CAN AUDIO-CAPABLE LANGUAGE MODELS PREDICT HUMAN EMOTION JUDGMENTS OF INFANT VOCALIZATIONS? RAW AUDIO ALONE IS NOT ENOUGH

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-capable large language models can process raw audio, but whether they reproduce fine-grained human emotion judgments for non-linguistic vocalizations is unclear. We compared four strategies on 178 matched transformed infant vocalizations rated for happiness, sadness, anger, fear, and disgust: direct GPT-Audio prediction from raw audio, supervised prediction from engineered acoustic features, and two feature-based in-context learning conditions using acoustically similar labeled examples with GPT-Audio and GPT-4o-mini. Direct GPT-Audio showed little correspondence with human ratings ($r = -0.173$ to 0.016), whereas supervised feature-based models reached $r = 0.883$. Feature-based in-context learning produced positive correspondence for both GPT-Audio ($r = 0.450$ to 0.725) and GPT-4o-mini ($r = 0.573$ to 0.786), although both remained below the strongest supervised models. These findings suggest that structured acoustic representations and labeled demonstrations support model prediction of human affective judgments, while conventional supervised feature-based approaches achieved higher correlations in this study.

1 INTRODUCTION

Human listeners can infer emotion from non-linguistic vocalizations such as laughter, crying, and cooing even in the absence of lexical information (Byrd et al., 2012; Holz et al., 2021). Previous work has shown that these judgments are systematically related to measurable acoustic properties, including temporal, spectral, and timbral characteristics (Bowman & Yamauchi, 2016). This relationship makes non-linguistic vocalizations a useful setting for examining whether artificial intelligence systems can reproduce human affective judgments from acoustic information.

Traditional computational approaches to affect prediction typically combine engineered acoustic features with supervised machine learning models (Bowman & Yamauchi, 2016; Baird et al., 2022). In contrast, recent audio-capable large language models can process raw audio directly, raising the possibility that affective information could be inferred without task-specific prediction models (Deshmukh et al., 2023; Tang et al., 2024; Kong et al., 2024). However, it remains unclear whether general-purpose audio models can recover the acoustic relationships underlying human emotional judgments, or whether explicit acoustic representations remain advantageous.

To investigate this question, we compared four strategies for predicting average human emotion ratings of transformed pre-linguistic infant vocalizations. The first used an audio-capable GPT model to predict ratings directly from raw audio. The second used conventional prediction methods based on engineered acoustic features. The third and fourth used GPT-based in-context learning with GPT-Audio and GPT-4o-mini, in which acoustically similar labeled examples were provided together with the feature representation of a held-out stimulus. By comparing these approaches against the same human ratings, we examined how raw-audio processing, conventional feature-based prediction, and feature-based in-context learning differ in their ability to reproduce human emotional judgments.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related work

Humans can infer affective information from non-linguistic vocalizations such as laughter, crying, screaming, and cooing even when semantic information is absent. Emotional judgments have been linked to measurable acoustic properties including timbre, spectral characteristics, and temporal structure (Holz et al., 2021; Brooks et al., 2023). This relationship has also been demonstrated in transformed infant vocalizations, where acoustic characteristics such as timbre, Mel-frequency cepstral coefficients, attack time, and attack slope were associated with human emotional ratings (Byrd et al., 2012; Bowman & Yamauchi, 2016).

Computational approaches to affect recognition have traditionally used engineered acoustic descriptors together with supervised prediction models. More recent systems increasingly rely on learned audio representations, including self-supervised approaches such as emotion2vec, which learn affect-relevant representations directly from large collections of audio (Baird et al., 2022; Li et al., 2023; Ma et al., 2024). Together, these approaches demonstrate that both engineered and learned acoustic representations can support accurate affective prediction, although they generally rely on models or representations designed specifically for audio or emotion-related tasks.

Audio-capable language models extend general-purpose language models to direct processing of speech and non-speech audio and have shown broad capability across audio understanding tasks (Deshmukh et al., 2023; Tang et al., 2024; Kong et al., 2024). However, general audio capability does not necessarily imply sensitivity to the acoustic cues underlying human emotional judgments. Chen et al. (2026) found that audio-language models often relied heavily on lexical information during emotion recognition, with performance deteriorating when lexical cues were neutral or absent. In-context performance can also depend on the relevance of the demonstrations provided to the model; Liu et al. (2022) showed that retrieving examples similar to a test input can improve in-context prediction. These findings motivate the present comparison of direct raw-audio prediction, conventional acoustic-feature-based prediction, and feature-based in-context learning for non-linguistic vocalizations.

