# Mixed Signals: Understanding Model Disagreement in Multimodal Empathy Detection

**Maya Srikanth**
Columbia University
ms6198@columbia.edu

**Run Chen**
Columbia University
runchen@cs.columbia.edu

**Julia Hirschberg**
Columbia University
julia@cs.columbia.edu

## Abstract

Multimodal models play a key role in empathy detection, but their performance can suffer when modalities provide conflicting cues. To understand these failures, we examine cases where unimodal and multimodal predictions diverge. Using fine-tuned models for text, audio, and video, along with a gated fusion model, we find that such disagreements often reflect underlying ambiguity, as evidenced by annotator uncertainty. Our analysis shows that dominant signals in one modality can mislead fusion when unsupported by others. We also observe that humans, like models, do not consistently benefit from multimodal input. These insights position disagreement as a useful diagnostic signal for identifying challenging examples and improving the robustness of empathy systems.

## 1 Introduction

Empathy recognition in human communication is a nuanced and multifaceted task and a core component of socially intelligent systems (Fung et al., 2016). Commonly defined as the capacity to understand others and share their emotional experiences, empathy encompasses both cognitive perspective-taking and affective resonance (Baumeister and Vohs, 2007). In human interactions, language, speech, and visual cues jointly convey emotional intent (Holler and Levinson, 2019). For example, a speaker's verbal message may appear neutral, yet their vocal prosody or facial expressions may signal warmth or concern. It is then the listener's responsibility to draw inferences about meaning based on a *combination* of these signals.

For AI systems, effectively interpreting these multimodal signals requires not only accurate unimodal representations but also robust integration of potentially conflicting information across modalities. Despite recent advances in multimodal emotion recognition (Jabeen et al., 2021), empathy recognition remains particularly complex, as unlike discrete emotions such as anger or joy, empathy often arises from subtle contextual cues that may not align across modalities (Hasan et al., 2023). For example, a neutral utterance might be perceived as warm or concerned when accompanied by a sympathetic tone or expression.

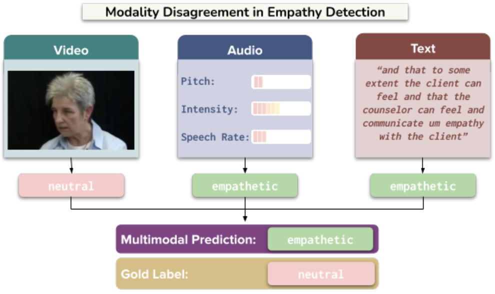

Modality Disagreement in Empathy Detection
Video
Audio
Text
Pitch:
Intensity:
Speech Rate:
"and that to some extent the client can feel and that the counselor can feel and communicate um empathy with the client"
neutral
empathetic
empathetic
Multimodal Prediction:
empathetic
Gold Label:
neutral

Figure 1: Given classifications provided by a single modality, we identify cases where integrating additional modalities leads to a different prediction. We analyze these differences to understand when and why they occur.

Our work investigates some of the complexities of multimodal empathy detection by examining instances of disagreement between multimodal models and their unimodal counterparts. In parallel, humans annotate unimodal and multimodal examples in our dataset for the presence of empathy. Our analyses reveal that instances of multimodal and unimodal model disagreement often correspond to examples that are difficult for human annotators as well, highlighting examples that are particularly challenging, ambiguous, or nuanced. By linking multimodal and unimodal *model* disagreement to *human* disagreement, we offer new insight into the limitations of current empathy modeling and highlight the value of disagreement-based analysis in socially grounded language tasks.

## 2 Related Work

**Empathy Modeling.** Early computational work on empathy has focused on generating emotionally

*Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1978–1991
December 20-24, 2025 ©2025 Association for Computational Linguistics

relevant textual responses (Rashkin et al., 2019; Li et al., 2019), but these approaches are inherently limited by the absence of non-verbal cues, which are critical to empathic understanding. Recent datasets such as EMPATHICSTORIES++ (Shen et al., 2024), MEDIC (Zhu et al., 2023), EMMI (Galland et al., 2024) and Chen et al. (2024b) address this limitation by incorporating context, speech, and facial expressions, enabling more comprehensive modeling of empathy. These resources have motivated frameworks such as PEGS (Zhang et al., 2024), EMOKNOB (Chen et al., 2024a) and SYNTHEMPATHY (Chen et al., 2025), which further extend multimodal empathetic generation by leveraging large language models (LLMs) and large audio-language models. Despite these advances, empathy still remains difficult to model due to its reliance on subtle, often conflicting signals across modalities. Prior work has largely focused on improving multimodal fusion strategies under the assumption that modalities are complementary (Zadeh et al., 2017; Tsai et al., 2019), but has paid less attention to when fusion may fail or introduce noise.

**Dataset Difficulty.** Complementary lines of work have investigated data difficulty and model disagreement as tools for understanding model behavior. Swayamdipta et al. (2020) propose *dataset cartography*, a method to identify hard or ambiguous training samples, showing how difficulty-aware instance selection improves benchmarking and reveals mislabeled or trivial examples. Saha et al. (2022) demonstrate that difficult instances are also harder for both humans and models to explain, and Wang et al. (2023)'s Learning-From-Disagreement (LFD) framework underscores the importance of examining disagreements between models to gain deeper, actionable insights into their behaviors. Although ambiguity is intrinsic to empathy modeling, disagreement-based diagnostics do remain underexplored. As such, we leverage modality disagreement to flag difficult examples that both mislead fusion models and elicit annotator uncertainty.

## 3 Experiment 1: Identifying Complex Examples from Modality Disagreement

Disagreement between models trained on different modalities can reveal challenging, nuanced, or ambiguous examples. Here, we identify and analyze such cases of disagreement in binary empathy detection using a multimodal English empathy speech dataset collected from Youtube (Chen et al., 2024b) (referred to as EMPSPEECH) consisting of 1,718 manually annotated English speech segments labeled as empathetic or neutral (Appendix A).

| Modality | Model | Accuracy | F1 |
|---|---|---|---|
| Text | **RoBERTa** | **0.75±0.02** | **0.73 ±0.02** |
| | DeBERTa | 0.69±0.02 | 0.68±0.02 |
| Audio | **HuBERT** | **0.72±0.01** | **0.71±0.01** |
| | Wav2Vec2 | 0.68±0.01 | 0.63±0.02 |
| Video | **VideoMAE** | **0.77±0.02** | **0.77±0.02** |
| | TimesFormer | 0.64±0.02 | 0.62±0.02 |
| **Fusion (All Modalities)** | | **0.76±0.02** | **0.72±0.02** |

Table 1: Fine-tuned model performance by modality on empathy classification (mean ± std over five runs).

