VocalAffectBench: Evaluating Vocal Emotion Recognition in AI Audio Models

Luc Debaupte, Tyler Baumgartner, Brandon Tai,
Candice Fan, Bill Wang, and Yi Zhong
Besimple AI, San Mateo, CA
{luc, yi}@besimple.ai

Abstract

Voice products increasingly need affective cues that are present in speech but absent from transcripts. We introduce **VocalAffectBench**, a public, test-only benchmark for evaluating whether AI audio models can identify expressed vocal emotion from raw audio. The benchmark contains 273 human-recorded English WAV clips from 51 speaker accounts totaling 1.95 hours across seven labels: `angry`, `disgusted`, `fearful`, `happy`, `neutral`, `sad`, and `surprised`, with 39 clips per class. All baselines are evaluated from audio alone, without transcripts or contextual metadata.

Across six released baselines, average accuracy is 35.5%. The strongest baseline, `gemini_3_5_flash`, reaches 46.5% on the seven-way task, above the 14.3% random baseline but far from robust emotion recognition. A secondary valence-bucket analysis maps labels into positive, neutral, and negative classes, excluding `surprised` because its valence is ambiguous. Aggregate accuracy under this coarser view is 50.9%. Performance is highly uneven across classes. By recall, neutral is identified most reliably at 75.6% averaged across baselines, while surprised and fearful reach only 10.7% and 15.4%, respectively. These results show that the evaluated baselines can extract some affective signal from speech, but discrete expressed-emotion recognition remains fragile, especially for non-neutral emotions that are often most important in voice agent workflows.

Introduction

Voice agents increasingly mediate interactions where tone, pacing, pauses, and intensity can change the meaning of otherwise identical words. A customer who says "that's fine" may be conceding, joking, withdrawing, or signaling frustration. A transcript preserves the phrase, but not the delivery that makes the interaction interpretable.

Automatic speech recognition (ASR) transcripts are often insufficient for this purpose. Text-only sentiment analysis can identify emotionally loaded words, but it answers a different question from speech emotion recognition. Voice agents that apply affect analysis only after transcription may therefore measure lexical sentiment rather than expressed vocal emotion.

Existing affective-speech resources have advanced the field through reusable corpora and shared evaluation tasks (Busso et al., 2008; Livingstone and Russo, 2018; Cao et al., 2014; Schuller et al., 2018). However, model comparisons are often difficult to interpret when task design and output spaces differ. For production teams choosing among hosted audio models, a compact raw-audio benchmark with public predictions and a fixed label set can be more actionable than a large training corpus.

VocalAffectBench addresses this gap with a small, auditable test set for single-label expressed vocal emotion in English speech. The benchmark is test-only, with equal class counts in the released set. Every baseline is scored on the same 273 clips under the same input protocol and label mapping policy, making the comparison easy to inspect and repeat.

Methods

Benchmark Design

VocalAffectBench is a test-only benchmark for expressed vocal emotion recognition. Each item contains a human-recorded English audio clip, a single verified target emotion label, and basic recording metadata. The released audio is 16 kHz mono WAV.

The target is *expressed* vocal emotion. During collection, speakers received script text and performed it under a specific assigned emotion. The script text was written to be compatible with multiple assigned emotions, so the label comes from the requested delivery rather than from emotional words in the script. We retained clips only when the delivered performance matched the assigned label during human review. The benchmark therefore measures whether a model identifies the emotion expressed in the recording, not whether it can infer

Table 1: Dataset composition. Each label has 39 clips.

Label	Clips	Minutes	Mean sec.
`angry`	39	16.7	25.7
`disgusted`	39	18.6	28.5
`fearful`	39	17.2	26.4
`happy`	39	15.6	24.0
`neutral`	39	15.1	23.2
`sad`	39	17.3	26.6
`surprised`	39	16.3	25.1
Total	273	116.8	25.7

what the speaker privately felt.

Collection and Curation

Audio clips were collected from human speakers performing emotional speech between May 15 and June 2, 2026. Collection used a same-script protocol. For each task, speakers received one short generated script and recorded it in three assigned emotions. The script was generated so the same words could plausibly be delivered in each of those emotions rather than lexically forcing one label.

After collection, one reviewer checked each clip for audio quality and whether the expressed delivery matched the requested emotion. Clips were excluded if audio quality was insufficient or if the performed emotion was incorrect or unclear. The reviewed pool available for benchmark construction contained 563 clips from 68 unique speakers.

