EMONET-VOICE: A Fine-Grained, Expert-Verified Benchmark for Speech Emotion Detection

Christoph Schuhmann*
LAION e.V.
christoph.schuhmann@laion.ai

Robert Kaczmarczyk*
LAION e.V.
Technical University of Munich

Gollam Rabby
L3S Research Center
Leibniz University of Hannover

Felix Friedrich
TU Darmstadt
Hessian.AI

Maurice Kraus
TU Darmstadt

Kourosh Nadi
LAION e.V.

Huu Nguyen
Ontocord
LAION e.V.

Kristian Kersting
TU Darmstadt
Centre for Cognitive Science
Hessian.AI
DFKI

Sören Auer
TIB–Leibniz Information Centre for Science and Technology
L3S Research Center
Leibniz University of Hannover

Abstract

The advancement of text-to-speech and audio generation models necessitates robust benchmarks for evaluating the emotional understanding capabilities of AI systems. Current speech emotion recognition (SER) datasets often exhibit limitations in emotional granularity, privacy concerns, or reliance on acted portrayals. This paper introduces EMONET-VOICE, a new resource for speech emotion detection, which includes EMONET-VOICE BIG, a large-scale pre-training dataset (featuring over 4,500 hours of speech across 11 voices, 40 emotions, and 4 languages), and EMONET-VOICE BENCH, a novel benchmark dataset with human expert annotations. EMONET-VOICE is designed to evaluate SER models on a fine-grained spectrum of 40 emotion categories with different levels of intensities. Leveraging state-of-the-art voice generation, we curated synthetic audio snippets simulating actors portraying scenes designed to evoke specific emotions. Crucially, we conducted rigorous validation by psychology experts who assigned perceived intensity labels. This synthetic, privacy-preserving approach allows for the inclusion of sensitive emotional states often absent in existing datasets. Lastly, we introduce EMPATHICINSIGHT-VOICE models that set a new standard in speech emotion recognition with high agreement with human experts. Our evaluations across the current model landscape exhibit valuable findings, such as high-arousal emotions like anger being much easier to detect than low-arousal states like concentration. [2]

1 Introduction

Synthetic speech technology has reached unprecedented fidelity, with state-of-the-art text-to-speech (TTS) and audio generation models, e.g., GPT-4 OmniAudio [25], achieving prosody, timbre, and expressiveness comparable to humans. These advancements significantly enhance human-computer interaction (HCI), enabling virtual assistants to convey appropriate emotional qualities across diverse

*Contributed equally and jointly supervised this project.
[2]links to our data, models and code

Preprint. Under review.

contexts [14]. However, this advancement remains asymmetric: while machines can to some extent effectively *synthesize* convincing affective speech, they still struggle to *recognize* the nuanced, context-dependent emotional information humans naturally convey [15, 34], a critical capability for truly conversational AI.

Despite steady progress in speech emotion recognition (SER) through deep architectures and self-supervised representations, evaluation remains constrained by datasets predominantly built around a limited set of "basic" emotions [8, 38]. Established benchmarks such as IEMOCAP [3], RAVDESS [19], and CREMA-D [4] have been invaluable for the field but exhibit three fundamental limitations:

(i) **Insufficient Granularity.** Coarse taxonomies fail to capture subtle or compound emotional states (e.g., *bittersweet*, *embarrassment*, *envy*) that are essential for naturalistic interaction [5].
(ii) **Limited Representativeness.** Current datasets predominantly consist of studio-quality acted speech, lacking linguistic diversity and omitting sensitive emotional states due to privacy constraints [20, 33].
(iii) **Restricted Scalability.** Licensing restrictions, privacy concerns, and annotation costs severely limit dataset size, impeding the data-intensive training regimes required by modern deep learning approaches [37, 28], specifically for open-source and -science.

These limitations are further compounded by evolving perspectives in affective science. Constructionist theories, particularly Barrett's *Theory of Constructed Emotion* [1], conceptualize emotions as context-dependent constructions rather than universal biological packages. This perspective aligns with dimensional frameworks such as Russell's valence–arousal circumplex [30] and supports multi-label approaches that treat affect as overlapping estimates rather than discrete categories [22, 23]. Consequently, SER research must evolve along two parallel trajectories: developing richer datasets with evaluation protocols that respect emotional complexity, and creating modeling strategies that go beyond simplistic classification paradigms.

To address these challenges, we introduce two complementary datasets. First, **EMONET-VOICE BIG**, a foundational dataset for pretraining models on SER. It is a comprehensive synthetic voice corpus exceeding 4,500 hours in four languages (English, German, Spanish, French), featuring 11 distinct voices with different gender identities and a fine-grained taxonomy of 40 emotion categories. As such, it provides an open, privacy-compliant foundation for emotional TTS research and multilingual speech analysis at scale. Second, from this corpus we curate **EMONET-VOICE BENCH**, comprising 12,600 audio clips annotated by psychology experts using a strict consensus protocol that evaluates *both* the presence *and* intensity of each target emotion across our 40-category emotion taxonomy. This approach yields a high-quality, multilingual benchmark for fine-grained SER while circumventing the privacy barriers that inhibit the collection of authentic sensitive vocal expressions.

Building on these datasets, we develop **EMPATHICINSIGHT-VOICE** (Small and Large), novel SER models that achieve state-of-the-art performance in fine-grained emotion recognition while demonstrating strong alignment with human expert judgments. Through comprehensive evaluation across the concurrent SER model landscape, we reveal critical insights into current SER capabilities, including systematic patterns in which emotions prove more challenging to recognize (e.g., low-arousal states like concentration versus high-arousal emotions like anger).

In summary, our contributions are four-fold: **(1)** We build **EMONET-VOICE BIG**, a pretraining, open-access, 4,500-hour multilingual synthetic speech corpus featuring 11 distinct synthetic voices across 4 languages and 40 emotion categories. **(2)** We introduce **EMONET-VOICE BENCH**, a meticulously curated and expert-verified benchmark dataset of 12,600 high-quality audio samples for fine-grained SER, featuring 40 emotion categories with 3 intensity levels. **(3)** We build **EMPATHICINSIGHT-VOICE** (Small and Large), novel SER models designed for nuanced emotion estimation. **(4)** We conduct comprehensively evaluations on our novel benchmark, providing critical insights into current SER capabilities and limitations.

2 Related Work

Current SER research operates on a constrained empirical foundation. The field still relies on a small set of *acted* corpora recorded in controlled studio conditions—IEMOCAP (12 h, 9 emotions) [3], RAVDESS (1 h, 8 emotions, speech & song) [19], SAVEE (0.8 h, 7 emotions, four male speakers) [12], the German EMODB [2], and the multi-ethnic CREMA-D [4]. While these corpora

Table 1: **Comparison of SER datasets**. Key aspects include licensing, scale, emotional range, speaker diversity, synthetic origin, and multilingual support. Open license means CC-BY 4.0 or equivalent; var. means varies across pooled corpora.

	Dataset	**Open Licence**	**Size (#Utts/Hours)**	**#Emo.**	**#Spk.**	**Synth.**	**Multilin.**
	IEMOCAP [3]	✗	10k / ∼12h	9	10 (5M/5F)	✗	✗
	RAVDESS [19]	✓	1.4k / ∼1h	8	24 (12M/12F)	✗	✗
	SAVEE [12]	✗	480 / <1h	7	4 (Male)	✗	✗
	EmoDB [2]	✗	535 / <1h	7	10 (5M/5F)	✗	✗
	CREMA-D [4]	✓	7.4k / ∼6h	6	91 (48M/43F)	✗	✗
	SERAB [31]	✗	9 corpora / var.	6	var.	✗	✓
	EmoBox [21]	✗	32 corpora / var.	≤8	var.	✗	✓
	SER Evals [26]	✗	18 corpora / var.	≤8	var.	✗	✓
	BERSt [36]	✓	∼4h	6	98	✗	✗
ours	**EmoNet-Voice Big**	✓	**>1M / >4,500h**	40	**11 (Synth)**	✓	✓
ours	**EmoNet-Voice Bench**	✓	**∼12k / 35.8h**	40	**11 (Synth)**	✓	✓

provide clean labels and high acoustic quality, they share four persistent weaknesses. *First*, they use *restrictive taxonomies*—typically Ekman's six basic emotions [8]—omitting compound or socially nuanced states such as *embarrassment*, *envy*, or *contemplation* [27, 5]. *Second*, their acted prosody exaggerates emotional cues and reduces generalization to spontaneous speech [20]. *Third*, privacy and ethics hinder collection of *intimate* or stigmatizing emotions (e.g. shame, desire, grief) [33]. *Fourth*, scale and linguistic diversity remain limited: most corpora contain < 100 speakers, just a few hours of audio, and are largely English-centric. Recent efforts to expand this foundation include early multilingual sets such as EmoReact and the parallel English–Mandarin ESD, which broaden language coverage but still cap labels at six basic categories [24, 39]. Aggregation benchmarks go further—SERAB pools nine legacy corpora in six languages [31]; EmoBox widens the scope to 32 datasets in 14 languages with turnkey evaluation splits [21]; SER Evals organises 18 minority-language corpora into in- and out-of-domain test beds for robustness analysis [26]; and BERSt collects ≈ 4 h of shouted and distanced English speech from 98 actors at 19 smartphone positions [36]. Yet these resources still inherit the core constraints of their sources: acted or scripted speech, narrow taxonomies (≤ 8 emotions), modest duration per language, and a lack of expert-validated intensity labels or sensitive affective states.