**Experimental Setup.** Examples in EMPSPEECH include video segments spanning three modalities: text (transcript), audio (speech), and video. The task is to predict whether the input contains empathetic (1) or neutral (0) speech.

We finetune two models per modality on the training set from EMPSPEECH: ROBERTA (Liu et al., 2019) and DEBERTA (He et al., 2021) for text, HUBERT (Hsu et al., 2021) and WAV2VEC2 (Baevski et al., 2020) for audio, and VIDEOMAE (Tong et al., 2022) and TIMESFORMER (Bertasius et al., 2021) for video (Appendix B.1).[1] Then, we extract 768-dimensional embeddings from each best-performing unimodal model (ROBERTA, HUBERT, and VIDEOMAE; Table 1) to train a multimodal fusion model that projects all three modalities into a shared latent space (Appendix B.2). Each modality embedding passes through an independent sigmoid gate that adaptively scales its contribution before fusion. The gated embeddings are then passed through an additive attention layer: each is projected into a shared attention space and scored against a learned attention vector. These scores are normalized across modalities to compute a weighted sum that forms the fused representation.

**Results.** We evaluate all models (unimodal and multimodal) on the test split of EMPSPEECH to identify *disagreements*, or examples where two models with varying input modalities assign *different* labels, highlighting those cases where different modalities may carry ambiguous, conflicting, or modality-specific signals.

Text shows the highest disagreement with audio and video (Table 2), while audio and video align

[1] The hidden layer dimensions of all models we consider are similar.

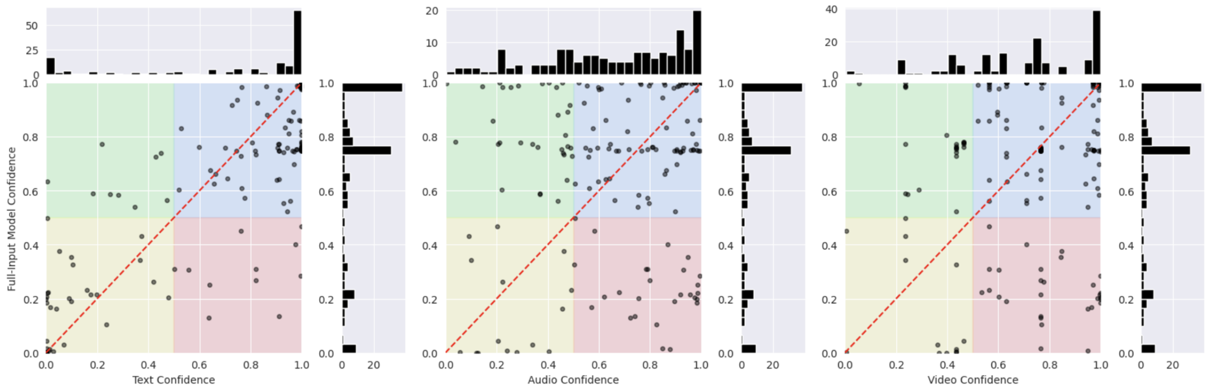


Figure 2: Comparing predictions between unimodal (text, audio, video) and multimodal models. We highlight regions where model predictions *agree* (blue and yellow quadrants) and disagree (red and green quadrants).

| **Modality** | **Text** | **Audio** | **Video** |
|---|---|---|---|
| **Text** | – | 0.338 | 0.318 |
| **Audio** | 0.338 | – | 0.253 |
| **Video** | 0.318 | 0.253 | – |
| **Full** | 0.214 | 0.383 | 0.331 |

Table 2: Pairwise disagreement rates among unimodal models and the fusion model, computed as the fraction of test examples with differing predictions.

| **Feature** | **Red vs. Blue** | | **Green vs. Blue** | |
|---|---|---|---|---|
| | p-value | Direction | p-value | Direction |
| **valence** | **0.0047** | $\mu_{\textbf{blue}} > \mu_{\textbf{red}}$ | 0.5166 | $\mu_{\text{green}} > \mu_{\text{blue}}$ |
| **arousal** | **0.0065** | $\mu_{\textbf{blue}} > \mu_{\textbf{red}}$ | **0.0136** | $\mu_{\textbf{blue}} > \mu_{\textbf{green}}$ |
| **Mean Pitch** | **0.0100** | $\mu_{\textbf{blue}} > \mu_{\textbf{red}}$ | **0.0001** | $\mu_{\textbf{blue}} > \mu_{\textbf{green}}$ |
| **dominance** | **0.0108** | $\mu_{\textbf{blue}} > \mu_{\textbf{red}}$ | 0.0667 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| **Min Pitch** | **0.0333** | $\mu_{\textbf{blue}} > \mu_{\textbf{red}}$ | **0.0001** | $\mu_{\textbf{blue}} > \mu_{\textbf{green}}$ |
| **Jitter** | **0.0347** | $\mu_{\textbf{red}} > \mu_{\textbf{blue}}$ | 0.0667 | $\mu_{\text{green}} > \mu_{\text{blue}}$ |
| **Max Intensity** | 0.1260 | $\mu_{\text{red}} > \mu_{\text{blue}}$ | **0.0023** | $\mu_{\textbf{green}} > \mu_{\textbf{blue}}$ |

Table 3: T-test results comparing red vs. blue and green vs. blue examples for audio features with $\alpha = 0.05$. Statistically significant results are bolded. See Appendix D for full table.

more closely. This difference likely reflects shared nonverbal cues such as prosody and facial expression. The fusion model's minimal disagreement with text suggests a bias toward verbal content, possibly mirroring the annotators' own reliance on textual signals.