We constructed the public benchmark by selecting 39 reviewed clips for each of the seven emotion labels. All public clips are single-speaker English recordings. Each clip is a one-person recording with an American-general accent and no background noise. We exported the released audio as 16 kHz mono WAV, with no noise reduction, normalization, or enhancement.

Evaluation Protocol

The six baselines are evaluated under a raw-audio-input protocol. At inference time, prompted audio models receive the audio file and a fixed instruction listing the allowed labels:

> Choose exactly one primary expressed emotion from the allowed label set. Base your answer only on the expressed vocal tone, prosody, pace, intensity, pauses, and wording. Do not infer the speaker's private internal state.

No transcript or contextual metadata is provided to the model. For provider emotion or prosody endpoints that do not accept arbitrary prompts, the evaluation submits the audio file to the endpoint with its default settings. The endpoint returns its native affect labels or scores, which are then mapped to the benchmark label set before scoring. The main evaluation uses a closed seven-label classification target:

```
angry   disgusted fearful   happy
neutral sad       surprised
```

This target set is intentionally conventional rather than exhaustive. The six non-neutral labels are following Ekman's widely used basic-emotion taxonomy (Ekman, 1992). We include `neutral` because prior speech-emotion benchmarks use it as a control or non-emotional class, including RAVDESS (Livingstone and Russo, 2018). It also reflects a practical requirement for deployed voice agents, which often need to separate marked affect from ordinary delivery. This choice does not assume that all emotion is reducible to seven universal categories. It provides a compact, familiar, auditable label space for comparing audio models.

Provider outputs may use native affect labels or scores. For scoring, each returned label is mapped to the seven-label benchmark set using the mapping implemented in the released evaluation script. For example, calm-style outputs map to `neutral`, and anxiety-style outputs map to `fearful`. The released `predictions.csv` preserves both the raw provider output and the normalized `mapped_label`. The full mapping is reported in the appendix.

Metric

The leaderboard reports overall accuracy, the fraction of clips whose mapped prediction matches the reference label. Class-level behavior is summarized with precision and recall. For a class c, let TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}.$$

Because each target class has 39 clips, per-class precision and recall are also reported to expose class-specific failure modes and prediction bias.

For aggregate accuracy, we also report a 95% binomial confidence interval. Each clip is treated as one Bernoulli trial, correct or incorrect. The interval estimates the plausible range for the true accuracy rate given the observed number of correct clips and total scored clips, using a binomial proportion interval with finite-sample correction. These intervals are descriptive. We do not claim statistically significant ordering between models whose intervals overlap substantially.

In addition to strict seven-way accuracy, we report a secondary valence-bucket analysis. The valence view asks whether a model captures a coarse affective direction even when it misses the exact emotion label. This analysis maps `angry`, `disgusted`, `fearful`, and `sad` to negative, `happy` to positive, and `neutral` to neutral, while excluding `surprised`.

Baseline Models

The released benchmark includes six baselines spanning general audio-capable models, speech-specific models, and provider emotion/prosody endpoints. Table 2 reports the aggregate leaderboard.

The strongest released baseline is `gemini_3_5_flash`, with 127 correct predictions out of 273 and 46.5% accuracy (95% CI: 40.7–52.4). The lowest released baseline is `openai_realtime`, with 70 correct predictions and 25.6% accuracy (95% CI: 20.8–31.1). Across all six baselines, 572 correct decisions, for an average accuracy of 34.9%.

Results

Class Difficulty

Performance varies substantially across emotion classes. Table 3 reports aggregate precision and recall by label over the six released baselines.

Precision and recall together help reveal label bias. `neutral` has the highest recall, 75.6%, but the lowest precision, 23.7%, because the baselines predict `neutral` 760 times across 1,638 decisions. That pattern indicates over-prediction of neutral rather than uniformly reliable neutral recognition, while `fearful` shows the opposite pattern, with 70.8% precision but 14.5% recall. Models rarely predict `fearful`, but when they do the prediction is often correct. These patterns show that model choice and safeguards should depend on the emotion classes that matter most in a specific application.

Neutral Bias

The dominant error pattern is over-prediction of `neutral`. Across all scored decisions, models predict `neutral` 760 times, or 46.4% of all outputs. Among incorrect predictions, 583 of 1,066 errors are mapped to `neutral`.

