EmotionLayer : A Multimodal Architecture for Empathic Voice Assistants

Based on Speech Emotion Recognition and Large Language Models

Alessio Bernardini

Department of Computer Science, University of Milan, Italy

alessio.bernardini@studenti.unimi.it

The Empathic Gap

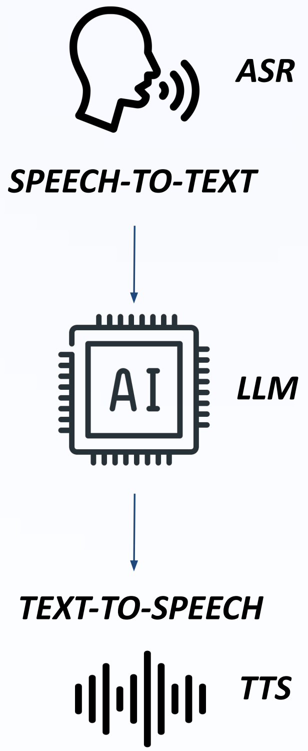
The Current Paradigm: Traditional voice systems (ASR → LLM → TTS) and newer Speech-to-Speech (STS) models process speech mainly as text, ignoring its ***prosodic*** and ***emotional components***.

The Missing Link: Valuable paralinguistic information such as ***tone***, ***rhythm***, ***hesitations***, and ***intensity*** is completely lost during transcription.

The consequence: The system “hears” the words but fails to “listen” to the emotional state.

Real Impact: In contexts where understanding the customer’s emotional state is essential, such as customer support, the lack of proper prosodic context leads to standardized, ***tone-deaf responses***, which can cause frustration and dissatisfaction.

Vocal Agent Pipeline



Questions & Contributions

Research Questions (RQs)

- **RQ1:** How effectively can LLMs ***detect*** and ***respond*** to complex emotions using multimodal input (text + emotional & audio context)?
- **RQ2:** Which LLM architectures offer ***optimal trade-offs*** between emotional ***intelligence, latency, and cost*** for production?

Key Contributions

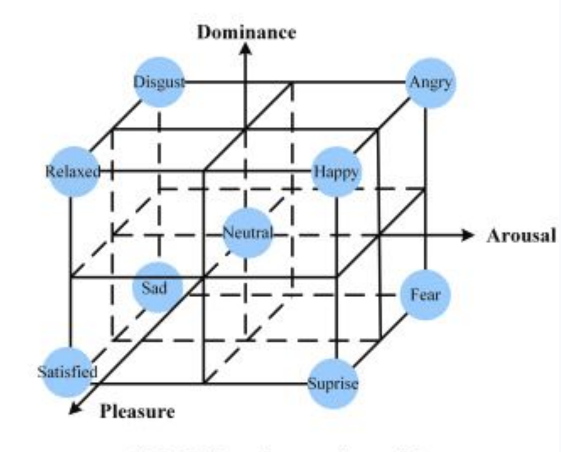
- **Italian SER Models:** Fine-tuning of Wav2Vec 2.0 for discrete ***emotion classification*** and continuous ***PAD regression***.
- **EmotionLayer Architecture:** Fuses text and prosody into a JSON payload ready for downstream LLMs.
- **Systematic Benchmark:** Evaluated 18 LLMs on 40 gold-standard scenarios via LLM-as-a-Judge to ***detect the correct user emotion*** and take the right decision.

Background Emotional Models

Ekman’s Categorical Model

Identifies 6 basic emotions:

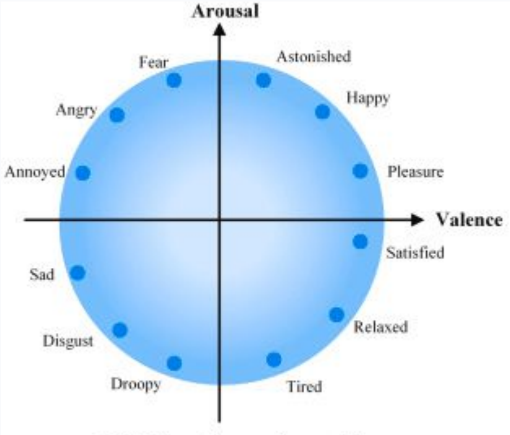
anger, disgust, fear, joy, sadness, *and* ***surprise***



(b) PAD dimension emotion model

PAD Dimensional Model

- **Valence (V):** Negative (-1) ↔ Positive (+1)
- **Arousal (A):** Calm (-1) ↔ Excited (+1)
- **Dominance (D):** Submissive (-1) ↔ Dominant (+1)



(a) VA dimension emotion model

State of the Art

Architectures

Wav2Vec 2.0 represents the state-of-the-art for self-supervised representations of the speech signal, utilizing advanced Transformer-based architectures.