3 Experiment

3.1 Dataset and human ratings

This study utilized the dataset introduced by Byrd et al. (2012), which investigated the relationship between acoustic properties of pre-linguistic infant vocalizations and human emotion perception. The dataset consists of 180 synthetic audio stimuli created from recordings of infant cooing, babbling, crying, laughing, and screaming. These synthetic sounds were designed to emulate novel pre-linguistic vocalizations while preserving acoustic characteristics relevant to emotion perception.

The 180 synthetic stimuli were generated from ten source infant vocalizations representing screaming, crying, laughing, cooing, and babbling. Using SPEAR (https://www.klingbeil.com/spear/), source recordings were decomposed into spectral components, after which pairs of recordings were recombined and adjusted in amplitude and pitch to produce four intended affective variants. Applying this procedure to 45 source-pair combinations yielded 180 synthetic stimuli.

Emotional ratings were collected from 145 undergraduate participants. To reduce participant fatigue, participants were randomly assigned to one of two groups, with each participant rating either 89 or 90 audio stimuli presented in a random order. Each sound was rated on five emotion dimensions: happiness, sadness, anger, fear, and disgust. Ratings were provided on a 7-point Likert scale ranging from 1 ("strongly disagree") to 7 ("strongly agree"), indicating the extent to which each emotion described the presented sound (Byrd et al., 2012).

Although 180 synthetic sounds were created, the original human-rating data contained ratings for 179 stimuli. Exact filename matching between the human-rating records and the available acoustic-feature records yielded 178 complete matches, which were retained for all subsequent analyses.

3.2 ACOUSTIC FEATURE EXTRACTION

Raw audio signals were converted into numerical acoustic features used by both the conventional machine-learning and GPT-based in-context approaches. Unlike direct `gpt-audio`, which operated on waveforms, these methods used the same structured representation, enabling direct comparison between conventional machine learning and GPT-based in-context prediction.

These features were extracted using a Python implementation designed to approximate the acoustic descriptors provided by MIRToolbox (Lartillot & Toiviainen, 2007). The resulting feature matrix contained 40 numeric feature columns spanning temporal/onset, spectral, timbral/harmonic, cepstral, and frequency-band measures. Representative features and the acoustic information captured by each family are summarized in Table 1. Together, these features provide a broad representation of the acoustic properties of each vocalization and include descriptors previously associated with human emotion judgments of non-linguistic musical sounds (Bowman & Yamauchi, 2016).

Table 1: Summary of the acoustic feature families used to represent each audio stimulus. Representative features are shown; the initial extracted feature matrix contained 40 numeric feature columns.

Feature family	**Representative features**	**Acoustic information represented**
Temporal/onset (4)	Attack time, attack slope, zero-crossing rate, envelope centroid	Onset and time-domain signal dynamics
Spectral distribution/shape (9)	Brightness, spectral rolloff, spectral centroid, entropy, spectral flatness, spread, kurtosis, skewness, high-/low-frequency ratio	Frequency distribution and shape of spectral energy
Timbral/ harmonicity (4)	Roughness, fluctuation, regularity, inharmonicity	Timbral texture, periodicity, and harmonic structure
Cepstral (13)	MFCCs 1–13	Compact representation of the spectral envelope
Frequency-band (10)	Sub-band features 1–10	Band-specific spectral content across frequency regions

Each stimulus was represented by one feature vector containing all extracted acoustic descriptors; this matrix served as input to both the traditional machine-learning and GPT-based in-context analyses.

3.3 GPT-AUDIO CONFIGURATIONS

Both GPT-Audio conditions used the `gpt-audio` model identifier (OpenAI, 2026). In the direct condition, the model received the stimulus waveform and rating instructions. In the feature-based condition, it received numerical acoustic descriptors and labeled examples as text; no stimulus waveform was supplied. The API configuration required audio on either the input or output side. Consequently, feature-based requests used text input and text-plus-audio output, and the five numerical predictions were parsed from the generated response transcript. The generated speech communicated the model's answer and provided no stimulus audio information.