Figure 2 visualizes disagreement regions between each unimodal model and the fusion model. We plot unimodal confidence (x-axis) against fusion confidence (y-axis) in the correct label; hence confidence greater than 0.5 results in a correct prediction. This yields four quadrants: green (multimodal correct, unimodal incorrect), red (multimodal incorrect, unimodal correct), blue (both correct), and yellow (both incorrect). Red and green quadrants are disagreement regions which we explore to identify complex examples.

| **AU** | **p (R vs B)** | **Dir** | **p (G vs B)** | **Direction** |
|---|---|---|---|---|
| **AU04** | **0.0106** | **red > blue** | 0.3682 | green > blue |
| **AU12** | **0.0174** | **blue > red** | 0.8977 | green > blue |
| **AU05** | 0.1837 | blue > red | **<0.0001** | **blue > green** |

Table 4: T-test results comparing AU activation rates between red vs. blue and green vs. blue with $\alpha = 0.05$. Statistically significant results are bolded. See Appendix D for full table

### 3.1 Modality-Based Feature Analysis

To better understand examples in disagreement regions, we extract and analyze modality-based human interpretable features.

**Audio.** We extract twelve prosodic and paralinguistic features from audio signals: nine low-level acoustic features using PRAAT (Boersma and Weenink, 1992–2022) and PARSELMOUTH (Jadoul et al., 2018), and three high-level affective dimensions: valence, arousal, and dominance using a finetuned WAV2VEC2 (Wagner et al., 2023) model. We compare feature distributions using t-tests for examples in disagreement quadrants ( red and green ) compared to those in the blue quadrant, signifying non-ambiguous, easy examples. Blue examples have several significantly elevated pitch-related values than red examples (Table 3), suggesting that stronger prosodic fluctuations are frequently corroborated by other modalities. Examples in the green quadrant show significantly higher *Max Intensity* than in blue, potentially reflecting the role of volume-based emphasis in aiding unimodal predictions. Both red and green examples exhibit significantly lower arousal than blue examples, suggesting that these less-aroused, subtler examples lack sufficient affective intensity, which misleads both unimodal and multimodal models.

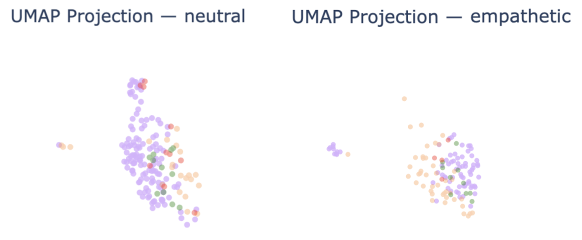


Figure 3: UMAP of text-only embeddings for empathetic (left) vs. neutral (right) examples, colored by modality disagreement; red and green points cluster near the decision boundary, marking ambiguous cases.

| Quadrant | Mean Entropy | St. Dev |
|---|---|---|
| Red | 0.670 | 0.017 |
| Blue | 0.259 | 0.222 |
| Yellow | 0.347 | 0.196 |
| Green | 0.565 | 0.192 |

Table 5: Mean entropy of the fusion model grouped by quadrant

**Video.** We examine facial action unit (AU) activations (Baltrušaitis et al., 2016) from video. AU04 (Brow Lowerer), AU12 (Lip Corner Puller), and AU05 (Upper Lid Raiser) show significant differences across example types, revealing how specific facial expressions contribute to perceptual ambiguity (Table 4). AU04 is more active in red examples than blue, indicating that, despite its visually strong presence, its signal conflicts with other modalities. In contrast, AU12, which is associated with positive affect, and AU05, which is linked to attentiveness (Friesen and Ekman, 1978), both show greater activation in blue examples than in red and green, respectively, suggesting that these expressions may serve as clearer cues that are more consistently interpreted across modalities. Our findings indicate that fine-grained facial signals may contribute to perceptual complexity in the visual stream.

**Text.** Visualizing UMAP (Sainburg et al., 2021) projections of text embeddings (Figure 3) reveals that examples in disagreement regions (red and green) cluster along the boundary between consistently correct (blue) and consistently incorrect (yellow) examples. Rather than forming isolated clusters, disagreement examples occupy transition zones in the embedding space: areas where semantic cues are weak. This underscores our finding that red and green examples are ambiguous and confirms modality disagreement as a reliable marker of challenging examples in empathy detection.

| Quadrant | Unimodal Judgment | Multimodal Judgment | $\Delta$ |
|---|---|---|---|
| Red | 0.301 | 0.164 | -0.137 |
| Blue | 0.379 | 0.646 | 0.267 |
| Yellow | 0.225 | 0.329 | 0.104 |
| Green | 0.482 | 0.218 | -0.264 |

Table 6: Cohen's Kappa between internal and external annotators, computed separately for each quadrant and prediction round.

### 3.2 Uncertainty Analysis

To ensure that the patterns observed in the disagreement quadrants are not simply a byproduct of model uncertainty, we compute the mean predictive entropy from the fusion model's posterior for examples of each quadrant (Table 5).

We observe a pronounced divergence in uncertainty: the disagreement quadrants (red and green) have a substantially higher mean predictive entropy than those of the combined agreement quadrants (blue and yellow). Independent-samples $t$-tests at $\alpha = 0.05$ confirm that the difference is statistically significant, with disagreement quadrants showing a higher mean predictive entropy than agreement quadrants ($p = 0.001$). This disparity indicates that model disagreement often co-occurs with high uncertainty, suggesting that examples in the red and green quadrants are both challenging and inherently ambiguous due to conflicting modality signals.

## 4 Experiment 2: Characterizing Complex Examples

We further assess whether model disagreements stem from data ambiguity using a human annotation study that tests whether examples from the two disagreement regions (red and green quadrants) are equally challenging for annotators.

**Annotation Setup.** We sample 204 examples evenly split across the four quadrants of each Figure 2 modality plot. For each example, annotators provide a binary judgment (empathetic or neutral) from a unimodal signal, then a judgment from the full multimodal version (instructions in Appendix C), allowing us to track how human predictions shift with additional modality signals and to understand the cognitive burden of multimodal integration. All examples were annotated by one author and one external annotator. Table 13 in the Appendix showcases frames and transcripts for four examples, along with annotator judgments.

**Results.** Annotator *disagreement*, measured with Cohen's Kappa (Cohen, 1960), can signal complex

| Quadrant | Unimodal Judgment | Multimodal Judgment | Δ |
|---|---|---|---|
| Red | 0.347 | 0.143 | -0.204 |
| Blue | 0.364 | 0.533 | 0.169 |
| Yellow | 0.304 | 0.548 | 0.244 |
| Green | 0.573 | 0.329 | -0.244 |

Table 7: Cohen's Kappa between internal and external annotators for examples of at least six tokens (the dataset median), computed separately for each quadrant and prediction round.

phenomena in examples (Jiang and de Marneffe, 2022; Pavlick and Kwiatkowski, 2019) such as uncertainty in meaning leading to discrepancies in reasoning. In disagreement regions (red and green), we see a *decrease* in annotator agreement between unimodal and multimodal judgments (Table 6), indicating that humans diverge when weighing signals across modalities. In contrast, annotator agreement *improves* upon examples where unimodal and multimodal model predictions are in agreement, supporting our hypothesis that these examples are relatively unambiguous and can be reliably interpreted once the full context is available (Table 6).