Italian Datasets

Resources are limited. We utilize: **EMOVO** (professional actors), **Emozionalmente** (crowdsourced, 6,902 samples) **AI4SER** (the only Italian dataset with native PAD annotations).

Empathic Benchmarks

Frameworks like **EQ-Bench** and **EmotionQueen** evaluate emotional intelligence, empathy generation, and the comprehension of complex dynamics in LLMs.

System Architecture

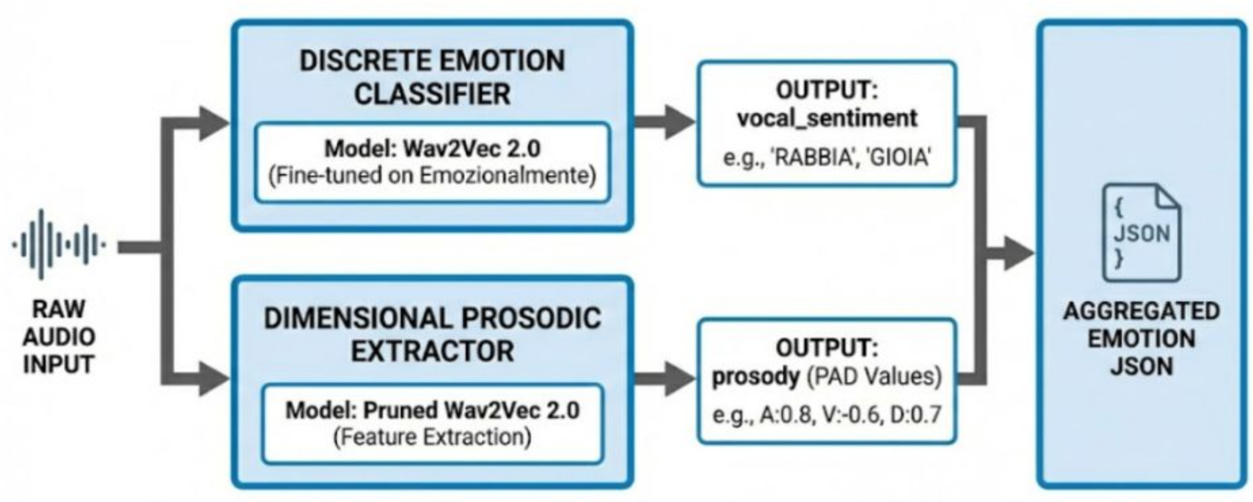
Parallel Audio Processing: The system analyzes raw audio streams simultaneously to extract distinct emotional vectors.

1. Discrete Emotion Classifier

Fine-tuned Wav2Vec 2.0 outputs a categorical label (e.g., "Anger: 96%") across 7 primary emotion classes.

2. PAD Prosodic Regressor

Pruned Wav2Vec 2.0 outputs continuous dimensional values (e.g., Arousal: 0.95, Valence: 0.05, Dominance: 0.90).


RAW AUDIO INPUT
DISCRETE EMOTION CLASSIFIER
Model: Wav2Vec 2.0
(Fine-tuned on Emozionalmente)
DIMENSIONAL PROSODIC EXTRACTOR
Model: Pruned Wav2Vec 2.0
(Feature Extraction)
OUTPUT: vocal_sentiment
e.g., 'RABBIA', 'GIOIA'
OUTPUT: prosody (PAD Values)
e.g., A:0.8, V:-0.6, D:0.7
{ JSON }
AGGREGATED EMOTION JSON

Emotion Taxonomy

Emotion	Status	Valence	Arousal	Dominance	Composition
Enthusiasm	😄	>0.6	>0.6	≈0.5	Joy + Surprise
Satisfaction	😄	>0.6	≈0.5	≈0.5	Joy + Good Text
Uncertainty	😐	<0.5	≈0.5	<0.5	Fear + Surprise
Frustration	😡	<0.4	>0.6	<0.5	Anger + Sadness
Sarcasm	☹	≈0.5	≈0.5	≈0.5	Discrepancy: Text vs Audio
Urgency	☹	<0.4	>0.8	<0.5	Fear + Anger
Insecurity	☹	<0.4	<0.4	<0.3	Fear + Sadness