3.4 FOUR LEARNING STRATEGIES

Four complementary approaches were evaluated for predicting human emotion ratings of the synthetic vocalization dataset. The first approach used conventional acoustic-feature-based ML methods. The second used an audio-capable GPT model (GPT-Audio) to predict emotion ratings directly from raw audio. The third and fourth used feature-based in-context learning with GPT-Audio and GPT-4o-mini.

3.4.1 GPT-4O-MINI FEATURE-BASED IN-CONTEXT LEARNING

The GPT-4o-mini in-context condition used the same 40 acoustic features, 10-fold split, and 25-neighbor retrieval procedure as the GPT-Audio feature condition. Features were z-standardized using

Table 2: Inputs and prediction procedures for the four learning strategies. All conditions target the same five mean human emotion ratings.

Strategy	Input and task information	Prediction procedure
Supervised models	Acoustic features and training ratings	Regression fitted within training folds; K-NN averages neighboring ratings.
Direct GPT-Audio	Raw waveform and rating instructions	Integer ratings averaged over 59 synthetic participants.
GPT-Audio feature ICL	Acoustic features and 25 labeled training neighbors	One continuous held-out prediction per stimulus; no weight updates.
GPT-4o-mini feature ICL	Acoustic features and 25 labeled training neighbors	One continuous held-out prediction per stimulus; no weight updates.

the complete feature matrix before folds were formed, and the standardized values were used only for Euclidean neighbor retrieval. The original feature values and the average human ratings for the 25 selected training examples were supplied in the prompt. GPT-4o-mini produced one continuous prediction from 1 to 7 for each held-out stimulus at temperature 0.2; no model weights were updated.

3.4.2 ML BASELINES

The acoustic feature vectors described in the previous section served as predictor variables for conventional supervised machine learning models, while the average human ratings for happiness, sadness, anger, fear, and disgust served as target variables. Separate regression models were constructed for each of the five emotion dimensions.

Four regression algorithms were evaluated: Random Forest Regression, XGBoost Regression, Ridge Regression, and Support Vector Regression (SVR) (Breiman, 2001; Chen & Guestrin, 2016; Hoerl & Kennard, 1970; Drucker et al., 1996). Hyper-parameters for all four models were selected using 5-fold cross-validation within each outer training fold, with R^2 used as the model-selection criterion. The selected configuration was then refit on the full outer training fold before predicting the corresponding held-out fold. Random Forest hyper-parameters included the number of trees, maximum tree depth, and minimum sample requirements for splitting and leaf nodes. XGBoost tuning included the number of estimators, tree depth, learning rate, sub-sampling rate, and feature sub-sampling rate. Ridge Regression tuned the regularization parameter α. SVR used a radial basis function (RBF) kernel, with C, γ, and ϵ selected through the inner cross-validation procedure.

Model-specific preprocessing was applied where appropriate. Preprocessing steps that estimated parameters from the data, including missing-value imputation and feature scaling, were fitted using only the current training data. Random Forest, Ridge Regression, and SVR used median imputation, whereas XGBoost used mean imputation. SVR features were standardized using a standard scaler, while Ridge Regression used robust scaling because of extreme feature values. For Ridge Regression, the high-to-low-frequency ratio was log-transformed before cross-validation, and a fluctuation feature was removed when it duplicated the roughness measure. Random Forest and XGBoost did not use feature scaling.

As an additional non-parametric comparison, we evaluated k-nearest-neighbor (K-NN) regression at two neighborhood sizes. We used $k = 25$ to match the number of labeled examples provided to the GPT-based in-context learning approach and additionally evaluated $k = 13$, approximately $\sqrt{178}$, to examine performance with a smaller neighborhood. For each held-out stimulus, the remaining 177 stimuli served as the candidate neighbor pool. Acoustic features were standardized using only these 177 stimuli, and Euclidean distance was used to identify the nearest neighbors. For each neighborhood size, the predicted rating for each emotion was the unweighted mean of the corresponding human ratings of the selected neighbors. This procedure was repeated until each of the 178 stimuli had been predicted once.

3.4.3 GPT-AUDIO

The direct-audio pipeline provided each .wav stimulus to an audio-capable GPT model together with task instructions matching the five emotion dimensions and 7-point rating scale used in the original human study. The model was instructed to assign an integer rating from 1 to 7 for happiness, sadness, anger, fear, and disgust and to return the five ratings in a structured JSON response.