These existing datasets, summarized and contrasted with our contributions in Table 1, highlight a clear gap. While valuable, they are often restricted by licensing, limited in scale (both in total hours and number of utterances), offer a narrow range of emotion categories (typically 9 or fewer), rely on human actors which limits the privacy-preserving access to sensitive emotions, and many lack multilingual support. EmoNet-Voice Big and EmoNet-Voice Bench directly address these shortcomings by providing a large-scale, openly licensed, synthetic, multilingual corpus with a significantly expanded emotion taxonomy.

Taxonomic limitations exacerbate data-scarcity and theoretical gaps. Modern affective science models emotions as context-dependent and graded rather than discrete [1, 18]. Dimensional (valence–arousal–dominance) and multi-label schemes [30, 37] better capture blended affect, yet almost all benchmarks still assign a *single discrete label* per clip. When intensity annotations exist, they typically rely on crowdsourcing and show low agreement [13, 35]. Consequently, the community lacks benchmarks that reflect contemporary understanding of emotion as multidimensional and graded, particularly for sensitive affective states that cannot be ethically collected from human participants.

Expert-validated intensity annotations across multidimensional affective spaces are missing from existing benchmarks, and we fill this critical gap by contributing EmoNet-Voice Bench with 12,600 carefully chosen clips whose emotional *presence* and *intensity* we had annotated by psychology experts, yielding a high-agreement subset. We overcome the taxonomic, scale, and ethical limitations of existing corpora by combining broad multilingual coverage, a 40-category taxonomy grounded in contemporary affective science [6, 1], and privacy-preserving synthetic speech generation, offering the first benchmark that provides *expert-validated* ratings across a multidimensional affective space.

Table 2: Overview of EMONET-VOICE BIG

Category	Hours
Playtime by Language	
English (en)	2,156
German (de)	716
Spanish (es)	888
French (fr)	881
Acting Chal. (en+de)	111
total	**4,752**
English Accent Distribution	
Louisiana	133
Valley Girl	159
British	132
Chinese	126
French	140
German	135
Indian	129
Italian	134
Mexican	131
Russian	134
Spanish	132
Texan	131
Vulgar Street	149
No accent specified	391

Table 3: Overview of EMONET-VOICE BENCH

Category	Value
Number of Clips	
English (en)	6,156 (48.9%)
German (de)	1,886 (15.0%)
Spanish (es)	2,193 (17.4%)
French (fr)	2,365 (18.8%)
Total Clips	**12,600**
Avg. Clip Duration	**10.36 s**
Total Playtime	**36.26 h**

Table 4: Number of voice audios annotated by human experts across batches for EMONET-VOICE BENCH. Mainly samples with at least positive weak agreement (emotion weakly / strongly present annotated by two human experts) were used in a next batch.

Batch	Unique Human Annotators	Annotated Voice Audios
1	2	4,538
2	3	7,719
3	4	343

3 The EMONET-VOICE Suite: Dataset Construction

This section describes how we built the EMONET-VOICE resources, beginning with our emotion taxonomy, followed by the creation of the large-scale dataset EMONET-VOICE BIG, and concluding with the expert-validated EMONET-VOICE BENCH subset used for final evaluation. Lastly, we introduce EMPATHICINSIGHT-VOICE models setting a new standard in SER.

3.1 Emotion Taxonomy

For EMONET-VOICE, we adopt the comprehensive 40-category emotion taxonomy originally developed for EMONET-FACE [32]. The taxonomy includes a diverse set of categories spanning positive emotions (e.g., *Elation*, *Contentment*, *Affection*, *Awe*), negative emotions (e.g., *Distress*, *Sadness*, *Bitterness*, *Contempt*), cognitive states (e.g., *Concentration*, *Confusion*, *Doubt*), physical states (e.g., *Pain*, *Fatigue*), and socially mediated emotions (e.g., *Embarrassment*, *Shame*, *Pride*, *Teasing*). This fine-grained structure enables the evaluation of models beyond binary or basic categorical classification. The full set of 40 emotion categories and their descriptive terms can be found in App.A.1. A comprehensive description of the methodology used to construct the taxonomy, including literature-based extraction and expert-guided refinement, is provided in App.A.4.

3.2 EMONET-VOICE BIG: Building a large-scale synthetic SER Dataset

The foundational dataset, EMONET-VOICE BIG, consists of emotionally expressive speech samples synthesized using the GPT-4 OmniAudio model[3]. An overview of EMONET-VOICE BIG's scale and language distribution is provided in Table 2. Our prompting strategy cast the model as an actor auditioning for a film, tasked with performing texts designed to evoke one of 40 emotion categories (from the taxonomy in Section 3.1). Key prompt elements included directives for strong emotional expression from the outset and naturalistic human speech patterns (e.g., varied rhythm, volume, tone, and appropriate vocal bursts). This aimed to ensure perceptible emotional content and avoid monotonous delivery. Audio was generated as 3- to 30-second, 24kHz WAV files, utilizing 11 synthetic voices (6 female/5 male) across English, German, French, and Spanish to build a diverse

[3]Accessed via the HyperLab API.

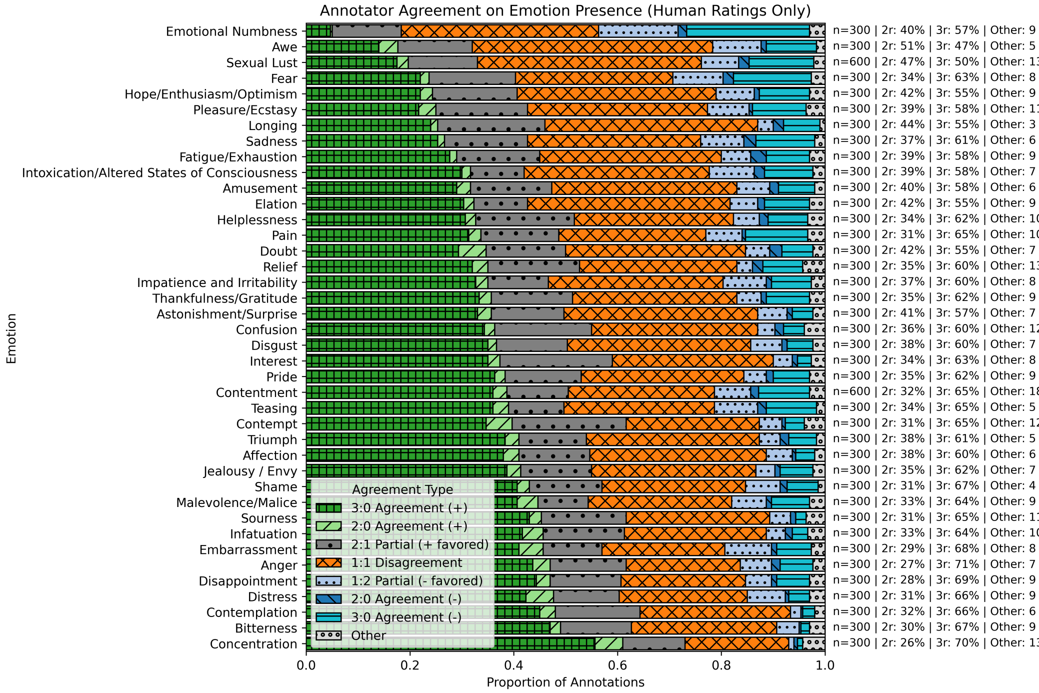

Figure 1: Annotator agreement for human ratings on perceived emotions in audio samples. Stacked horizontal bars display the proportion of audio-emotion instances for each emotion, categorized by agreement type. These categories include full agreement on emotion presence (e.g., '3:0 (+)', '2:0 (+)'), partial agreement where presence is favored (e.g., '2:1 (+ favored)'), disagreement (e.g., '1:1'), partial agreement where absence is favored (e.g., '1:2 (- favored)'), and full agreement on emotion absence (e.g., '2:0 (-)', '3:0 (-)'). Instances with other rating configurations are grouped under 'Other'. The numbers to the right of each bar indicate the total number of instances (n) for that emotion, along with the percentage of these instances rated by two (%2r) or three (%3r) annotators; 'Other' denotes instances with four annotators. The annotation process ensured all audio-emotion pairs were initially rated by two annotators. If both these annotators marked an emotion as present (rating > 0), the instance was subsequently rated by a third annotator. Additionally, a random subset of instances received a fourth annotation.

multilingual corpus. The full prompting template and detailed methodology, including the importance of specific instructions and language-specific adaptations for vocal burst generation, are presented in the Supplement.