Strategy Taxonomy

Target Strategy	Emotion Target	Speed	Tone	Pitch
Rapid Resolution	Urgency	1.25x	Professional	default
Assertive Neutrality	Sarcasm	1.00x	Professional	low
Active De-escalation	Anger, Frustration, Disgust	0.80x	Apologetic	low
Empathic Validation	Uncertainty, Insecurity, Sadness	0.85x	Empathetic	low
Positive Mirroring	Enthusiasm, Satisfaction, Joy	1.10x	Cheerful	high
Direct Efficiency	Neutral	1.00	Informative	balance

Evaluation Framework

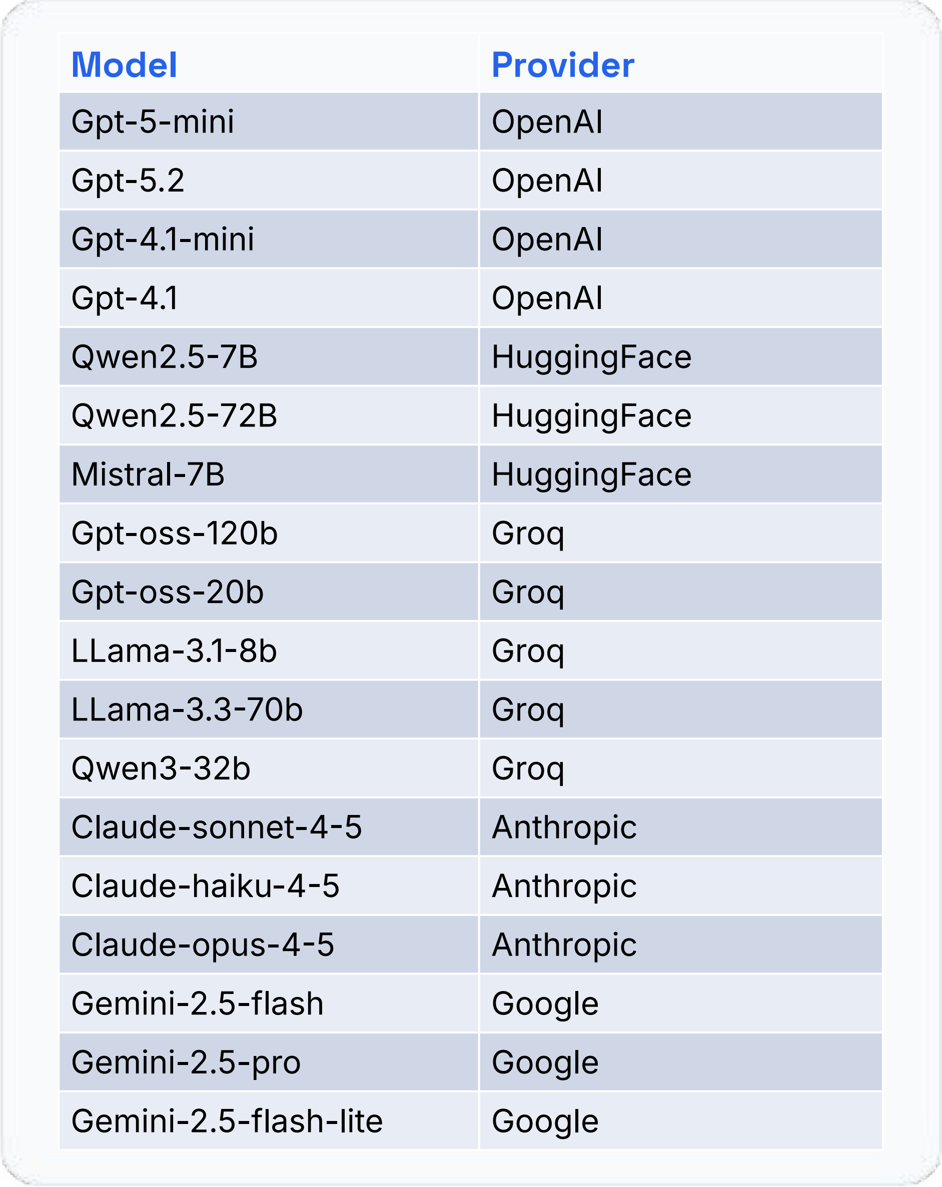
We evaluated 18 candidate LLMs over 40 gold-standard scenarios using an automated **LLM-as-a-Judge** methodology (by Claude Opus 4.6).