We repeat the analysis on the subset of clips containing at least six tokens (the median utterance length in EMPSPEECH). Examples below the median often include short phrases and backchannels, while examples above the median are often complete sentences with richer lexical and syntactic structure. As shown in Table 7, the red and green quadrant utterances continue to exhibit a substantial drop in Cohen's $\kappa$ compared to blue and yellow quadrant utterances, which exhibit substantial gains with additional information. These results collectively corroborate our hypothesis that modality disagreement can serve as a valuable signal for identifying ambiguous, challenging, or complex instances that are also difficult for human annotators.

## 5 Discussion and Conclusion

We have demonstrated how disagreement, both between modalities and between humans and models, can serve as a diagnostic lens to understand the complexity of multimodal empathy detection, challenging the assumption that more signals from other modalities reliably yields better performance. Our analysis reveals that disagreement between unimodal and multimodal models is often not arbitrary, but instead marks the presence of subtle, ambiguous, or context-sensitive cues that challenge fusion models and human annotators alike.

While our study focuses on speaker-centric empathy (evaluating speakers' empathic expression), our diagnostic can be generalized to listener-centric tasks, which dominate existing empathy datasets and capture listeners' emotional responses to each utterance (Appendix A.2). These findings emphasize the necessity for high-quality annotation in socially complex tasks like empathy detection, where model errors may reflect genuine human uncertainty or disagreement. This framework provides a scalable method for identifying ambiguity and enhancing model reliability, especially in recognizing complex emotional states.

Beyond diagnosis, disagreement offers a foundation for improving multimodal learning. Cross-modal conflict can guide labeling efforts toward informative and ambiguous examples, making annotation more efficient when resources are limited. Patterns of disagreement can also inform curriculum design (Qian et al., 2025), where models first learn from consistent, low-disagreement examples before tackling more ambiguous ones to build nuanced reasoning and robustness. Furthermore, insights from disagreement can inspire more adaptive fusion approaches that dynamically re-weight or downplay misleading modalities when they conflict (Huang et al., 2023), reducing over-reliance on a single signal. High-disagreement examples can also serve as realistic adversarial test cases that expose systematic vulnerabilities and strengthen fusion strategies under genuine multimodal conflict (Yang et al., 2022). Finally, this diagnostic perspective extends beyond empathy detection to other socially complex tasks such as persuasion (Bai et al., 2021), rapport (Baihaqi et al., 2024), or sarcasm (Zhou et al., 2024), where multimodal cues and subjective judgments often diverge. In such settings, disagreement between unimodal and fusion models highlights genuinely ambiguous cases that can guide targeted annotation, evaluation, and model refinement.

Ultimately, treating disagreement as a meaningful signal rather than an error reframes how we evaluate and improve multimodal models. By revealing when and why models diverge, this perspective lays the foundation for building systems that reason more like humans do. Beyond empathy detection, this framework also opens broader pathways toward socially intelligent multimodal systems that can recognize uncertainty, resolve conflicting evidence, and adapt their reasoning to the inherent ambiguity of human affective communication.

## Limitations

We acknowledge several limitations in our study. Our analyses are based on a limited dataset and a small number of human annotators. Given that empathy is inherently subjective, annotations may vary due to individual interpretations, potentially introducing biases rather than reflecting universal properties of the data. Additionally, we rely upon a single dataset, and future work should investigate whether the patterns we observe hold across other datasets and domains.

Our data is also derived from U.S.-based, English-language television and interview content. As such, the generalizability of our findings to multilingual or culturally diverse settings may be limited. Future research should investigate these patterns in varied cultural and linguistic environments to better assess the broader applicability of our conclusions.

## Ethics Statement

We used a publicly available dataset and strictly use open-source models for analysis.

All annotations were conducted by an author and an individual affiliated with the research team. No participants were recruited via crowdsourcing or external platforms, and no monetary compensation was provided, as the annotators were contributing in a research capacity. We provided detailed information on what we asked the annotators to annotate and how we planned to use the data. The annotators willingly agreed to participate with full knowledge of the task. No sensitive or identifying information was collected from annotators.

We note that empathy expression may vary across cultures, and our findings may not generalize to non-English or non-Western contexts. We encourage future work to explore these questions in more diverse settings.

We will release all code and experimental resources at `https://github.com/mayasrikm/multimodal-empathy-disagreement` to support reproducibility.

## Acknowledgments

This work was supported in part by a grant from the Columbia Center of AI Technology (CAIT) in collaboration with Amazon. The views, opinions and/or findings expressed are solely those of the authors.

## References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and V. S. Subrahmanian. 2021. M2p2: Multimodal persuasion prediction using adaptive fusion.

Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, and Koichiro Yoshino. 2024. Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Pablo Barros, Nikhil Churamani, Angelica Lim, and Stefan Wermter. 2019. The omg-empathy dataset: Evaluating the impact of affective behavior in storytelling.

Roy F. Baumeister and Kathleen D. Vohs. 2007. *Encyclopedia of Social Psychology*, volume 1. Sage.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Paul Boersma and David Weenink. 1992–2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.2.14, retrieved 24 May 2022.

Haozhe Chen, Run Chen, and Julia Hirschberg. 2024a. EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8170–8180, Miami, Florida, USA. Association for Computational Linguistics.

Run Chen, Haozhe Chen, Anushka Kulkarni, Eleanor Lin, Linda Pang, Divya Tadimeti, Jun Shin, and Julia Hirschberg. 2024b. Detecting empathy in speech. In *Proceedings of Interspeech 2024*, Dublin, Ireland. ISCA.

Run Chen, Jun Shin, and Julia Hirschberg. 2025. Synthempathy: A scalable empathy corpus generated using llms without any crowdsourcing.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Wallace V. Friesen and Paul Ekman. 1978. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press.

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. *arXiv preprint arXiv:1605.04072*.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024. Emmi: Empathic multimodal motivational interviews dataset: Analyses and annotations. *arXiv preprint arXiv:2406.16478*.