3.3 EmoNet-Voice Bench: A Human Expert Benchmark for SER

From EmoNet-Voice Bench, we created a subset of 12,600 unique audio files annotated for emotion by human experts on a three-point annotation scale, summarized in Table 3. We depict the annotation platform for our human experts in Appendix Figures 2 and 3. The dataset features 11 distinct synthetic voices (6 female and 5 male) across four languages: English (48.9%), German (15.0%), Spanish (17.4%), and French (18.8%). The average clip duration is 10.36 seconds, resulting in a total playtime of 36.26 hours.

Table 4 summarizes our annotation procedure. Ensuring the quality and reliability of the emotion annotations was a central priority in constructing the EmoNet-Voice Bench. We recruited a team of six human experts with at least a Bachelor's degree in Psychology to serve as benchmark annotators, thereby guaranteeing familiarity with emotional theory and terminology. In total, 33,605 single-emotion labels across 12,600 unique audio samples were contributed — some samples ultimately received more than three annotations. Each audio clip was first labeled independently by two experts who were presented with the audio alongside one specific target emotion category from our taxonomy in addition to a three-point scale: 0 indicating the emotion was not perceived, 1 indicating it was mildly present at low intensity, and 2 indicating it was intensely present and clearly perceptible. If both human experts agreed that the emotion was present (either "weakly present" or "strongly present"), the clip was sent to a third expert for confirmation. Additionally, we randomly selected a subset of clips to receive a third or even a fourth annotation regardless of whether the first two annotators agreed. To reduce potential gender biases in emotional perception, each group assigned per snippet was balanced in gender composition. Importantly, annotators performed their assessments independently and were blinded to the ratings of others.

Figure 1 illustrates inter-annotator agreement patterns across emotion categories, showing the distribution of full agreement, partial agreement, and disagreement for each emotion-audio pair. The numbers alongside each bar indicate total instances and rating distributions across multiple annotators. The analysis reveals clear consensus patterns: emotions like *concentration* and *bitterness* achieve strong expert agreement, while others such as *numbness* and *awe* show notable disagreement even among psychology professionals. The overall inter-rater reliability measured by Cronbach's α is 0.14 (95% CI [0.12, 0.15]), with per-emotion values detailed in Appendix Table 9. While this low α might initially suggest poor reliability, it actually reflects the inherent complexity of fine-grained emotion perception rather than annotation deficiencies. Unlike simpler emotion taxonomies, our 40-category framework captures subtle distinctions that legitimately evoke different interpretations among experts. These patterns demonstrate that while human agreement is robust for many emotions, certain categories naturally elicit diverse interpretations—underscoring the nuanced nature of affective expression in speech. Rather than indicating weak annotation quality, this variability highlights EMONET-VOICE's sensitivity to the inherent complexity of emotional perception. Our annotations thus capture both the challenges and opportunities in modeling authentic emotional diversity at scale.

3.4 EMPATHICINSIGHT-VOICE: Training state-of-the-art SER models

Another contribution of this work, based on the datasets we built, is to establish a novel state-of-the-art speech emotion recognition model.

First linear probing experiments as well as previous works [17, 7] show that the off-the-shelf Whisper encoders [29] are not capable of reflecting on emotions—an essential capability for emotion-aware audio generation and captioning. Specifically, at a fine-grained level, existing TTS models fail to recognize emotions effectively, as we will discuss later. To address this limitation, we continually pre-trained Whisper encoders as the backbone of our EMPATHICINSIGHT-VOICE. Specifically, we leverage EMONET-VOICE BIG as a pretraining dataset and train emotion-experts in two stages. We base our experiments on Whisper-Small to optimize for the performance-efficiency tradeoff.

In the first stage, the Whisper encoder is trained on a combination of EMONET-VOICE BIG and another 4,500 hours of public emotion-related content[4] to develop general emotional acoustic representations. This data was annotated using an iterative process with Gemini Flash 2.0 to obtain emotion scores (0–4 scale) for all audio snippets. In the second stage, we freeze the Whisper encoder and train MLP expert heads—one per emotion dimension—on top of the fixed encoder embeddings. This way, each MLP receives the full voice audio sequence from the Whisper encoder as sequence flattened token embeddings and then regresses a single emotion intensity score. We propose two model sizes to accommodate different performance requirements, namely EMPATHICINSIGHT-VOICE SMALL with 74M paremeter MLP heads and EMPATHICINSIGHT-VOICE LARGE with 148M paremeter MLP heads. We optimize them using mean absolute error (MAE) on the Gemini Flash 2.0–generated emotion scores.

Through this two-stage fine-tuning and dedicated MLP ensemble, EMPATHICINSIGHT-VOICE effectively captures and predicts fine-grained emotional content from speech with high human alignment, as we demonstrate in the following. Further details are outlined in Appendix A.2.

4 Experiments: Do they hear what we hear?

In this section, we evaluate current SER models on our novel benchmark. Before that, we start by introducing our experimental setup.

[4]https://huggingface.co/datasets/mitermix/audiosnippets

Table 5: Performance comparison of audio language models on the EMONET-VOICE BENCH. Models are evaluated against human emotion ratings using correlation metrics (Spearman and Pearson r, higher is better) and error metrics (MAE and RMSE, lower is better). Our EMPATHICINSIGHT-VOICE models demonstrate superior performance across all metrics, with LARGE achieving the highest Pearson correlation and lowest error and refusal rates. Refusal rates indicate the percentage of samples where models declined to provide emotion assessments. Best scores in bold.

	Model	Refusal (↓)	Spearman (↑)	Pearson (↑)	MAE (↓)	RMSE (↓)
	Gemini 2.0 Flash	0.01%	0.355	0.350	3.608	4.453
	Gemini 2.5 Pro	**0.00**%	0.417	0.416	3.008	3.785
	GPT-4o Mini Audio Preview	2.26%	0.326	0.327	3.320	4.124
	GPT-4o Audio Preview 2024-12-17	27.59%	0.337	0.336	3.432	4.247
	Hume Voice	39.16%	0.274	0.231	4.744	5.474
ours	EMPATHICINSIGHT-VOICE SMALL	**0.00**%	**0.418**	0.414	2.997	3.757
ours	EMPATHICINSIGHT-VOICE LARGE	**0.00**%	0.415	**0.421**	**2.995**	**3.756**

Experimental Setup. EMONET-VOICE BENCH assesses a model's proficiency in discerning emotional intensity from audio. To facilitate a nuanced comparison across models, many of which output continuous scores, our primary evaluation employs metrics suited for regression and correlation analysis on a common scale. The 3-level intensity human judgments (0: Not Present, 1: Mildly Present, 2: Intensely Present) are mapped to a 0-10 scale for this evaluation, becoming 0, 5, and 10, respectively. Model predictions are likewise generated or normalized to this 0-10 continuous scale.

We benchmarked general-purpose multimodal models (e.g., Gemini, GPT-4o) via zero-shot prompting, as well as specialized speech models (e.g., Hume Voice). Hume Voice was subject to constraints on input length ($\leq$5s) and taxonomy coverage. Initial experiments with Whisper failed, due to a general lack of emotion understanding, which led to our development of EMPATHICINSIGHT-VOICE, which pair continually pre-trained Whisper encoders with MLP regressors on our EMONET-VOICE dataset.