Each complete pass through the source audio files was treated as one synthetic participant. At the beginning of each synthetic participant, a sampling temperature was pseudo-randomly assigned from 0.2 to 1.0 and remained fixed for all stimuli within that pass, providing between-participant variation in model responses.

The final direct-audio evaluation used OpenAI's `gpt-audio` (OpenAI, 2026). Automatic response validation and retry procedures were used to handle malformed or incomplete outputs. A total of 59 completed synthetic participants were retained for analysis. For each stimulus and emotion dimension, `gpt-audio` ratings were averaged across these 59 synthetic participants before comparison with the corresponding average human ratings.

3.4.4 GPT-AUDIO FEATURE-BASED IN-CONTEXT LEARNING

The final prediction approach investigated whether a large language model could infer the relationship between acoustic features and human emotional ratings through in-context learning. Unlike the direct `gpt-audio` approach, this method did not provide audio waveforms to the model. Instead, OpenAI's `gpt-audio` model received numerical acoustic feature values together with labeled examples showing the corresponding average human emotion ratings. The model weights were not modified or retrained; adaptation to the task occurred only through the examples supplied within each prompt (Liu et al., 2022).

Predictions were generated within a 10-fold cross-validation framework. The 178 stimuli were randomly divided into ten approximately equal folds using a fixed random seed of 42. For nearest-neighbor retrieval, acoustic features were z-standardized using statistics computed across the complete feature matrix prior to construction of the cross-validation folds. In each fold, nine folds formed the training pool and the remaining fold served as the held-out test set. Each held-out stimulus was processed individually. Its standardized acoustic feature vector was compared using Euclidean distance only with the standardized feature vectors of stimuli in the training pool, and the 25 nearest training stimuli were selected as in-context examples. Standardization was used only to calculate nearest-neighbor distances; the original, non-standardized acoustic feature values were supplied to GPT in the prompt.

Each of the 25 selected training examples contained an anonymized filename, its acoustic feature values, and its average human ratings for happiness, sadness, anger, fear, and disgust. After these labeled examples, GPT was provided with the anonymized filename and acoustic feature values of the held-out stimulus, but not its human ratings. The model was then instructed to predict continuous ratings from 1 to 7 for each of the five emotion dimensions. Predictions outside the valid 1–7 range were constrained to the nearest valid endpoint. A temperature of 0.2 was used for all GPT-based in-context predictions.

This procedure was repeated separately for every stimulus in the held-out fold and across all ten folds so that each stimulus received one prediction while held out from the pool of labeled examples. Only stimuli from the training folds were eligible for nearest-neighbor selection. Selecting the 25 acoustically closest training examples rather than supplying the entire training set reduced prompt length and API cost while providing demonstrations that were acoustically similar to the stimulus being predicted.

3.4.5 EVALUATION

Evaluation was based on one prediction for each of the 178 matched stimuli. The four supervised models and GPT-based in-context learning were each evaluated using 10-fold held-out prediction, whereas K-NN used leave-one-stimulus-out prediction and direct GPT-Audio required no training split. For the cross-validated and leave-one-out approaches, held-out predictions were combined so that each stimulus contributed one prediction to the final evaluation; direct GPT-Audio predictions were compared directly with the corresponding average human ratings. Performance for each emotion

was summarized using Pearson's correlation coefficient (r) between predicted and average human ratings. Statistical significance was assessed using two-tailed tests, and 95% confidence intervals were calculated using Fisher's z transformation.

4 RESULTS

4.1 ML BASELINES

The supervised acoustic-feature models showed strong correspondence with average human ratings (Figure 1). Random Forest and XGBoost achieved Pearson correlations of 0.728–0.883 across the five emotion dimensions. SVR also produced consistently strong correlations, while Ridge Regression showed more moderate performance. The K-NN baseline with $k = 25$ ranged from $r = 0.543$ for disgust to $r = 0.747$ for sadness. Using $k = 13$ yielded higher correlations for all five emotions, ranging from $r = 0.576$ for disgust to $r = 0.789$ for sadness. These results show that local similarity in the acoustic-feature space captured substantial predictive information even without a language model.