Md Rakibul Hasan, Md Zakir Hossain, Shreya Ghosh, Aneesh Krishna, and Tom Gedeon. 2023. Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods. *arXiv preprint arXiv:2311.00721*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Judith Holler and Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Maochun Huang, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. 2023. Context-based adaptive multimodal fusion network for continuous frame-level sentiment prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3468–3477.

Summaira Jabeen, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Abdul Jabbar. 2021. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087*.

Yannick Jadoul, Bart de Boer, and Peter Thompson. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Chengxuan Qian, Kai Han, Jiaxin Liu, Zhenlong Yuan, Zhengzhong Zhu, Jingchao Wang, Chongwen Lyu, Jun Chen, and Zhe Liu. 2025. Dyncim: Dynamic curriculum for imbalanced multimodal learning.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Swarnadeep Saha, Peter Hase, Nazneen Rajani, and Mohit Bansal. 2022. Are hard examples also harder to explain? a study with human and model-generated explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2121–2131, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tim Sainburg, Leland McInnes, and Timothy Q Gentner. 2021. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.

Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Park. 2024. EmpathicStories++: A multimodal dataset for empathy towards personal experiences. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4525–4536, Bangkok, Thailand. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Junpeng Wang, Liang Wang, Yan Zheng, Chin-Chia Michael Yeh, Shubham Jain, and Wei Zhang. 2023. Learning-From-Disagreement: A Model Comparison and Visual Analytics Framework . *IEEE Transactions on Visualization & Computer Graphics*, 29(09):3809–3825.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. 2022. Defending multimodal fusion models against single-source adversaries.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. STICKERCONV: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733, Bangkok, Thailand. Association for Computational Linguistics.

Bingzhe Zhou, Hannan Wang, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. 2024. Simulate, refine and integrate: Strategy synthesis for efficient smt solving. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 7976–7984. International Joint Conferences on Artificial Intelligence Organization.

Zhou'an Zhu, Xin Li, Jicai Pan, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal dataset for empathic dialogue in counseling. *arXiv preprint arXiv:2305.14221*.

# A Dataset

## A.1 Dataset Specifications

We use a multimodal empathy dataset (Chen et al., 2024b) consisting of 346 English-language videos totaling approximately 53 hours, collected from YouTube between 2020 and 2022 using keywords like "empathy" and "empathetic training." The dataset includes empathy training sessions, therapy roleplays, interviews, TED Talks, and TV/movie scenes, comprising both acted (62%) and spontaneous (38%) speech. Each video was labeled by at least three expert annotators as either empathetic or neutral, with final labels determined by majority vote. Metadata such as speaker gender, topic, and emotional context was manually annotated, covering themes like therapy, parenting, workplace dynamics, and social relationships. From this collection, a subset of 65 videos was transcribed, diarized, and manually re-aligned using Praat to ensure accurate speaker segmentation and time alignment. This process resulted in 1,718 annotated segments with speaker labels, timestamps, transcripts, and empathy stage annotations, enabling fine-grained analysis of empathy in naturalistic and semi-scripted settings. The median utterance length of the dataset is six tokens. Table 8 shows example utterances below and above the median; examples below the median often include short phrases and backchannels, while examples above the median are often complete sentences with richer lexical and syntactic structure.

## A.2 Dataset Comparison

To the best of our knowledge, our work is the first to evaluate multimodal disagreement on speaker-centric empathy detection datasets. Most publicly available empathy datasets (such as EMPATHICSTORIES++ (Shen et al., 2024) and OMG-EMPATHY (Barros et al., 2019)) are fundamentally structured around listener response, not speaker expression. In these datasets, the task is to predict how empathetic a listener feels after hearing a story, rather than to assess whether the speaker themselves is expressing empathy. For instance, in EMPATHICSTORIES++, participants record personal stories and then rate their own emotional responses, framing empathy as a *reaction* to the content rather than as a property of the speaker's delivery. Similarly, OMG-EMPATHY evaluates listener self-reported affective states following brief monologues, again focusing on perceived empathy rather than expressed empathy. This distinction matters because listener-focused tasks inherently entangle speaker behavior with listener subjectivity, making it difficult to isolate which cues (textual, audial, or visual) are directly responsible for empathy expression. In contrast, the dataset we use in this study (Chen et al., 2024b) is one of the only accessible resources that explicitly asks annotators to evaluate the *speaker's* empathy, based solely on the speech segment itself, across modalities. This framing allows us to analyze how empathy is expressed in real time by the speaker, independent of listener interpretation, and enables direct comparisons between modalities on their ability to convey empathetic intent. Our current dataset offers a uniquely valuable lens into the structure of empathy as a speaker-side communicative behavior: something that remains underexplored in the literature. In future work, our modality-disagreement diagnostic could be used to flag nuanced, high-ambiguity segments that challenge *listener* empathy models. They could serve as an effective proxy for identifying segments that elicit high listener variance in empathy judgments, enabling targeted annotation and model refinement on exactly those ambiguous utterances where listener-centric prediction systems struggle most.

# B Model Training Details

Data was split into train, test and validation sets using random sampling, with an 80-10-10 split. We run fine-tuning and inference for all open-source models on an A100 GPU in Google Colab.

## B.1 Unimodal Model Training Details

Each model is trained on a binary empathy classification task using precomputed 768-dimensional embeddings. We freeze all but the final two transformer layers and train for fifteen epochs with a learning rate of 5e-6 and batch size of eight.

## B.2 Fusion Model Details

Each unimodal model representation is independently gated and passed through an additive attention mechanism that computes modality-specific weights. The weighted embeddings are aggregated and classified using a three-layer feedforward network with max pooling. The fusion model is trained for ten epochs using a learning rate of 1e-4 and includes modality dropout during training. To characterize how the model balances each of the three modalities at inference time, we computed the

| < 6 Tokens | ≥ 6 Tokens |
|---|---|
| *“there’s no way”* (3 tokens) | *“No, no he’s a good guy go easy on him he’s lost his son, Fabio”* (15 tokens) |
| *“You lost it?”* (3 tokens) | *“You kids have the biggest hearts I’ve ever seen.”* (9 tokens) |
| *“I can understand that.”* (4 tokens) | *“congrats my dude, on everything man”* (6 tokens) |

Table 8: Example utterances with fewer than six tokens (left) versus at least six tokens (right).

per-modality gate-weight distributions over the full test set (Table 9). The mean gate weights indicate that our model allocates substantial importance to each modality, with only a slight preference toward audio and video. The high variance also shows that the model dynamically adapts its reliance on each modality on a per-sample basis. Thus, our fusion model draws substantially on all three streams; no single modality is systematically favored.

| Modality | Mean | Standard Deviation |
|---|---|---|
| **Text** | 0.430 | 0.297 |
| **Audio** | 0.502 | 0.227 |
| **Video** | 0.476 | 0.315 |

Table 9: Per-modality gate-weight distributions over the full test set.