We report four key metrics: *Mean Absolute Error (MAE)* and *Root Mean Squared Error (RMSE)* to quantify the average magnitude and larger deviations of prediction error on this 0-10 scale. Additionally, *Pearson Correlation (Pearson r)* and *Spearman Rank Correlation (Spearman r)* are used to assess the linear and monotonic agreement, respectively, between model-predicted intensities and human judgments. These metrics collectively provide a comprehensive view of how well models capture both the absolute values and the relative ordering of perceived emotional intensities.

4.1 Evaluating Speech Emotion Recognition Models

Table 5 presents performance across seven models, revealing clear performance tiers. Our EMPATHICINSIGHT-VOICE models achieve state-of-the-art results, with EMPATHICINSIGHT-VOICE LARGE obtaining the highest Pearson correlation (0.421) and lowest error rates (MAE: 2.995, RMSE: 3.756). EMPATHICINSIGHT-VOICE SMALL demonstrates competitive performance with the highest Spearman correlation (0.418). Gemini 2.5 Pro emerges as the strongest foundation model competitor (Pearson r: 0.416, Spearman r: 0.417), while other commercial models show significantly lower correlations and higher error and refusal rates. This shows that current audio models show decent alignment with human expert ratings on speech emotion recognition. Notably, refusal rates vary dramatically across models. While EMPATHICINSIGHT-VOICE models and Gemini variants process all samples (0-0.01% refusal), GPT-4o Audio Preview refuses 27.59% of samples, and Hume Voice refuses 39.16%—reflecting safety constraints around sensitive emotional content, such as intoxication and pleasure/ecstasy.

Overall, this indicates that our specialized AI models can, to some extent, "hear" what humans hear and demonstrate reasonable alignment with human emotion ratings, while several (general-purpose) models struggle in this task. Yet, this recognition capability proves more complex than initially apparent, as we will explore further next.

Emotion-Specific Performance Patterns. Per-emotion analysis in Table 6 reveals clear performance hierarchies. High-arousal emotions prove most detectable across all models: *teasing* (average Spearman r: 0.617), *embarrassment* (0.585), and *anger* (0.536) show strong human-model alignment,

Table 6: Spearman's ρ by Emotion for Audio Models. Emotions are sorted by average performance across models, with best values in bold, runner-up underlined, and color-coded by correlation strength (gradient between red $= -1$ and blue $= 1$, NaN in gray). Several key patterns emerge: (i) Audio models demonstrate strong alignment with human annotations for high-arousal emotions like teasing. (ii) Our EMPATHICINSIGHT-VOICE models consistently outperform existing audio models across most emotions, or score second best. (iii) Some commercial models show systematic refusal patterns (NaNs) for sensitive emotions (sexual content, intoxication). (iv) Performance dramatically drops for subtle, low-arousal emotions like concentration. (v) Even state-of-the-art models struggle with complex cognitive-emotional states (e.g., contemplation), suggesting current audio models may be limited to more physiologically manifest emotions.

emotion	GPT-4o Mini Audio	GPT-4o Audio	Hume Voice	Gemini 2.0 Flash	Gemini 2.5 Pro	EMPATHICINSIGHT-VOICE SMALL (ours)	EMPATHICINSIGHT-VOICE LARGE (ours)	avg.
Teasing	0.569	0.636	NaN	0.556	0.626	0.649	**0.662**	0.617
Embarrassment	0.550	0.637	0.416	0.529	0.618	0.669	**0.678**	0.585
Anger	0.496	0.555	0.418	0.526	**0.602**	0.578	0.577	0.536
Impatience and Irritability	0.455	0.471	NaN	0.448	0.504	0.554	**0.570**	0.500
Malevolence/Malice	0.345	NaN	NaN	0.333	0.529	0.562	**0.615**	0.477
Shame	0.437	0.393	0.441	0.419	0.516	0.552	**0.558**	0.474
Sadness	0.470	0.404	0.357	0.466	**0.529**	0.483	0.521	0.461
Helplessness	0.347	0.375	NaN	0.462	0.483	**0.536**	0.535	0.457
Astonishment/Surprise	**0.487**	NaN	NaN	0.454	0.459	0.451	0.428	0.456
Pleasure/Ecstasy	0.364	NaN	NaN	0.342	0.462	**0.538**	0.529	0.447
Disgust	0.421	**0.493**	0.330	0.419	0.483	0.419	0.460	0.432
Contempt	0.412	0.433	0.324	0.407	0.466	**0.478**	0.469	0.427
Fear	0.355	0.367	0.437	0.353	0.441	**0.470**	0.458	0.411
Amusement	0.412	0.362	0.380	0.355	0.432	0.454	**0.462**	0.408
Relief	0.317	0.361	0.398	0.349	0.463	0.462	**0.501**	0.407
Pain	0.365	0.345	0.370	0.386	0.413	0.472	**0.474**	0.404
Jealousy/ Envy	0.334	0.361	0.264	0.425	**0.487**	0.469	0.471	0.402
Elation	0.390	0.330	0.313	0.344	0.466	0.475	**0.487**	0.401
Pride	0.348	0.308	0.259	0.415	0.482	**0.484**	0.474	0.396
Confusion	0.379	0.339	0.331	0.358	**0.451**	0.423	0.451	0.390
Disappointment	0.301	**0.466**	0.249	0.370	0.426	0.432	0.461	0.386
Doubt	0.379	0.347	0.241	0.403	0.402	0.459	**0.463**	0.385
Triumph	0.333	0.279	0.216	0.370	**0.482**	0.460	0.455	0.371
Infatuation	0.315	0.317	NaN	0.354	**0.413**	0.392	0.408	0.367
Bitterness	0.330	0.324	NaN	0.286	0.360	**0.411**	0.404	0.352
Fatigue/Exhaustion	0.221	NaN	NaN	0.297	0.400	**0.455**	0.384	0.351
Thankfulness/Gratitude	0.297	NaN	NaN	0.281	**0.418**	0.358	0.379	0.347
Intoxication/ Altered States of Consciousness	0.198	NaN	NaN	0.269	0.241	0.486	**0.487**	0.336
Distress	0.374	0.369	-0.138	0.375	**0.450**	0.432	0.430	0.327
Sexual Lust	0.203	0.279	NaN	0.356	**0.450**	0.332	0.334	0.326
Affection	0.310	**0.390**	0.182	0.330	0.349	0.359	0.356	0.325
Longing	0.289	0.330	0.214	0.326	0.348	**0.365**	0.350	0.317
Awe	0.298	0.276	0.058	0.314	0.314	0.329	**0.332**	0.275
Hope/Enthusiasm/Optimism	0.250	NaN	NaN	0.175	0.203	**0.345**	0.343	0.263
Sourness	0.158	0.180	NaN	0.250	0.303	**0.331**	0.323	0.258
Interest	0.161	0.169	0.119	0.148	0.287	**0.351**	0.315	0.221
Contemplation	0.187	0.128	0.177	0.252	**0.282**	0.263	0.247	0.219
Contentment	-0.044	-0.019	0.195	0.140	0.224	0.231	**0.330**	0.151
Emotional Numbness	0.139	0.092	NaN	0.099	0.125	0.139	**0.145**	0.123
Concentration	0.085	0.019	**0.262**	0.186	0.151	0.055	0.068	0.118

suggesting these acoustic signatures are most reliably encoded in prosody. Conversely, performance drops dramatically for subtle, low-arousal states like *concentration* (0.118) and *emotional numbness* (0.123), highlighting fundamental limitations in detecting nuanced emotional states from audio alone.

Moreover, the table reveals systematic differences in emotion detection across our 40-category taxonomy. It demonstrates that EMPATHICINSIGHT-VOICE models consistently outperform competitors across most emotions, particularly excelling in complex states often missed by other systems. For instance, EMPATHICINSIGHT-VOICE achieves superior performance on challenging emotions like *intoxication* (where EMPATHICINSIGHT-VOICE scores 0.48 compared to 0.269 by the runner-up and many commercial models often completely refuse assessment), and similar for *malevolence*—emotions that require nuanced prosodic understanding.

Commercial Model Limitations. Commercial models exhibit systematic refusal patterns for sensitive content, with GPT-4o Audio and Hume Voice showing nearly identical NaN patterns for emotions like *sexual content* and *intoxication*—indicating shared (safety) constraints. This creates evaluation gaps precisely where human emotional complexity is especially relevant for applications. Even state-of-the-art models struggle with complex cognitive-emotional states (*contemplation*, *interest*, *contentment*), suggesting current architectures may be fundamentally limited to more physiologically manifest emotions rather than subtle internal states.