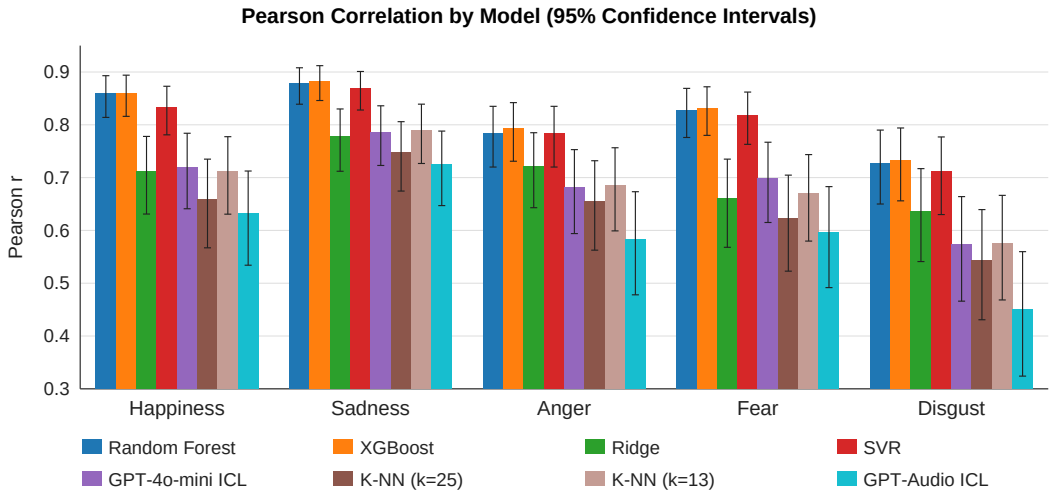

Figure 1: Pearson correlations between predicted and mean human emotion ratings for eight acoustic-feature approaches. Error bars show stimulus-level 95% Fisher-z confidence intervals, not tests of differences between models. K-NN used leave-one-stimulus-out prediction; the other methods used ten-fold prediction.

4.2 GPT-4O-MINI FEATURE-BASED IN-CONTEXT LEARNING

Figure 2 shows GPT-4o-mini predictions against the mean human ratings. Correlations were positive for all five emotions: $r = 0.720$ for happiness, $r = 0.786$ for sadness, $r = 0.682$ for anger, $r = 0.699$ for fear, and $r = 0.573$ for disgust. GPT-4o-mini produced higher correlations than GPT-Audio feature ICL for all five emotions under the evaluated prompting procedures.

4.3 DIRECT GPT-AUDIO PREDICTION

Figure 3 illustrates the relationship between direct `gpt-audio` predictions and average human emotion ratings. Pearson correlations were close to zero or negative for every emotion, ranging from $r = -0.173$ for disgust to $r = 0.016$ for fear (Table 3). The model therefore showed little correspondence with variation in the human ratings under this prompting procedure.

Direct predictions frequently occupied narrow bands, assigning similar ratings to different stimuli. The dashed identity lines in Figure 3 indicate exact agreement; correlation instead measures linear association and does not establish absolute rating accuracy.