## C Annotation Instructions

We employed two annotators, one of the paper’s authors and an non-author, both fluent English speakers based in the United States. No additional demographic information was collected, as the annotation was conducted internally for research purposes.

Annotators were asked to provide two judgments per example, labeling each as either empathetic or neutral (Figure 4). A excerpt describing empathy (drawn from the Encyclopedia of Social Psychology, Volume 1, (Baumeister and Vohs, 2007)) was provided to ensure a consistent conceptual foundation for annotation:

> Empathy is often defined as understanding another person’s experience by imagining oneself in that other person’s situation: One understands the other person’s experience as if it were being experienced by the self, but without the self actually experiencing it. There are three commonly studied components of emotional empathy. The first is feeling the same emotion as another person (sometimes attributed to emotional contagion, e.g., unconsciously “catching” someone else’s tears and feeling sad oneself). The second component, personal distress, refers to one’s own feelings of distress in response to perceiving another’s plight. The third emotional component, feeling compassion for another person, is the one most frequently associated with the study of empathy. Cognitive empathy refers to the extent to which we perceive or have evidence that we have successfully guessed someone else’s thoughts and feelings.

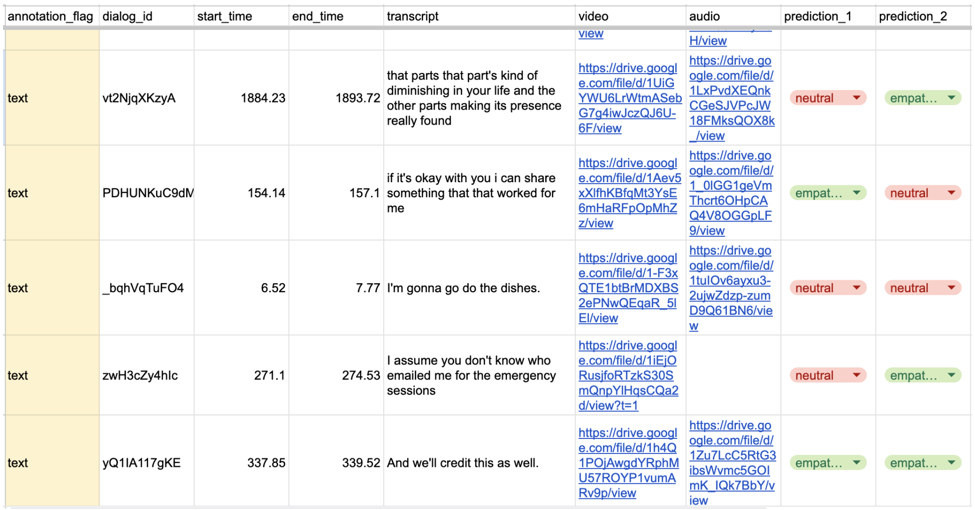
Annotators were given an annotation flag indicating which modality to use for the first pass; for instance, if the flag was text, only the transcript was to be used to make the first prediction. After submitting the first judgment, annotators were then given access to the full video, including all available audio, visual, and textual information. They were then asked to provide a second prediction.

## D Full Feature Comparisons

Tables 10, 11 and 12 provide additional results from the t-tests comparing examples across different confidence quadrants. Table 10 provides an internal comparison between the disagreement quadrants. Table 11 presents the full version of the audio feature comparisons summarized in Table 3. Table 12 expands on the facial feature comparisons shown in Table 4.

## E Feature Distributions

Figures 5 and 6 visualize the distributions of key features across confidence quadrants. Figure 5 presents the distribution of selected audio features (e.g., pitch, intensity) for red, green, and blue examples, highlighting acoustic patterns associated with model disagreement. Figure 6 shows activation rates for facial Action Units (AUs) in red, green, and blue examples, illustrating how specific facial expressions vary across agreement conditions. These visualizations complement the statistical comparisons reported in Tables 11 and 12, providing a more interpretable view of the underlying feature dynamics.



| annotation_flag | dialog_id | start_time | end_time | transcript | video | audio | prediction_1 | prediction_2 |
|---|---|---|---|---|---|---|---|---|
| | | | | | view | H/view | | |
| text | vt2NjqXKzyA | 1884.23 | 1893.72 | that parts that part's kind of diminishing in your life and the other parts making its presence really found | https://drive.google.com/file/d/1UiGYWU6LrWtmASebG7g4iwJczQJ6U-6F/view | https://drive.google.com/file/d/1LxPvdXEQnkCGeSJVPcJW18FMksQOX8k_/view | neutral ▾ | empat… ▾ |
| text | PDHUNKuC9dM | 154.14 | 157.1 | if it's okay with you i can share something that that worked for me | https://drive.google.com/file/d/1Aev5xXlfhKBfqMt3YsE6mHaRFpOpMhZz/view | https://drive.google.com/file/d/1_0lGG1geVmThcrt6OHpCAQ4V8OGGpLF9/view | empat… ▾ | neutral ▾ |
| text | _bqhVqTuFO4 | 6.52 | 7.77 | I'm gonna go do the dishes. | https://drive.google.com/file/d/1-F3xQTE1btBrMDXBS2ePNwQEqaR_5lEl/view | https://drive.google.com/file/d/1tulOv6ayxu3-2ujwZdzp-zumD9Q61BN6/view | neutral ▾ | neutral ▾ |
| text | zwH3cZy4hlc | 271.1 | 274.53 | I assume you don't know who emailed me for the emergency sessions | https://drive.google.com/file/d/1iEjORusjfoRTzkS30SmQnpYlHqsCQa2d/view?t=1 | | neutral ▾ | empat… ▾ |
| text | yQ1IA117gKE | 337.85 | 339.52 | And we'll credit this as well. | https://drive.google.com/file/d/1h4Q1POjAwgdYRphMU57ROYP1vumARv9p/view | https://drive.google.com/file/d/1Zu7LcC5RtG3ibsWvmc5GOImK_IQk7BbY/view | empat… ▾ | empat… ▾ |