Table 4 shows the most frequent aggregate confusions. All six true non-neutral classes are often collapsed to `neutral`. The largest confusion is `fearful` to `neutral`, followed by `disgusted` to `neutral` and `happy` to `neutral`. This matters for applications because a conservative neutral prediction can make a voice product miss the very affective states that should trigger escalation, empathy, or additional caution.

Valence as a Coarser Target

Some products do not need a seven-way emotion label. They may only need a coarser valence signal. As a secondary descriptive analysis, we map `angry`, `disgusted`, `fearful`, and `sad` to negative, `happy` to positive, and `neutral` to neutral. We exclude `surprised` because its valence is ambiguous.

Under this coarse mapping, aggregate accuracy rises to 52.3%. The best model on this valence view is `gemini_3_5_flash` at 69.2%, followed by `tr_voxtral_small` at 61.3% and `tr_qwen3_5_omni_plus` at 53.7%. This suggests that valence is more tractable than discrete emotion for current models, but still not solved.

Accuracy alone does not show whether a model tends to emit positive, neutral, or negative labels. We therefore also describe each model's output valence skew in Table 5. This should not be read as an intrinsic property of the model. It is a descriptive summary of the labels the model outputs under this benchmark and mapping.

By this definition, three baselines are negative-skewed and three are neutral-skewed. None of the evaluated baselines is positive-skewed. This matters because two models with similar aggregate accuracy can create different product risks if one over-predicts negative affect and another over-predicts neutral delivery.

Discussion

The baseline results show partial but brittle affect recognition. The leading model is roughly three times the random baseline, but still misses more than half of the clips. Average performance across models is only 34.9%. These results make the benchmark useful as a stress test: it exposes the gap between detecting some vocal affect and relying on a seven-way emotion label in production.

The class-level pattern is as important as the aggregate ranking. Models performed best on `neutral`, while `surprised`, `fearful`, and `disgusted` remain difficult for most baselines. This creates a practical risk. A model may appear useful when averaged over a neutral-heavy workload, yet fail on the high-salience states that matter most for escalation or intervention.

The neutral-bias result also cautions against treating speech emotion outputs as facts about users. A neutral prediction can mean that the model did not detect salient affect, but it can also reflect model uncertainty or calibration error. Deployed products should therefore be conservative when affect signals would change material outcomes.

For voice agents, emotion recognition should be treated as a probabilistic auxiliary signal after product-specific validation, not as a sole controller of action. The current baselines are better viewed as measurement targets than as deployable decision mechanisms.

Intended Use

VocalAffectBench is intended for diagnostic evaluation of audio emotion recognition models. Appropriate uses include provider comparison, error analysis, regression tracking, and class-specific product validation.

Table 2: Seven-class VocalAffectBench baseline results. All baselines scored all 273 clips.

Benchmark ID	Provider model	Correct	Acc.	angry	disgust.	fearful	happy	neutral	sad	surpr.
gemini_3_5_flash	gemini-3.5-flash	127	46.5	56.4	28.2	33.3	43.6	74.4	89.7	0.0
hume_prosody	speech_prosody	105	38.5	51.3	5.1	0.0	79.5	84.6	20.5	28.2
tr_qwen3_5_omni_plus	qwen3.5-omni-plus	98	35.9	35.9	25.6	15.4	23.1	64.1	87.2	0.0
tr_voxtral_small	mistralai/voxtral-small-24b-2507	88	32.2	15.4	74.4	17.9	23.1	46.2	33.3	15.4
inworld_voice_profile	inworld/inworld-stt-1	84	30.8	35.9	0.0	12.8	41.0	97.4	23.1	5.1
openai_realtime	gpt-realtime-2	70	25.6	33.3	2.6	7.7	10.3	87.2	35.9	2.6

Table 3: Aggregate per-class precision and recall across six baselines.

Emotion	True	Pred.	Correct	Precision (%)	Recall (%)
neutral	234	760	177	23.3	75.6
sad	234	256	113	44.1	48.3
angry	234	163	89	54.6	38.0
happy	234	206	86	41.7	36.8
disgusted	234	138	53	38.4	22.6
fearful	234	48	34	70.8	14.5
surprised	234	67	20	29.9	8.5

Table 4: Most frequent aggregate confusions across all baselines.