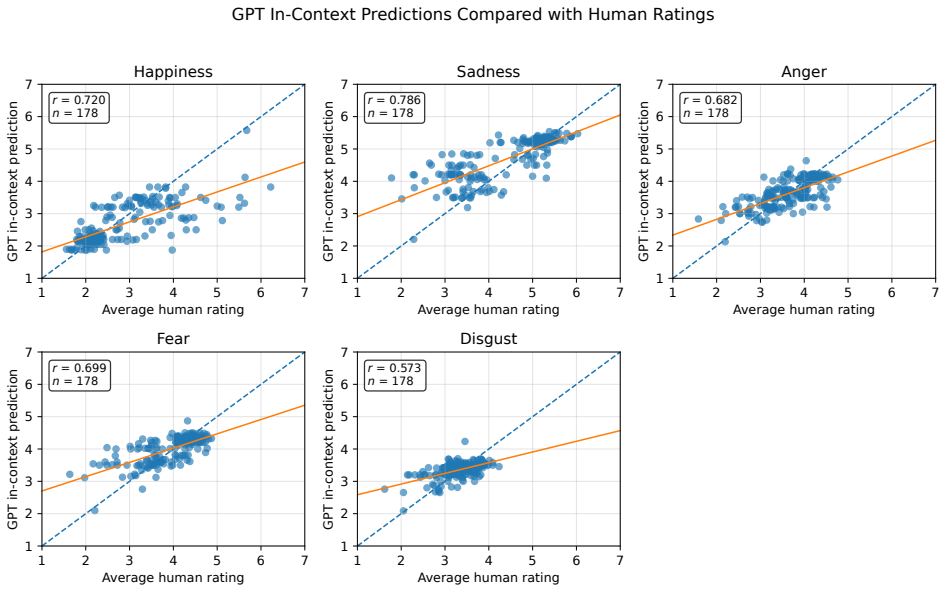

Figure 2: Average human ratings versus GPT-4o-mini feature-based in-context predictions for 178 matched stimuli. Dashed lines indicate perfect agreement, and solid lines show fitted linear regressions.

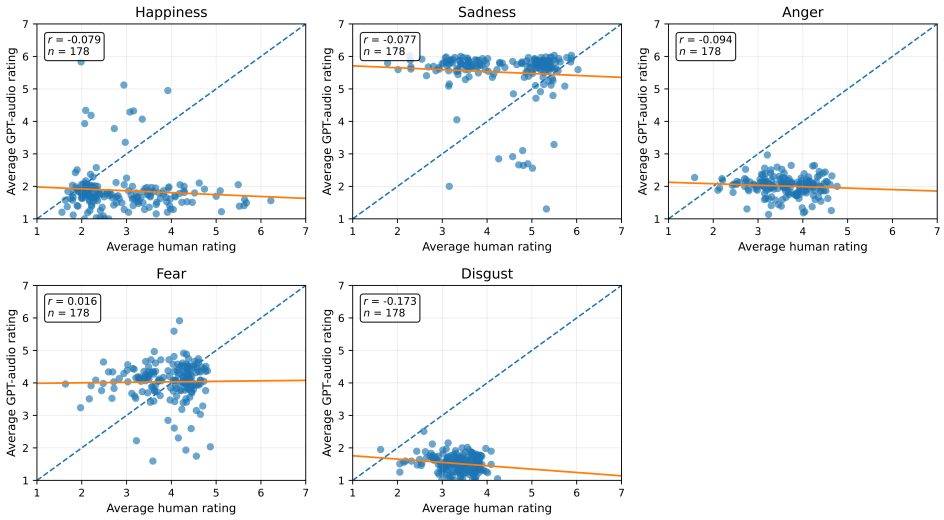

Figure 3: Average human ratings versus direct `gpt-audio` predictions for 178 matched stimuli. Dashed lines indicate perfect agreement, and solid lines show fitted linear regressions.

4.4 GPT-AUDIO FEATURE-BASED IN-CONTEXT LEARNING

Figure 4 shows the per-stimulus predictions against the human ratings. Feature-based GPT-Audio predictions were positively correlated with human ratings for all five emotions, with correlations ranging from $r = 0.450$ for disgust to $r = 0.725$ for sadness. Table 3 reports the emotion-specific estimates and 95% Fisher-z confidence intervals. Each correlation differed from zero at $p < .001$ under the two-tailed stimulus-level test. These tests concern association with human ratings; they do not test differences between prediction strategies.

Both feature-based in-context conditions showed substantially stronger correspondence than direct audio for every emotion. GPT-4o-mini was higher than GPT-Audio feature ICL for all five emotions in these runs, although these values come from separate model runs and are not a paired significance test. GPT-Audio feature ICL fell below the reported Random Forest/XGBoost range, while the GPT-4o-mini range overlapped its lower portion. K-NN also remained a strong comparison, with its reported ranges overlapping and extending above the GPT-Audio feature-ICL range. Because K-NN used leave-one-stimulus-out prediction and GPT-Audio ICL used ten-fold prediction, these comparisons do not isolate the contribution of language-model inference under identical training pools.

