = 0.607$ between predicted and true

intensity ratings. XLSR-wav2vec achieves an MAE of 1.333 with a Pearson correlation of $r = 0.710$. Overall, the results suggest that the dataset contains a learnable acoustic signal for aggression intensity and supports effective parameter-efficient adaptation of pretrained speech encoders. Across both absolute error and association with human judgments, XLSR-wav2vec shows stronger performance, indicating better alignment with the relative ordering and magnitude of aggression ratings in this dataset.

To compare this regression result to the classification baseline models, we convert the predictions to the binary label of *anger present/absent* when the predicted aggression value is higher/lower than five. Figure 4 illustrates the recall performance across the threshold sweep before and after fine-tuning. We observe that the dotted curves (fine-tuned models) much more closely resemble the desirable behavior as described in the previous section. Once again the fine-tuned XLSR-wav2vec performs best. Nevertheless, the improvement of the fine-tuned HuBERT model is also clearly visible. Overall, the characteristics of both models after fine-tuning on *HaessigDB* are much more similar and useful than before.

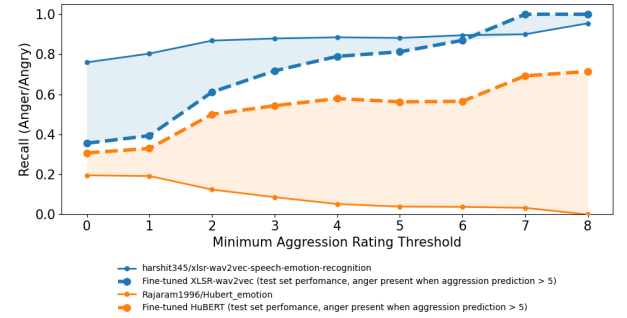


Figure 4: Irritability detection performance on the *HaessigDB* aggression subset of two emotional speech detection models from Huggingface before and after fine-tuning.

We emphasize that these results rely on an internal 80/20 split and therefore do not, on their own, establish out-of-sample generalizability across speakers, domains, or recording conditions. Nonetheless, given the absence of established benchmarks for irritable speech detection with ordinal intensity ratings, the observed error levels and correlations provide initial evidence that pretrained speech encoders can be adapted to predict fine-grained intensity labels for irritability subdimensions. This capability is salient for applied settings such as customer service, where early and graded detection of escalating affect may be more actionable than categorical emotion labels.

5. Conclusion

This paper introduced *HaessigDB*, a curated corpus of temporally ordered call-center speech snippets with ordinal intensity ratings for annoyance, frustration, and aggression. The database enables modeling irritability as graded and evolving over interaction time. Initial experiments suggest that pretrained speech encoders can be adapted to predict these fine-grained intensity labels, underscoring the feasibility of intensity-aware affect modeling for escalation management in conversational systems.

²The selected models are (1) *harshit345/xlsr-wav2vec-speech-emotion-recognition* [32], (2) *Rajaram1996/HuBERT_emotion* [33], (3) *speechbrain/emotion-recognition-wav2vec2-IEMOCAP* [34], (4) *ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition* [35], (5) *firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3* [36]

6. Acknowledgements

This work was supported by Innosuisse and is part of the project 104.954 IP-SBM.

7. Generative AI Use Disclosure

We used generative AI tools solely for language editing, manuscript polishing (e.g., improving clarity, grammar, and style), and coding assistance. We did not use generative AI to generate original scientific content, develop the study design, or produce results. All substantive intellectual contributions, including the study conception, methodology, data collection, analysis, interpretation, and final wording and conclusions, were produced and verified by the authors.