Confusion	Count
fearful → neutral	113
disgusted → neutral	106
happy → neutral	100
sad → neutral	96
surprised → neutral	91
angry → neutral	77
surprised → happy	49
fearful → sad	43

The benchmark is not intended as a training corpus, a hidden leaderboard, a universal measure of emotional intelligence, or a biometric dataset. Since labels and baseline predictions are public, reports should include enough evaluation detail to reproduce the result.

Limitations

VocalAffectBench is intentionally focused. First, it is English-only and all released clips are spoken in an general American accent. Emotion expression, prosody, speech rhythm, and label interpretation vary across languages and cultures. Multilingual and cross-accent evaluation will require additional collection and label validation rather than direct translation.

Second, the benchmark uses a seven-label discrete taxonomy. Human affect is richer than a single class. A primary label is useful for reproducible benchmarking, but it cannot capture every state a listener might perceive. The included valence analysis partly addresses this by showing a coarser alternative, but it is not a replacement for richer affect modeling.

Third, clips were selected to give each target class the same number of examples. This is useful for fair comparison and per-class analysis, but it is not a natural estimate of emotion prevalence in production conversations. Models used in real workloads should be evaluated against their own traffic distribution as well.

Fourth, the dataset contains acted or performed emotional speech. This matches the benchmark target of expressed vocal emotion, but it limits what the results can claim about spontaneous real-world conversations. Results should not be interpreted as evidence that a model can infer what a speaker truly feels.

Fifth, the reference labels are assigned performance targets verified by one reviewer. This provides a practical benchmark signal, but it does not measure population-level listener perception and does not support inter-rater reliability analysis.

Ethical and Privacy Considerations

Released speech can contain acoustic characteristics that may identify or profile speakers, even when the text content is not sensitive. Contributors consented to public release of their recordings for research and benchmarking use. The dataset is nevertheless released for evaluation, not for biometric or high-stakes use.

The benchmark labels expressed performance rather than inner state. This framing should be preserved in downstream reporting. Models evaluated on VocalAffectBench should avoid claims that they can determine a person's inner state or personal risk level from voice alone.

Table 5: Output valence skew based on mapped predictions. Percentages are shares of all 273 model outputs.

Model	Neg. out.	Pos. out.	Neutral out.	Ambig. out.	Output skew
`gemini_3_5_flash`	56.0	9.2	34.8	0.0	Negative
`hume_prosody`	22.3	32.2	35.9	9.5	Neutral
`tr_qwen3_5_omni_plus`	50.5	5.9	42.5	1.1	Negative
`tr_voxtral_small`	55.7	10.3	23.8	10.3	Negative
`inworld_voice_profile`	16.1	16.1	66.7	1.1	Neutral
`openai_realtime`	20.9	1.8	74.7	2.6	Neutral

Data Availability

VocalAffectBench is a public, test-only benchmark. The dataset and evaluation harness use the MIT License, the same release license used for Voice Code Bench. Because the dataset contains human voice recordings, users should still follow the intended-use and ethics constraints in this paper and the dataset card. The dataset repository is available at https://huggingface.co/datasets/besimple-ai/vocal-affect-bench.

The release includes the audio, metadata, predictions, aggregate results, and documentation. The main metadata file is `data/metadata.jsonl`. Baseline predictions are stored in `data/predictions.csv` and `baselines/predictions/`. Aggregate results are stored in `data/leaderboard-summary.csv` and `baselines/results.csv`. The public release does not include source materials or demographic metadata.

Conclusion

VocalAffectBench focuses attention on a practical voice-agent requirement that transcript-based evaluation can obscure: audio models must preserve affective cues carried by vocal delivery, not only the words being spoken. Its contribution is not another training corpus for speech emotion recognition, but a test of whether raw-audio models can recover expressed vocal emotion under a fixed, auditable protocol. By combining controlled label construction, transcript-free evaluation, released predictions, and per-class error analysis, the benchmark connects emotion recognition performance to product risk.

The baseline results show why this distinction matters. Current audio models extract some affective signal, but seven-way expressed-emotion recognition remains brittle; neutral predictions are overused; and several non-neutral classes are frequently missed. Reporting aggregate accuracy alongside per-class precision, recall, confusions, and valence skew therefore gives model developers and application teams a more actionable view of quality: which models detect broad affective signal, which miss high-salience emotions, and where additional validation or safeguards are needed before affect predictions influence user-facing behavior.