Binary Metrics (40% Weight)

- **Emotion Match (0/1):** Correctly identifying the emotion from multimodal JSON.
- **Strategy Match (0/1):** Selecting the appropriate strategy.

Continuous Metrics (60% Weight, Scale 1-5)

- **Relevance:** Technical accuracy of the response.
- **TTS Alignment:** Consistency of TTS parameters (speed/pitch).
- **Voice Suitability:** Conciseness and conversational naturalness.
- **Empathic Response:** Overall quality of empathy expressed.



Model	Provider
Gpt-5-mini	OpenAI
Gpt-5.2	OpenAI
Gpt-4.1-mini	OpenAI
Gpt-4.1	OpenAI
Qwen2.5-7B	HuggingFace
Qwen2.5-72B	HuggingFace
Mistral-7B	HuggingFace
Gpt-oss-120b	Groq
Gpt-oss-20b	Groq
LLama-3.1-8b	Groq
LLama-3.3-70b	Groq
Qwen3-32b	Groq
Claude-sonnet-4-5	Anthropic
Claude-haiku-4-5	Anthropic
Claude-opus-4-5	Anthropic
Gemini-2.5-flash	Google
Gemini-2.5-pro	Google
Gemini-2.5-flash-lite	Google

Model Performance Results

Provider	Models	Score	Latenza (s)	Costo ($)
Gemini	**gemini-2.5-flash**	**91.70**	4.328	0.001000
Gemini	**gemini-2.5-pro**	**90.50**	10.617	0.003944
OpenAI	**gpt-4.1-2025-04-14**	**90.30**	3.769	0.004647
Anthropic	claude-sonnet-4-5	89.46	8.475	0.010305
Anthropic	claude-opus-4-5	87.64	8.270	0.016995
Groq	**openai/gpt-oss-120b**	87.10	**1.594**	**0.000633**
Groq	qwen/qwen3-32b	86.20	**1.623**	0.000819
HuggingFace	Qwen2.5-72B-Instruct	85.60	8.927	**0.000647**
Gemini	gemini-2.5-flash-lite	85.40	**1.494**	**0.000235**
Groq	openai/gpt-oss-20b	84.00	**1.041**	**0.000364**
OpenAI	gpt-4.1-mini	82.75	3.362	0.000909
Groq	llama-3.3-70b-versatile	82.60	**0.758**	0.001051
OpenAI	gpt-5-mini-2025-08-07	80.30	19.667	0.003060
Anthropic	claude-haiku-4-5	80.06	3.909	0.003355

Model Performance Results

Leadership Gemini

Gemini 2.5 Flash-Preview (91.70) and **Pro** (90.50) dominate in terms of quality, achieving the highest overall scores and even outperforming OpenAI solutions.

Extreme Speed

The **Groq** infrastructure remains unrivalled in terms of latency. The **llama-3.1-8b-instant** model achieves record-breaking minimum response times of just 0.511 seconds.

Cost Efficiency

Gemini 2.5 Flash-Lite stands out for its exceptional value for money: a solid score (85.40) combined with extremely low operating costs ($0.000235).

Conclusions & Future Work

Key Conclusions

- **Bridging the Gap:** Solving the "Empathic Gap" in voice assistants is technically viable in production by injecting paralinguistic data pre-LLM.
- **Voice Suitability Matters:** LLMs optimized purely for complex reasoning (e.g., GPT-5) often generate overly verbose text that ruins conversational TTS naturalness.

Future Work

- **SER Improvements:** Enhance the current Speech Emotion Recognition accuracy (66.81%) by fine-tuning on spontaneous, non-acted speech datasets.
- **Human Evaluation:** Expand the LLM benchmark with multi-rater human evaluations on live interactions.
- **E2E Systems:** Transition towards End-to-End Speech-to-Speech models capable of natively and directly controlling prosodic output parameters.

Thank you for your attention !

Alessio Bernardini

alessio.bernardini@studenti.unimi.it