Figure 4: Annotation interface

| Feature | t-stat | p-value | Mean Comparison |
|---|---|---|---|
| **Mean Pitch** | **2.453** | **0.0159** | $\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$ |
| **Max Intensity** | **-2.124** | **0.0366** | $\boldsymbol{\mu_{\text{green}} > \mu_{\text{red}}}$ |
| **Max Pitch** | **2.016** | **0.0465** | $\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$ |
| **Min Pitch** | **2.007** | **0.0475** | $\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$ |
| valence | -1.908 | 0.0593 | $\mu_{\text{green}} > \mu_{\text{red}}$ |
| arousal | 1.827 | 0.0705 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| speaking_rate | 1.773 | 0.0807 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| dominance | 1.712 | 0.0899 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| Shimmer | 0.773 | 0.4416 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| Jitter | 0.622 | 0.5355 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| Mean Intensity | 0.544 | 0.5886 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| HNR | 0.508 | 0.6129 | $\mu_{\text{red}} > \mu_{\text{green}}$ |
| Min Intensity | -0.429 | 0.6685 | $\mu_{\text{green}} > \mu_{\text{red}}$ |

Table 10: T-test results comparing audio features between red and green examples. Statistically significant results are bolded.

## F Quadrant Examples

Table 13 shows video frames, transcripts, annotator judgments, and the true labels for examples from each confidence plot quadrant. These examples illustrate that disagreement quadrants often contain more ambiguous instances for both humans and models where cues from different modalities may conflict, while examples from agreement quadrants typically display alignment between modalities.

| Feature | p (Red vs Blue) | Direction | p (Green vs Blue) | Direction |
|---|---|---|---|---|
| **valence** | **0.0047** | **$\mu_{\text{blue}} > \mu_{\text{red}}$** | 0.5166 | $\mu_{\text{green}} > \mu_{\text{blue}}$ |
| **arousal** | **0.0065** | **$\mu_{\text{blue}} > \mu_{\text{red}}$** | **0.0136** | **$\mu_{\text{blue}} > \mu_{\text{green}}$** |
| **Mean Pitch** | **0.0100** | **$\mu_{\text{blue}} > \mu_{\text{red}}$** | **0.0001** | **$\mu_{\text{blue}} > \mu_{\text{green}}$** |
| **dominance** | **0.0108** | **$\mu_{\text{blue}} > \mu_{\text{red}}$** | 0.0667 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| **Min Pitch** | **0.0333** | **$\mu_{\text{blue}} > \mu_{\text{red}}$** | **0.0001** | **$\mu_{\text{blue}} > \mu_{\text{green}}$** |
| **Jitter** | **0.0347** | **$\mu_{\text{red}} > \mu_{\text{blue}}$** | 0.0667 | $\mu_{\text{green}} > \mu_{\text{blue}}$ |
| Max Intensity | 0.1260 | $\mu_{\text{red}} > \mu_{\text{blue}}$ | **0.0023** | **$\mu_{\text{green}} > \mu_{\text{blue}}$** |
| Mean Intensity | 0.1599 | $\mu_{\text{red}} > \mu_{\text{blue}}$ | 0.5329 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| HNR | 0.2217 | $\mu_{\text{blue}} > \mu_{\text{red}}$ | 0.2055 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| speaking_rate | 0.2723 | $\mu_{\text{blue}} > \mu_{\text{red}}$ | 0.9991 | $\mu_{\text{green}} > \mu_{\text{blue}}$ |
| Shimmer | 0.4122 | $\mu_{\text{red}} > \mu_{\text{blue}}$ | 0.1541 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| Max Pitch | 0.6845 | $\mu_{\text{red}} > \mu_{\text{blue}}$ | 0.2647 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |
| Min Intensity | 0.7999 | $\mu_{\text{blue}} > \mu_{\text{red}}$ | 0.1571 | $\mu_{\text{blue}} > \mu_{\text{green}}$ |

Table 11: T-test results comparing audio features between red vs. blue and green vs. blue examples. Statistically significant p-values are bolded.

| AU | p (Red vs Blue) | Direction | p (Green vs Blue) | Direction |
|---|---|---|---|---|
| **AU04: Brow Lowerer** | **0.0106** | **red > blue** | 0.3682 | green > blue |
| **AU12: Lip Corner Puller** | **0.0174** | **blue > red** | 0.8977 | green > blue |
| **AU05: Upper Lid Raiser** | 0.1837 | blue > red | **<0.0001** | **blue > green** |
| AU17: Chin Raiser | 0.2256 | red > blue | 0.9802 | blue > green |
| AU10: Upper Lip Raiser | 0.2275 | blue > red | 0.6700 | green > blue |
| AU45: Blink | 0.3200 | blue > red | 0.7462 | green > blue |
| AU07: Lid Tightener | 0.3252 | blue > red | 0.9318 | blue > green |
| AU14: Dimpler | 0.4593 | red > blue | 0.0652 | green > blue |
| AU20: Lip Stretcher | 0.5701 | blue > red | 0.7907 | blue > green |
| AU09: Nose Wrinkler | 0.6211 | blue > red | 0.7639 | green > blue |
| AU25: Lips Part | 0.6227 | blue > red | 0.7492 | blue > green |
| AU01: Inner Brow Raiser | 0.6529 | blue > red | 0.4674 | green > blue |
| AU23: Lip Tightener | 0.6630 | red > blue | 0.3474 | green > blue |
| AU28: Lip Suck | 0.6735 | red > blue | 0.9846 | green > blue |
| AU26: Jaw Drop | 0.6851 | red > blue | 0.4596 | blue > green |
| AU06: Cheek Raiser | 0.7097 | blue > red | 0.3201 | green > blue |
| AU15: Lip Corner Depressor | 0.9528 | red > blue | 0.4834 | green > blue |
| AU02: Outer Brow Raiser | 0.9647 | blue > red | 0.6677 | green > blue |

Table 12: T-test results comparing AU activation rates between red vs. blue and green vs. blue. Bolded p-values are statistically significant.