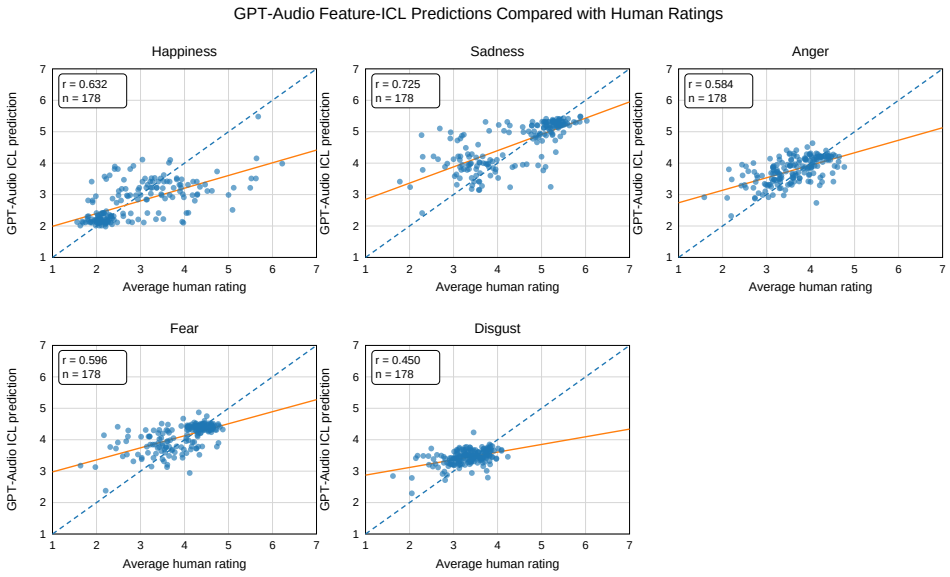

Figure 4: Average human ratings versus GPT-Audio feature-based in-context predictions for 178 matched stimuli. Dashed lines indicate perfect agreement, and solid lines show fitted linear regressions.

Table 3: Pearson correlations between the three GPT conditions and mean human ratings across 178 matched stimuli. The feature-based conditions combine numerical acoustic descriptors with 25 labeled demonstrations. Confidence intervals apply to the GPT-Audio feature-ICL correlations; no confidence intervals for differences between conditions are reported.

Emotion	**Direct audio** r	**GPT-Audio ICL** r	**GPT-4o-mini ICL** r	**GPT-Audio ICL 95% CI**
Happiness	−0.079	0.632	0.720	[0.534, 0.712]
Sadness	−0.077	0.725	0.786	[0.647, 0.788]
Anger	−0.094	0.584	0.682	[0.478, 0.673]
Fear	0.016	0.596	0.699	[0.492, 0.683]
Disgust	−0.173	0.450	0.573	[0.324, 0.560]

5 DISCUSSION

The evaluated models showed substantially different correspondence with human ratings across the four prediction strategies. Direct waveform-based ratings had correlations near zero or below zero, whereas both feature-based in-context conditions were positively correlated with human ratings for every emotion. The GPT-Audio comparison uses the same model identifier for direct and feature-based conditions, while the GPT-4o-mini condition provides a complementary model comparison.

The comparison does not isolate representation alone. The feature-based condition also supplies 25 labeled examples, requests continuous rather than integer ratings, and uses a fixed temperature of 0.2 rather than averaging 59 passes with varying temperatures. The evidence therefore supports a difference between the complete evaluated procedures; it does not establish that acoustic features alone caused the improvement or that raw-audio models cannot learn the task under other forms of supervision.

Conventional feature-based models remained strong. Random Forest and XGBoost produced the highest reported correlations, and K-NN demonstrated that simple local averaging already captured substantial predictive structure. Neither in-context condition required task-specific weight updates, but this procedural advantage did not translate into better correspondence than the strongest supervised models.

The target of prediction is an aggregate perceptual judgment about transformed sounds, rather than an infant's internal emotional state. These results suggest that explicit acoustic representations and labeled examples are useful for this constrained task. Future work should compare strategies under matched data splits and supervision, include raw-audio demonstrations, and test whether hybrid representations improve prediction on additional datasets.

6 LIMITATIONS

The analysis used one small dataset of transformed infant vocalizations, and the findings may not generalize to natural speech, music, environmental sounds, or other affective audio domains. The stimuli were derived from ten source vocalizations; stimulus-level splits do not establish generalization to entirely new source recordings. The GPT results are specific to the evaluated model identifier, API configuration, and prompting procedures.