5 Discussion

Our analysis reveals a fundamental relationship between human annotation consensus and model performance in audio-based emotion recognition, with implications that extend beyond the specific task to the broader understanding of machine learning on subjective human judgments.

We demonstrated that EMPATHICINSIGHT-VOICE models advance the state-of-the-art significantly, with best error and ordering values. Yet, MAE values around 3.0 on a 0-10 scale indicate substantial room for improvement even in our best model. The consistent pattern of high-arousal emotions being more detectable than low-arousal states across all architectures suggests this represents a fundamental challenge in audio-based emotion recognition rather than a limitation of specific models.

ASR models don't (yet) understand emotions. ASR models like Whisper currently lack the ability to accurately understand and represent nuanced emotions [17, 7]. However, by continually pretraining these models, we can enable them to perceive and interpret emotions in a way that supports more human-like predictions, as we demonstrated. Our EMONET-VOICE BIG dataset represents a crucial first step toward equipping foundation models with this emotional understanding.

Annotation Ambiguity Predicts Model Performance. The most striking finding from our comparative analysis is the systematic correlation between inter-annotator agreement and model performance across the emotional spectrum. Emotions exhibiting strong human consensus, such as *Teasing* (Spearman's $\rho = 0.617$), *Embarrassment* ($\rho = 0.583$), and *Anger* ($\rho = 0.536$), demonstrate both high agreement rates (predominantly green regions in Figure 1) and superior model alignment (top, dark blue in Table 6). Conversely, cognitively complex emotions like *Concentration* ($\rho = 0.118$), *Contemplation* ($\rho = 0.151$), and *Contentment* ($\rho = 0.123$) exhibit substantial human disagreement and correspondingly poor model performance.

This pattern suggests that model failures may not represent algorithmic inadequacies but rather reflect genuine perceptual ambiguities inherent in the emotional recognition task itself. We propose that inter-annotator agreement might establish a practical upper bound for model performance, as it is not be expected from computational systems to exceed human consensus on subjective human judgments.

Arousal-Dependent Recognition Bias. Our results demonstrate a clear arousal-based performance hierarchy, with high-energy emotions consistently outperforming their low-arousal counterparts. This bias appears across all model architectures, from transformer-based systems (GPT-4o variants) to specialized audio models (Hume Voice), suggesting a fundamental limitation in current acoustic feature extraction paradigms.

High-arousal emotions like *Anger*, *Embarrassment*, and *Impatience and Irritability* likely produce more distinctive acoustic signatures—increased pitch variance, amplitude fluctuations, and prosodic changes—that are readily captured by existing audio processing pipelines. In contrast, low-arousal states such as *Contemplation* and *Concentration* may manifest through subtle changes in speech patterns that fall below current model sensitivity thresholds.

This finding has significant implications for real-world applications: current audio emotion recognition systems may be inherently biased toward detecting emotional extremes while systematically underperforming on the nuanced, everyday emotional states that characterize much of human interaction. Furthermore, the arousal-dependent performance bias indicates that current audio processing architectures may be learning *acoustic stereotypes* of emotions rather than developing genuine emotional understanding. Models excel at detecting prototypical emotional expressions while failing on subtle variations, suggesting they may be capturing surface-level patterns rather than underlying emotional concepts.

The Cognitive Emotion Recognition Gap. A particularly noteworthy pattern emerges for cognitively-oriented emotions—states that require contextual understanding beyond immediate acoustic features. Emotions such as *Contemplation*, *Interest*, and *Concentration* represent mental processes rather than affective responses, and their recognition may fundamentally require understanding *why* someone is in a particular state, not merely *how* they sound while experiencing it.

This limitation points to a broader challenge in current emotion recognition paradigms: the reliance on acoustic features alone may be insufficient for detecting emotions that are primarily cognitive rather than affective. Future architectures might need to incorporate contextual information, dialogue history, or multimodal inputs to bridge this gap, going toward multimodal AI assistants.

5.1 Limitations and Future Directions

While our analysis provides valuable insights, several limitations should be acknowledged and addressed in future work.

The **fidelity of synthetic data** from GPT-4o Audio generations underlying our datasets, while state-of-the-art, may still exhibit subtle differences from genuine human vocalizations, meaning model performance on this benchmark might not directly generalize to spontaneous real-world speech. The benchmark primarily evaluates the recognition of **emotional portrayals** in synthetic speech driven by acting scenarios, which differs from the often more nuanced or mixed cues in authentic, spontaneous expressions. Furthermore, the dataset inherently reflects the capabilities and potential biases of the specific audio generation model used. While EMONET-VOICE incorporates 11 voices and 4 languages, this speaker and linguistic diversity does not yet encompass the full spectrum of human identities, accents, dialects, or age ranges. Finally, **emotion perception is inherently subjective**; while expert consensus minimizes variability, the labels represent a reliable approximation of perceived emotion in synthetic stimuli rather than an objective internal state.

5.2 Ethical Considerations

This work responds to growing concerns about the unintended effects of emotionally uncalibrated AI systems. As AI models become more capable of producing emotionally charged content, it is essential to understand how people interpret and respond to these synthetic expressions. Our datasets offer a basis for exploring potential risks, including miscommunication and emotional manipulation. We recognize the ethical challenges, especially regarding misuse for manipulative ends—concerns that underscore our commitment to transparency and safety. In response, we advocate for the development of safeguards to mitigate such misuse [10].

The development of EMONET-VOICE was guided by a strong ethical commitment, primarily addressed through the exclusive use of synthetic voice generation. This approach deliberately avoids the privacy risks associated with collecting real human emotional expressions, particularly those tied to sensitive or deeply personal experiences—such as pain, shame, or sexual desire—that would be difficult, if not impossible, to collect ethically and at scale from human participants. All voice samples in EMONET-VOICE are artificially generated using TTS models, with manual filtering and prompt diversification to reflect a broad range of gender, demographic, and accent representations while minimizing problematic content, motivated by Friedrich et al. [9]. Although the likelihood is extremely low, we acknowledge the remote possibility that some samples may resemble real individuals [11]; however, no personally identifiable data was used at any stage.

We release EMONET-VOICE as a research artifact with the recommendation to use it for academic purposes and encourage thorough examination of potential downstream biases and ethical implications. We invite users to engage with our tools, transparently report any unexpected behaviors, and contribute feedback to help advance responsible data curation and safer AI development.

6 Conclusion

We introduced EMONET-VOICE, novel datasets for fine-grained speech emotion estimation, designed to address critical limitations in existing SER resources. We create a large-scale, pretraining datasets EMONET-VOICE BIG, which is a synthetic multilingual voice dataset. Derived from it, we establish EMONET-VOICE BENCH which has psychology expert annotation, utilizing a 40-category emotion taxonomy with 3-level ratings. Their synthetic nature, combined with diverse voice and language coverage (11 voices, 4 languages, with balanced representations), prevents privacy concerns inherent in collecting authentic sensitive emotional data and broadens diversity. Furthermore, we create

EMPATHICINSIGHT-VOICE models (Small and Large), which establish state-of-the-art in speech emotion recognition. Existing foundation models like Gemini, GPT4o and Hume Voice perform significantly worse.

Our results indicate gaps in current emotion recognition and hint to several future research paths. Future research should investigate whether the agreement-performance relationship holds across different modalities (text, video, physiological signals) and develop targeted architectures to handle low-agreement emotional categories more effectively. The development of context-aware models that can leverage situational information may be particularly promising for addressing the cognitive emotion recognition gap.

Expanding EMONET-VOICE with more samples, languages, and speaker profiles using next-generation voice synthesis represents a key priority, along with exploring multiple generative models to mitigate single-model bias. Investigating cross-modal consistency by generating corresponding facial expressions or scenarios for the same emotional prompts offers a path toward richer multimodal benchmarks. Further analysis could also explore model performance variations across different languages or speaker voices within the current dataset to better understand the scope and limitations of current approaches.

7 Acknowledgements

We gratefully acknowledge the support of Intel (oneAPI Center of Excellence), DFKI, Nous Research (providing cluster access and compute), TU Darmstadt, TIB–Leibniz Information Centre for Science and Technology, and hessian.AI (providing compute and helpful discussions), and the open-source community for contributing to emotional AI. This work benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA No 952215), the Hessian research priority program LOEWE within the project WhiteBox, the HMWK cluster projects "Adaptive Mind" and "Third Wave of AI", and from the NHR4CES. Furthermore, this work was partly funded by the Federal Ministry of Education and Research (BMBF) project "XEI" (FKZ 01IS24079B).