8. References

- [1] B. Li, L. Liu, W. Mao, Y. Qu, and Y. Chen, "Voice artificial intelligence service failure and customer complaint behavior: The mediation effect of customer emotion," *Electronic Commerce Research and Applications*, vol. 59, p. 101261, May 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1567422323000261>
- [2] C. Crolic, F. Thomaz, R. Hadi, and A. T. Stephen, "Blame the Bot: Anthropomorphism and Anger in Customer-Chatbot Interactions," *Journal of Marketing*, vol. 86, no. 1, pp. 132–148, Jan. 2022. [Online]. Available: <https://doi.org/10.1177/00222429211045687>
- [3] R. W. Zhang, X. Liang, and S.-H. Wu, "When chatbots fail: exploring user coping following a chatbots-induced service failure," *Information Technology & People*, vol. 37, no. 8, pp. 175–195, Jul. 2024. [Online]. Available: <https://doi.org/10.1108/ITP-08-2023-0745>
- [4] U. Gnewuch and F. Reinkemeier, "Overcoming Breakdowns in Customer-Chatbot Interaction: Design and Impact of Collaborative Repair Strategies," *Management Information Systems Quarterly*, pp. 1–33, Jul. 2025. [Online]. Available: <https://doi.org/10.25300/MISQ/2025/18742>
- [5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, and others, "A database of german emotional speech," in *Proc. of interspeech*, vol. 5, 2005, pp. 1517–1520.
- [6] F. Burkhardt, O. Schrüfer, U. Reichel, H. Wierstorf, A. Derington, F. Eyben, and B. Schuller, "EmoDB 2.0: a database of emotional speech in a world that is not black or white but grey," in *Proc. of interspeech*, Aug. 2025, pp. 4488–4492.
- [7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008. [Online]. Available: <https://doi.org/10.1007/s10579-008-9076-6>
- [8] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, May 2018. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391>
- [9] S. A. Chowdhury, E. A. Stepanov, and G. Ricciardi, "Predicting User Satisfaction from Turn-Taking in Spoken Conversations," 2016, pp. 2910–2914. [Online]. Available: https://www.isca-archive.org/interspeech_2016/chowdhury16_interspeech.html
- [10] M. M. M. Miah, U. Schnaithmann, A. Raghuvanshi, and Y. Son, "Multimodal Contextual Dialogue Breakdown Detection for Conversational AI Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Y. Yang, A. Davani, A. Sil, and A. Kumar, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 303–314. [Online]. Available: <https://aclanthology.org/2024.naacl-industry.25/>
- [11] M. Walker, I. Langkilde, J. Wright, A. Gorin, and D. Litman, "Learning to predict problematic situations in a spoken dialogue system: experiments with how may I help you?" in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, ser. NAACL 2000. USA: Association for Computational Linguistics, Apr. 2000, pp. 210–217. [Online]. Available: <https://dl.acm.org/doi/10.5555/974305.974333>
- [12] M. Hernandez Caralt, I. Sekulic, F. Carevic, N. Khau, D. N. Popa, B. Guedes, V. Guimaraes, Z. Yang, A. Manso, M. Reddy, P. Rosso, and R. Mathis, "'Stupid robot, I want to speak to a human!'" User Frustration Detection in Task-Oriented Dialog Systems," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, K. Darwish, and A. Agarwal, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 276–285. [Online]. Available: <https://aclanthology.org/2025.coling-industry.23/>
- [13] B. Schuller, A. Mallol-Ragolta, A. P. Almansa, I. Tsangko, M. M. Amin, A. Semertzidou, L. Christ, and S. Amiriparian, "Affective computing has changed: the foundation model disruption," *npj Artificial Intelligence*, vol. 2, no. 1, p. 16, Jan. 2026. [Online]. Available: <https://www.nature.com/articles/s44387-025-00061-3>
- [14] P. Larrouy-Maestri, D. Poeppel, and M. D. Pell, "The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions," *Perspectives on Psychological Science*, vol. 20, no. 4, pp. 623–638, Jan. 2024. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12231869/>
- [15] A. Cuadra, M. Wang, L. A. Stein, M. F. Jung, N. Dell, D. Estrin, and J. A. Landay, "The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–18. [Online]. Available: <https://dl.acm.org/doi/10.1145/3613904.3642336>
- [16] A. Mari, A. Mandelli, and R. Algesheimer, "Empathic voice assistants: Enhancing consumer responses in voice commerce," *Journal of Business Research*, vol. 175, p. 114566, Mar. 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0148296324000705>
- [17] A. Juquelier, I. Poncin, and S. Hazée, "Empathic chatbots: A double-edged sword in customer experiences," *Journal of Business Research*, vol. 188, p. 115074, Feb. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0148296324005782>
- [18] A. Perry, "AI will never convey the essence of human empathy," *Nature Human Behaviour*, vol. 7, no. 11, pp. 1808–1809, Nov. 2023. [Online]. Available: <https://www.nature.com/articles/s41562-023-01675-w>
- [19] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020. [Online]. Available: <https://aclanthology.org/2020.cl-1.2/>
- [20] J. Hu, Y. Huang, X. Hu, and Y. Xu, "The Acoustically Emotion-Aware Conversational Agent With Speech Emotion Recognition and Empathetic Responses," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 17–30, Jan. 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/9888003>
- [21] Y. Shou, T. Meng, W. Ai, F. Fu, N. Yin, and K. Li, "A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning," *ACM Trans. Inf. Syst.*, vol. 44, no. 2, pp. 47:1–47:48, Jan. 2026. [Online]. Available: <https://dl.acm.org/doi/10.1145/3786343>
- [22] P. C. Barata, S. Holtzman, S. Cunningham, B. P. O'Connor, and D. E. Stewart, "Building a definition of irritability from academic