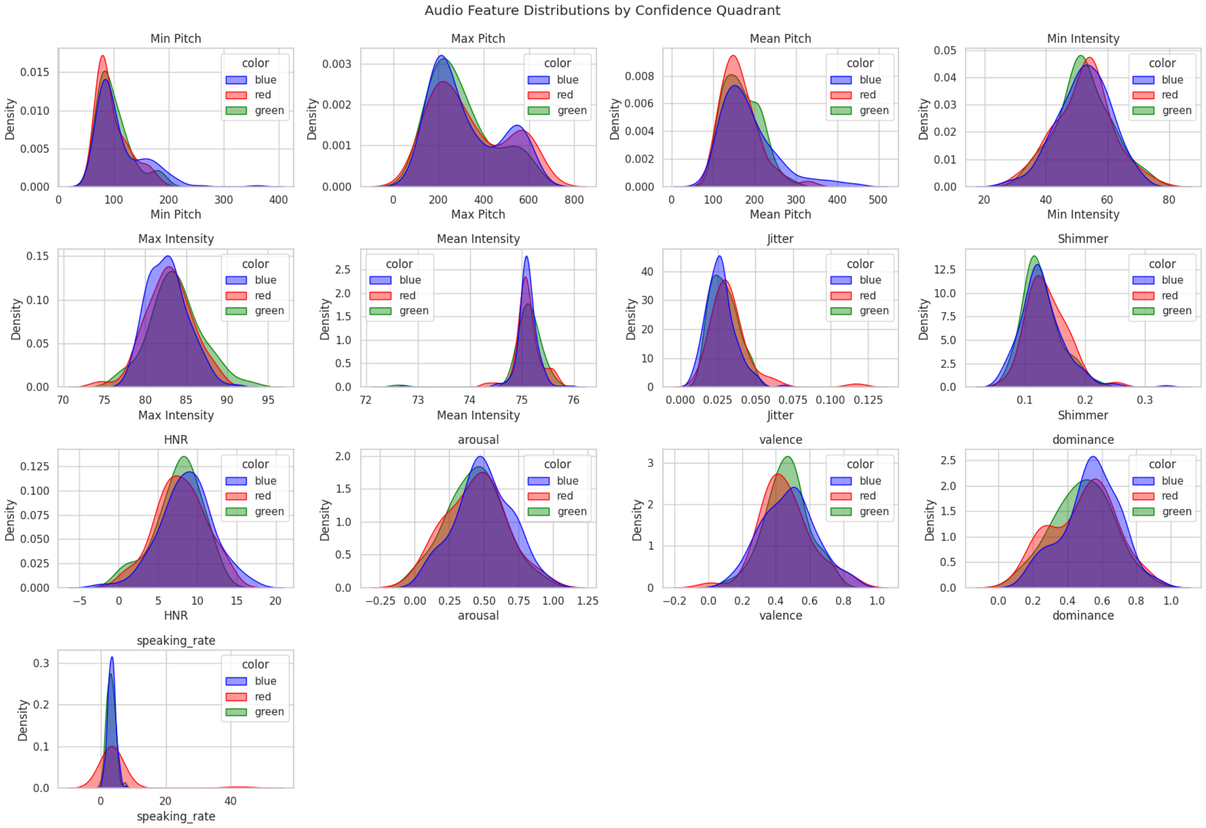


Figure 5: Distribution of audio features for red, green and blue examples across the confidence quadrants. Red examples are those correctly classified by the unimodal audio model but misclassified by the multimodal model; green examples represent the reverse. Blue examples represent those correctly classified by both the unimodal audio model and the multimodal model. Significant differences appear in pitch and intensity-based features.

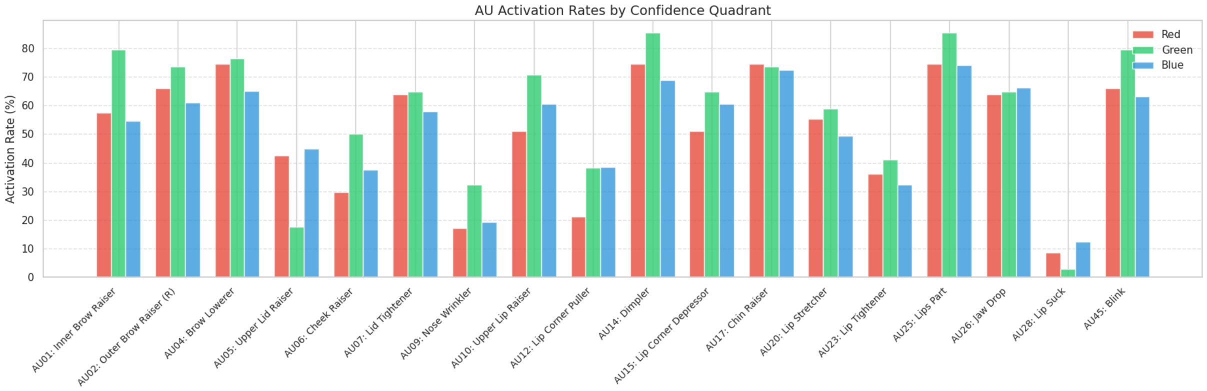


Figure 6: AU activation rates for red, green, and blue examples. Red bars indicate examples where the unimodal visual model predicted correctly but the multimodal model did not (Red: Unimodal $> 0.5$, Multimodal $< 0.5$). Green bars show the reverse. Blue bars indicate examples where both the unimodal and multimodal models correctly predicted the label.

| | |
|---|---|
| <br>**Transcript:** "I assume you don't know who emailed me for the emergency sessions"<br>**Quadrant:** Red<br>**True Label:** Empathetic<br>**Annotator 1:** Neutral<br>**Annotator 2:** Empathetic | <br>**Transcript:** "In fact research suggests we spend about 55 percent of our day..."<br>**Quadrant:** Blue<br>**True Label:** Neutral<br>**Annotator 1:** Neutral<br>**Annotator 2:** Neutral |
| <br>**Transcript:** "One of the reasons I wanted to come here tonight was to discuss our future."<br>**Quadrant:** Yellow<br>**True Label:** Neutral<br>**Annotator 1:** Empathetic<br>**Annotator 2:** Empathetic | <br>**Transcript:** "It's good to have you here um especially to talk about a topic that i think is one of the more sensitive topics that we we're discussing in society today..."<br>**Quadrant:** Green<br>**True Label:** Empathetic<br>**Annotator 1:** Neutral<br>**Annotator 2:** Empathetic |

Table 13: Example clips from each disagreement quadrant with transcript and labels.