References

[1] Lisa Feldman Barrett. The theory of constructed emotion. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.

[2] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

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

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 2014.

[5] Alan S Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698, 2019.

[6] Alan S Cowen, Petri Laukka, Dacher Keltner, and Klaus R Scherer. The sound of 24 emotions: A new audio dataset of emotional vocalizations and a cross-cultural validation. *Behavior research methods*, 52:1101–1119, 2020.

[7] Soumya Dutta and Sriram Ganapathy. Leveraging content and acoustic representations for efficient speech emotion recognition. *arXiv preprint arXiv:2409.05566*, 2024.

[8] Paul Ekman. *An Argument for Basic Emotions*. Psychology Press, 1992.

[9] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. Multilingual text-to-image generation magnifies gender stereotypes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL, main)*, 2025.

[10] Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Llavaguard: An open vlm-based framework for safeguarding vision datasets and models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2025.

[11] Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Does clip know my face? *Journal of Artificial Intelligence Research (JAIR)*, 2024.

[12] Philip Jackson and Syed Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

[13] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. Wrime: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104. Association for Computational Linguistics, 2021.

[14] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A. Hale. Why human–ai relationships need socioaffective alignment. *Nature, Humanities and Social Sciences Communications*, 2025.

[15] Clara Lee and Rafael Gomez. Emotion recognition in human-computer interaction: Why affective understanding matters. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3):1–25, 2021.

[16] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett. *Handbook of Emotions*. Guilford Press, 4th edition, 2016.

[17] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. *arXiv preprint arXiv:2305.16065*, 2023.

[18] Kristen A. Lindquist. Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model. *Emotion Review*, 5(4):356–368, 2013. Place: US Publisher: Sage Publications.

[19] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 2018.

[20] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. Acted vs. natural emotional speech: How different are they? *Proceedings of Interspeech*, 2017.

[21] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. arXiv:2406.07162, 2024.

[22] Stacy Marsella, Jonathan Gratch, and Paolo Petta. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2010.

[23] Emily Mower, Maja J Matarić, and Shrikanth Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070, 2011.

[24] Niloofar Nojavan and Mohammad Soleymani. Emoreact: Detecting emotions in children's speech and gestures. `https://github.com/bnojavan/EmoReact`, 2021. Accessed: 2025-05-06.

[25] OpenAI. Gpt-4 omniaudio: Unified audio generation for speech and sound. `https://openai.com`, 2024. Accessed via HyperLab reseller.

[26] Mohamed Osman, Daniel Z. Kaplan, and Tamer Nadeem. SER Evals: In-domain and out-of-domain benchmarking for speech emotion recognition. arXiv:2408.07851, 2024.

[27] Robert Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001.

[28] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *arXiv preprint arXiv:2005.00357*, 2020.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[30] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[31] Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. SERAB: A multi-lingual benchmark for speech emotion recognition. *arXiv preprint arXiv:2110.03414*, 2021.

[32] Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Felix Friedrich, Maurice Kraus, Krishna Kalyan, Kourosh Nadi, Huu Nguyen, Kristian Kersting, and Sören Auer. EmoNet-Face: An expert-annotated benchmark for synthetic emotion recognition, 2025.

[33] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Interspeech 2013*, ISCA, August 2013. ISCA.

[34] Björn W Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

[35] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Bjoern W Schuller. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd International Multimodal Sentiment Analysis Challenge and Workshop*, 2021.

[36] Paige Tuttösí, Mantaj Dhillon, Luna Sang, Shane Eastwood, Poorvi Bhatia, Quang Minh Dinh, Avni Kapoor, Yewon Jin, and Angelica Lim. BERSting at the screams: A benchmark for distanced, emotional and shouted speech recognition. arXiv:2505.00059, 2025.

[37] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multimodal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593. Association for Computational Linguistics, 2020.

[38] Lin Zhao and Anand Kumar. Speech emotion recognition: A review of datasets, methods, and challenges. *IEEE Transactions on Affective Computing*, 13(1):305–325, 2022.

[39] Kun Zhou, Jia Liu, and Yuexian Chen. Emotional speech dataset (esd): A parallel corpus in English and Mandarin. In *Interspeech*, 2021.

A Appendices

A.1 EMONET-VOICE Taxonomy

The 40 emotion categories used in EMONET-VOICE, adapted from EMONET-FACE [32], are listed below with associated descriptive terms used during conceptualization and prompting:

- **Amusement:** 'lighthearted fun', 'amusement', 'mirth', 'joviality', 'laughter', 'playfulness', 'silliness', 'jesting'
- **Elation:** 'happiness', 'excitement', 'joy', 'exhilaration', 'delight', 'jubilation', 'bliss', 'Cheerfulness'

- **Pleasure/Ecstasy:** 'ecstasy', 'pleasure', 'bliss', 'rapture', 'Beatitude'
- **Contentment:** 'contentment', 'relaxation', 'peacefulness', 'calmness', 'satisfaction', 'Ease', 'Serenity', 'fulfillment', 'gladness', 'lightness', 'serenity', 'tranquility'
- **Thankfulness/Gratitude:** 'thankfulness', 'gratitude', 'appreciation', 'gratefulness'
- **Affection:** 'sympathy', 'compassion', 'warmth', 'trust', 'caring', 'Clemency', 'forgiveness', 'Devotion', 'Tenderness', 'Reverence'
- **Infatuation:** 'infatuation', 'having a crush', 'romantic desire', 'fondness', 'butterflies in the stomach', 'adoration'
- **Hope/Enthusiasm/Optimism:** 'hope', 'enthusiasm', 'optimism', 'Anticipation', 'Courage', 'Encouragement', 'Zeal', 'fervor', 'inspiration', 'Determination'
- **Triumph:** 'triumph', 'superiority'
- **Pride:** 'pride', 'dignity', 'self-confidently', 'honor', 'self-consciousness'
- **Interest:** 'interest', 'fascination', 'curiosity', 'intrigue'
- **Awe:** 'awe', 'awestruck', 'wonder'
- **Astonishment/Surprise:** 'astonishment', 'surprise', 'amazement', 'shock', 'startlement'
- **Concentration:** 'concentration', 'deep focus', 'engrossment', 'absorption', 'attention'
- **Contemplation:** 'contemplation', 'thoughtfulness', 'pondering', 'reflection', 'meditation', 'Brooding', 'Pensiveness'
- **Relief:** 'relief', 'respite', 'alleviation', 'solace', 'comfort', 'liberation'
- **Longing:** 'yearning', 'longing', 'pining', 'wistfulness', 'nostalgia', 'Craving', 'desire', 'Envy', 'homesickness', 'saudade'
- **Teasing:** 'teasing', 'bantering', 'mocking playfully', 'ribbing', 'provoking lightly'
- **Impatience and Irritability:** 'impatience', 'irritability', 'irritation', 'restlessness', 'short-temperedness', 'exasperation'
- **Sexual Lust:** 'sexual lust', 'carnal desire', 'lust', 'feeling horny', 'feeling turned on'
- **Doubt:** 'doubt', 'distrust', 'suspicion', 'skepticism', 'uncertainty', 'Pessimism'
- **Fear:** 'fear', 'terror', 'dread', 'apprehension', 'alarm', 'horror', 'panic', 'nervousness'
- **Distress:** 'worry', 'anxiety', 'unease', 'anguish', 'trepidation', 'Concern', 'Upset', 'pessimism', 'foreboding'
- **Confusion:** 'confusion', 'bewilderment', 'flabbergasted', 'disorientation', 'Perplexity'
- **Embarrassment:** 'embarrassment', 'shyness', 'mortification', 'discomfiture', 'awkwardness', 'Self-Consciousness'
- **Shame:** 'shame', 'guilt', 'remorse', 'humiliation', 'contrition'
- **Disappointment:** 'disappointment', 'regret', 'dismay', 'letdown', 'chagrin'
- **Sadness:** 'sadness', 'sorrow', 'grief', 'melancholy', 'Dejection', 'Despair', 'Self-Pity', 'Sullenness', 'heartache', 'mournfulness', 'misery'
- **Bitterness:** 'resentment', 'acrimony', 'bitterness', 'cynicism', 'rancor'
- **Contempt:** 'contempt', 'disapproval', 'scorn', 'disdain', 'loathing', 'Detestation'
- **Disgust:** 'disgust', 'revulsion', 'repulsion', 'abhorrence', 'loathing'
- **Anger:** 'anger', 'rage', 'fury', 'hate', 'irascibility', 'enragement', 'Vexation', 'Wrath', 'Peevishness', 'Annoyance'
- **Malevolence/Malice:** 'spite', 'sadism', 'malevolence', 'malice', 'desire to harm', 'schadenfreude'
- **Sourness:** 'sourness', 'tartness', 'acidity', 'acerbity', 'sharpness' (Note: Primarily gustatory, vocal correlates might be subtle reactions)
- **Pain:** 'physical pain', 'suffering', 'torment', 'ache', 'agony'
- **Helplessness:** 'helplessness', 'powerlessness', 'desperation', 'submission'
- **Fatigue/Exhaustion:** 'fatigue', 'exhaustion', 'weariness', 'lethargy', 'burnout', 'Weariness'
- **Emotional Numbness:** 'numbness', 'detachment', 'insensitivity', 'emotional blunting', 'apathy', 'existential void', 'boredom', 'stoicism', 'indifference'
- **Intoxication/Altered States of Consciousness:** 'being drunk', 'stupor', 'intoxication', 'disorientation', 'altered perception'
- **Jealousy & Envy:** 'jealousy', 'envy', 'covetousness'