- definitions and lay descriptions,” *Emotion Review*, vol. 8, no. 2, pp. 164–172, 2016.
- [23] J. Dollard, L. W. Doob, N. E. Miller, O. H. Mowrer, and R. R. Sears, *Frustration and Aggression*. New Haven, CT: Yale University Press, 1939.
- [24] L. Berkowitz, “Frustration–aggression hypothesis: Examination and reformulation,” *Psychological Bulletin*, vol. 106, no. 1, pp. 59–73, 1989.
- [25] H. Yazdi, C. Wickman, M. Ljung Aust, I. Selbing, L. Kowalski, and J. Axelsson, “Understanding frustration triggers and emotional responses in driving situations,” *Scientific Reports*, vol. 14, no. 1, p. 28613, 2024.
- [26] G. N. Yannakakis, R. Cowie, and C. Busso, “The Ordinal Nature of Emotions: An Emerging Approach,” *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, Jan. 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/8521685>
- [27] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *Proceedings of ICSLP 2002*, 2002.
- [28] J. I. Manso and M. I. Torres, “Analyzing the expression of annoyance during phone calls to complaint services,” in *7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2016)*, Wroclaw, Poland, October 16–18, 2016. IEEE, 2016, pp. 103–106. [Online]. Available: <https://doi.org/10.1109/CogInfoCom.2016.7804533>
- [29] J. Irastorza and M. I. Torres, “Tracking the expression of annoyance in call centers,” in *Cognitive Infocommunications, Theory and Applications*, ser. Topics in Intelligent Engineering and Informatics, R. Klempous, J. Nikodem, and P. Z. Baranyi, Eds. Cham: Springer, 2019, vol. 13, pp. 131–151.
- [30] D. R. King, S. M. Ritchie, M. Sandhu, S. Henderson, and B. Boland, “Temporality of emotion: Antecedent and successive variants of frustration when learning chemistry,” *Science Education*, vol. 101, no. 4, pp. 639–672, 2017.
- [31] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 4th ed. SAGE Publications, 2019.
- [32] A. Conneau, A. Baeovski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised Cross-Lingual Representation Learning for Speech Recognition,” 2021, pp. 2426–2430. [Online]. Available: https://www.isca-archive.org/interspeech.2021/conneau21_interspeech.html
- [33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9585401>
- [34] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [35] E. H. Calabrés, “wav2vec2-lg-xlsr-en-speech-emotion-recognition (revision 17cf17c),” 2024. [Online]. Available: <https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition>
- [36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models.” *ICLR*, vol. 1, no. 2, p. 3, 2022.