References

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. 2014. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Ekman, P. 1992. An Argument for Basic Emotions. *Cognition and Emotion*, 6(3–4):169–200.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Livingstone, S. R., and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. *PLOS ONE*, 13(5):e0196391.

Russell, J. A. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., and Zafeiriou, S. 2018. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats. In *Proceedings of Interspeech 2018*, 122–126.

Appendix: Released File Schema

Each row in `data/metadata.jsonl` is a JSON object with the following fields:

```
{
  "audio_id": "06Eg9dXO99fAAti4HB34",
  "file_name": "audio/06Eg9dXO99fAAti4HB34.wav",
  "required_emotion": "neutral",
  "duration_seconds": 12.4,
```

```
  "sample_rate": 16000,
  "channels": 1
}
```

The main prediction file uses the following CSV columns:

```
audio_id,required_emotion,model_name,provider_model,
predicted_label,mapped_label,confidence,correct,error
```

Appendix: Valence View

Table 6 reports the secondary valence analysis described in Section . The mapping is: `angry`, `disgusted`, `fearful`, and `sad` map to negative. `happy` maps to positive. `neutral` maps to neutral. `surprised` is excluded.

Table 6: Secondary valence accuracy, excluding `surprised`.

Model	Scored	Correct	Accuracy
`gemini_3_5_flash`	234	162	69.2
`tr_voxtral_small`	234	131	56.0
`tr_qwen3_5_omni_plus`	234	124	53.0
`hume_prosody`	234	117	50.0
`inworld_voice_profile`	234	96	41.0
`openai_realtime`	234	84	35.9
Aggregate	1404	714	50.9

Appendix: Exact Prompt

Prompted audio-model baselines use the following instruction:

```
You are evaluating vocal expression in an audio clip.

Choose exactly one primary expressed emotion from this
↪  allowed label set:
[angry, disgusted, fearful, happy, neutral, sad,
↪  surprised]

Use one of those labels verbatim. Do not use synonyms or
↪  any emotion
outside the allowed label set.

Do not infer the speaker's private internal state.
Base your answer only on the expressed vocal tone,
↪  prosody, pace,
intensity, pauses, and wording.

Return only valid JSON:
{
  "primary_emotion": "...",
  "confidence": 0.0,
  "evidence": "brief explanation"
}
```

Appendix: Label Mapping

Provider-native labels and prompted-model synonyms are normalized to the seven-label benchmark set before scoring. The Hume-specific mapping is:

```
aesthetic appreciation -> happy
amusement -> happy
anger -> angry
anxiety -> fearful
boredom -> neutral
calmness -> neutral
concentration -> neutral
confusion -> surprised
contemplation -> sad
contentment -> happy
disappointment -> sad
determination -> angry
disgust -> disgusted
distress -> sad
excitement -> happy
fear -> fearful
horror -> fearful
interest -> surprised
joy -> happy
pride -> happy
realization -> surprised
sadness -> sad
satisfaction -> happy
surprise (negative) -> surprised
surprise (positive) -> surprised
tiredness -> neutral
```

The general alias mapping used for prompted and other provider outputs is:

```
anger/angry -> angry
anxious/anxiety/fear/fearful/scared -> fearful
bored/boredom/calm/calmness/concentration/tender/tiredne
↪  ss ->
↪  neutral
contemplation/sad/sadness -> sad
confused/confusion/interest/realization/surprise/surpris
↪  ed ->
↪  surprised
determination/frustrated/frustration -> angry
disgust/disgusted -> disgusted
excited/excitement/happy/happiness/joy/joyful/pride/sati
↪  sfaction ->
↪  happy
unclear -> unscored
```

Appendix: Citation

```
@misc{vocalaffectbench2026,
  title  = {VocalAffectBench: Evaluating Vocal Emotion
  ↪  Recognition in AI Audio Models},
  author = {Debaupte, Luc and Baumgartner, Tyler and Tai,
  ↪  Brandon and Fan, Candice and Wang, Bill and Zhong,
  ↪  Yi},
  year   = {2026},
  note   = {Benchmark dataset},
  url    = {https://huggingface.co/datasets/besimple-ai/
  ↪  vocal-affect-bench},
  license = {MIT}
}
```