A.2 More Details on SOTA SER Model Training Methodology

This section provides an in-depth description of the training procedures for the models discussed in Section 4: i.e. the Whisper backbone and the EMPATHICINSIGHT-VOICE ensembles.

Data Curation and Fine-tuning for Emotion Captioning. Our goal was to adapt pre-trained Whisper models [29] for the task of generating nuanced emotional captions from speech. The data generation and fine-tuning pipeline involved several key steps:

1. **Initial Large-Scale Data Sources:** The primary data source was the EMONET-VOICE BIG synthetic voice-acting dataset. This was augmented with approximately 4,500 hours of audio extracted from publicly available online videos (vlogs, diaries, documentaries). We applied voice activity detection (VAD) to isolate speech segments ranging from 3 to 12 seconds.
2. **Dimensional Emotion Scoring with Gemini Flash 2.0:** All audio snippets—both from EMONET-VOICE BIG and the VAD-extracted clips—were annotated using Gemini Flash 2.0. A complex, multi-shot prompt (detailed in the supplementary materials) guided the model to produce intensity scores on a 0–4 scale (0 = absent, 4 = extremely present) for each of our 40 emotion dimensions simultaneously. This provided a structured, dimensional representation of perceived emotional content.
3. **Iterative Caption Generation for Whisper Training:**
   - Our initial attempt was to fine-tune Whisper to *directly regress* these 40-dimensional scores (i.e., to output numerical values), but this approach consistently collapsed into predicting nonsensical sequences of numbers. Similarly, training a specialized output head to perform ordinal regression utilizing a Wasserstein distance loss did not yield more sophisticated or coherent captions.
   - We then converted the dimensional scores into *procedurally generated string captions* using predefined templates (e.g., "The speaker sounds strongly amused and slightly joyful."). Training on these templated captions improved over direct regression, but the resulting Whisper outputs still tended toward repetitive or syntactically unnatural phrasing.
   - The most effective strategy was to take those procedurally generated captions and run them back through Gemini Flash 2.0 for *paraphrasing.* This second pass introduced significant linguistic diversity and more natural sentence structures, while preserving the original 40-dimensional semantics. The paraphrasing prompt specifically encouraged varied wording and sentence complexity.
4. **Training Data Preparation:** All EMONET-VOICE BIG audio segments longer than 30 seconds were truncated to their first 30 seconds, to meet Whisper's input constraints. Very long segments were further subdivided at silent regions into shorter clips, resulting in a final training pool of over 2 million audio–caption pairs when combined with the processed VAD data.
5. **Whisper Fine-tuning:** Various sizes of OpenAI's Whisper models were then fine-tuned on this dataset of audio paired with the paraphrased emotional captions. The objective was to teach Whisper to generate fluid, context-sensitive descriptions of emotional content given raw speech input. Iteratively refining the captions via paraphrasing proved crucial for yielding outputs that were both semantically accurate and linguistically natural. We also experimented with incorporating synthetic "emotion bursts" during fine-tuning, but this led to degraded embedding quality and was therefore not used in the final models.

EMPATHICINSIGHT-VOICE: MLP Ensembles for Dimensional Emotion Prediction. The EMPATHICINSIGHT-VOICE models were designed to provide direct predictions for each of the 40 emotion dimensions—complementing the captioning approach with explicit scalar estimates.

1. **Feature Extraction:** We used the encoder from our best-performing Whisper variant as a fixed feature extractor. For any input audio, we ran it through the Whisper encoder and collected the full sequence of token embeddings (sequence length = 1,500; embedding dimension = 768), yielding 1,152,000 features when flattened. Preliminary experiments showed that preserving the entire unpooled sequence outperformed all tested pooling strategies (mean, max, min, concatenation) for downstream MLP regression.
2. **MLP "Expert" Heads:** We trained an ensemble of 40 independent MLP models. Each MLP served as an "expert" head dedicated to regressing the intensity score for exactly one

of the 40 emotion dimensions using the corresponding flattened Whisper embeddings as input.

3. **Training Targets:** The regression targets were the direct 0–4 intensity scores produced by Gemini Flash 2.0 (via the multi-shot prompt described in the supplementary files). During the *encoder fine-tuning* stage, we experimented with injecting synthetic "emotion bursts"—artificially boosting certain dimension signals in the audio—to encourage a more robust embedding space. However, this augmentation degraded the underlying Whisper embeddings and ultimately hurt downstream MLP performance. Consequently, no synthetic bursts were used for final training.

4. **MLP Architecture:** Both the Small and Large EMPATHICINSIGHT-VOICE variants share the same overall architectural pattern for regressing from the high-dimensional flattened embeddings:

   - *Input Projection:* A first linear layer reduces the 1,152,000-dimensional input to a much smaller embedding space.
   - *Hidden Layers:* Three fully connected layers with ReLU activations, each followed by dropout for regularization to mitigate overfitting.
   - *Output Layer:* A final linear projection that outputs a single continuous value in [0, 4], corresponding to the predicted intensity for that emotion.

5. **Model Sizes:**

   - EMPATHICINSIGHT-VOICE SMALL: The initial projection reduces 1,152,000 inputs to 64 dimensions. The subsequent hidden layer sizes are $64 \rightarrow 32 \rightarrow 16$. Each MLP head has about 73.73 million trainable parameters, the vast majority residing in that first projection layer.
   - EMPATHICINSIGHT-VOICE LARGE: The initial projection reduces 1,152,000 inputs to 128 dimensions. The subsequent hidden layers are $128 \rightarrow 64 \rightarrow 32$. This yields approximately 147.48 million trainable parameters per head, again dominated by the input projection.

6. **Parallel Inference and Training Loss:** At inference time, we evaluate all 40 MLP experts in parallel to predict the full 40-dimensional emotion profile (i.e., different strengths of emotionality across dimensions). During training, each MLP head is optimized independently using the mean absolute error (MAE) between predicted and target emotion strength.

All trained EMPATHICINSIGHT-VOICE models (Small and Large) and the associated inference code are available via our project page.

A.3 Hume Voice mapping

A.4 Detailed Taxonomy Construction Methodology

The 40-category emotion taxonomy utilized in both the EMONET-VOICE foundation and benchmark datasets was originally developed for the EmoNet-Face Benchmark [32].

The primary objective was to create a taxonomy that supports a more fine-grained and nuanced understanding of affective states in AI, moving beyond the limitations of traditional basic emotion models. This development was rooted in contemporary psychological research and significantly informed by the principles of the Theory of Constructed Emotion (TCE) [1].

The taxonomy was designed to encompass a wide array of affective experiences, including not only common positive and negative emotions but also intricate social emotions (e.g., *Embarrassment*, *Shame*, *Pride*), cognitive states (e.g., *Concentration*, *Doubt*, *Confusion*), and bodily states (e.g., *Pain*, *Fatigue*, *Intoxication*). Less typical but experientially relevant categories like *Sourness* and *Helplessness* were also incorporated. The full list of 40 categories and their descriptive word clusters can be found in App. A.1 (cross-referencing the list you already have, which is similar to App. Tab. 4 from the EmoNet-Face paper).