Feature standardization for ICL retrieval used the complete feature matrix before cross-validation. Held-out ratings were excluded from demonstrations, but held-out feature information contributed to scaling. Future evaluation should fit standardization within each training fold. K-NN used a different evaluation split from the ten-fold methods, and the direct and feature-based GPT-Audio procedures differed in demonstrations, output scale, sampling, and aggregation. These design differences limit causal attribution and strictly matched model comparisons.

Pearson correlation quantifies association rather than calibration or absolute prediction error. Fisher-z intervals and correlation tests treat stimuli as independent and do not account for dependence arising from shared source recordings or overlapping training sets. No paired significance test of differences between strategies is reported.

7 CONCLUSION

Direct GPT-Audio predictions showed little correspondence with human emotion ratings for transformed infant vocalizations. Both GPT-Audio and GPT-4o-mini feature-based in-context learning yielded positive correlations across all five emotions, while the strongest supervised acoustic-feature models achieved higher correlations. The results support evaluating structured acoustic information and labeled demonstrations alongside direct audio processing, with matched experimental controls needed to separate their contributions.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, were used for literature retrieval and discovery, feedback on experimental methodology, implementation and debugging of analysis and data-processing code, support in checking and interpreting results, and assistance with drafting, editing, formatting, and figure preparation. GPT-Audio was separately evaluated as an experimental prediction model, as described in Section 3. Generative AI was not used to generate the underlying transformed-vocalization dataset or human ratings. The authors reviewed and verified AI-assisted code, numerical results, citations, analyses, and manuscript text and take responsibility for the final content of this work.

ETHICS STATEMENT

This study analyzes previously collected aggregate human emotion ratings and transformed infant vocalizations; no new human participants were recruited for the present analyses. The prediction target is listeners' judgments of sounds, and the results do not establish the internal emotional states of infants or validate deployment in clinical or other consequential settings. Release of underlying recordings and ratings is subject to applicable consent, permissions, and redistribution terms.

REPRODUCIBILITY STATEMENT

Section 3 describes stimulus matching, acoustic-feature families, model-specific preprocessing, validation procedures, neighbor retrieval, and GPT prediction settings. It explicitly identifies full-matrix scaling in the ICL pipeline. Section 4 reports the outcome measures, and Section 6 describes evaluation limitations. Analysis materials are planned for release through the Open Science Framework following double-blind review, with data sharing subject to the applicable permissions. An anonymous code and data repository is not currently provided.

REFERENCES

Alice Baird, Panagiotis Tzirakis, Jeffrey A. Brooks, Christopher B. Gregory, Björn Schuller, Anton Batliner, Dacher Keltner, and Alan Cowen. The acii 2022 affective vocal bursts workshop & competition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2022. doi: 10.1109/ACIIW57231.2022.10086002.

Casady Bowman and Takashi Yamauchi. Perceiving categorical emotion in sound: The role of timbre. *Psychomusicology: Music, Mind, and Brain*, 26(1), 2016. doi: 10.1037/pmu0000105.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Jeffrey A. Brooks, Panagiotis Tzirakis, Alice Baird, et al. Deep learning reveals what vocal bursts express in different cultures. *Nature Human Behaviour*, 7:240–250, 2023. doi: 10.1038/s41562-022-01489-2.

Michael Byrd, Casady Bowman, and Takashi Yamauchi. Cooing, crying, and babbling: A link between music and prelinguistic communication. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pp. 1392–1397, 2012.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.274. URL https://aclanthology.org/2026.eacl-long.274/.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18090–18108. Curran Associates, Inc., 2023. doi: 10.52202/075280-0795. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3a2e5889b4bbef997ddb13b55d5acf77-Abstract-Conference.html.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems*, volume 9, pp. 155–161. MIT Press, 1996.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific Reports*, 11:9663, 2021. doi: 10.1038/s41598-021-88431-0.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25125–25148. PMLR, 2024. URL https://proceedings.mlr.press/v235/kong24a.html.

Olivier Lartillot and Petri Toiviainen. MIR in matlab (ii): A toolbox for musical feature extraction from audio. In *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

Jinchao Li, Xixin Wu, Kaitao Song, Dongsheng Li, Xunying Liu, and Helen Meng. A hierarchical regression chain framework for affective vocal burst recognition. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096395.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.931.

OpenAI. Gpt audio model documentation. https://platform.openai.com/docs/models, 2026. Accessed 2026.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.