The construction process involved several key stages:

Hume Voice Label	**Our Taxonomy**
Joy	Elation
Empathic Pain	Distress
Guilt	-
Nostalgia	Longing
Determination	-
Surprise (positive)	Surprise
Horror	Fear
Calmness	Contentment
Desire	Sexual Lust
Awkwardness	Embarrassment
Satisfaction	Pleasure
Aesthetic Appreciation	Awe
Entrancement	Concentration
Romance	Infatuation
Love	Affection
Excitement	Arousal
Realization	Contemplation
Tiredness	Fatigue
Envy	Jealousy & Envy
Anxiety	-
Boredom	-
Adoration	-
Sympathy	-
Admiration	Admiration
Craving	Craving
Surprise (negative)	Astonishment

Table 7: Mapping of Hume Voice labels to our emotion taxonomy. Note that if one Hume Voice label fits to more than one emotion from our taxonomy, only one item was chosen.

1. **Literature-Driven Candidate Extraction:** The comprehensive "Handbook of Emotions" (946 pages) [16] was digitized using Optical Character Recognition (OCR). The digitized text was then divided into manageable 500-word segments.
2. **AI-Assisted Term Identification:** GPT-4 was employed to analyze these text segments and extract potential nouns representing emotion concepts.
3. **Refinement and Deduplication:** The initially extracted terms were aggregated, and duplicates were removed, resulting in a candidate list of approximately 170 unique emotion-related nouns.
4. **Expert-Guided Clustering and Categorization:** This refined list of terms underwent an iterative process of clustering. This involved independent categorization efforts by team members, followed by critical reviews and discussions. Psychologists and researchers in affective computing provided expert guidance throughout this phase to ensure the semantic coherence and psychological relevance of the emerging categories. Each of the final 40 categories represents a cluster of these semantically related emotion words.

In line with the Theory of Constructed Emotion, this taxonomy does not presuppose the biological universality or fixedness of these emotional categories. Instead, it is intended to facilitate context-aware and socially informed interpretations of affective expressions by AI systems. Recognizing the inherent ambiguity in perceiving emotions (e.g., a high-arousal vocal expression might be interpreted as amusement, elation, or excitement depending on context and observer), the taxonomy was specifically designed to support plausible multi-label annotations rather than forcing rigid, single-label classifications. This approach aims to enable richer and more contextually sensitive representations of emotion in AI.

B Annotation Platform Instructions and UI

Table 8: Summary of key dataset statistics for EMONET-VOICE. *Hume Voice provides 46 emotions on a continuous scale from 0-1, of which we were able to map 29 to our emotion taxonomy. Human annotators voted on a discrete scale: 0 (emotion not present), 1 (emotion weakly present), 2 (emotion strongly present). All scales were transformed to a 0-10 scale for further analysis. Note that GPT-4o Audio Preview was not able to process 2,100 samples (e.g., returned an empty response).

Annotator	Unique Audio Files	Emotions per Annotation	Scale
Human 1	6837	1	0-2
Human 2	6620	1	0-2
Human 3	2600	1	0-2
Human 4	11605	1	0-2
Human 5	343	1	0-2
Human 6	5600	1	0-2
EMPATHICINSIGHT-VOICE LARGE	12600	40	0-4
EMPATHICINSIGHT-VOICE SMALL	12600	40	0-4
GPT-4o Audio Preview 2024-12-17	10500	40	0-10
GPT-4o Mini Audio Preview	12600	40	0-10
Gemini 2.0 Flash	12600	40	0-10
Gemini 2.5 Pro	12600	40	0-10
Hume Voice	12600	*29	0-1

Audio Emotion Annotation

Instructions

In this task, you'll be assessing whether a specific emotion appears to be present in the audio recordings. Each recording will be associated with a single emotion label, and you need to decide whether that emotion is:

- **Not Present** - The emotion is not detectable in the audio
- **Weakly Present** - The emotion is somewhat present but not strong
- **Strongly Present** - The emotion is clearly and strongly expressed

Listen carefully to each recording and make your selection based on your perception of the emotion in the audio.

Figure 2: Instructions given to the human annotator for the expert annotation of EMONET-VOICE BENCH.

emotion	alpha	alpha ci lower	alpha ci upper	n items
Embarrassment	0.272	0.186	0.368	300
Teasing	0.271	0.178	0.362	300
Pain	0.247	0.160	0.334	300
Anger	0.220	0.129	0.310	300
Shame	0.216	0.122	0.297	300
Sadness	0.211	0.111	0.301	300
Distress	0.208	0.121	0.297	300
Malevolence	0.204	0.098	0.294	300
Contentment	0.197	0.109	0.281	300
Relief	0.196	0.098	0.280	300
Jealousy / Envy	0.194	0.095	0.282	300
Intoxication	0.193	0.104	0.279	300
Authenticity	0.185	0.093	0.279	300
Disappointment	0.176	0.071	0.271	300
Fear	0.161	0.066	0.241	300
Impatience and Irritability	0.159	0.057	0.247	300
Helplessness	0.158	0.070	0.246	300
Pride	0.156	0.057	0.241	300
Sexual Lust	0.149	0.048	0.243	300
Triumph	0.145	0.043	0.246	300
Elation	0.138	0.040	0.229	300
Overall	0.138	0.124	0.152	12600
Fatigue	0.129	0.034	0.217	300
Concentration	0.103	0.023	0.186	300
Disgust	0.103	0.004	0.195	300
Thankfulness	0.088	-0.008	0.178	300
Pleasure	0.082	-0.011	0.177	300
Doubt	0.078	-0.020	0.171	300
Amusement	0.068	-0.031	0.155	300
Infatuation	0.063	-0.031	0.150	300
Confusion	0.060	-0.027	0.148	300
Contempt	0.046	-0.045	0.130	300
Affection	0.043	-0.052	0.133	300
Bitterness	0.033	-0.051	0.116	300
Astonishment	0.021	-0.068	0.109	300
Contemplation	0.021	-0.065	0.104	300
Sourness	0.004	-0.088	0.083	300
Hope	-0.005	-0.106	0.090	300
Longing	-0.046	-0.149	0.046	300
Arousal	-0.066	-0.170	0.030	300
Interest	-0.094	-0.178	-0.009	300
Emotional Numbness	-0.099	-0.179	-0.017	300
Awe	-0.127	-0.218	-0.035	300

Table 9: Cronbach's α inter-rater reliability (0 = emotion absent; 1 = weakly present; 2 = strongly present) for each emotion category ($n = 300$ items per label), with 95% confidence intervals obtained via non-parametric bootstrap (1 000 resamples, seed = 42). "Overall" reports α and CI computed across all 40 emotion categories + 2 extra categories (12 000 + 600 = 12 600 total annotations). Note that the analysis contains two extra categories (authenticity and arousal) that is not present in the narrow emotion category definition A.1.

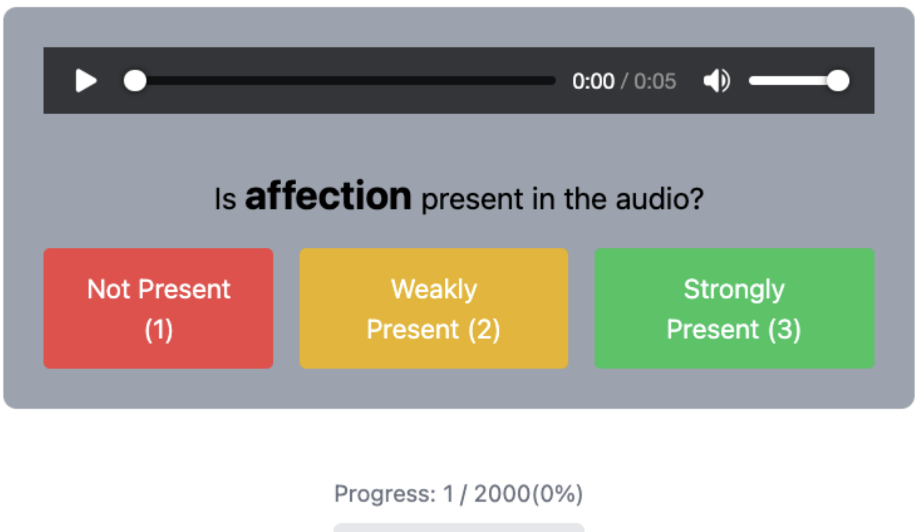

Figure 3: UI of our expert annotation tool for EMONET-VOICE BENCH.