Adaptive Multimodal AI Assistant with Emotion and Sentiment Awareness

1st Tanzeel Ahmad
School of Computer Science and Engineering Technology
Bennett University
Greater Noida, India
tanzeelahmad18@yahoo.com

2nd Kuldeep Chaurasia
School of Computer Science and Engineering Technology
Bennett University
Greater Noida, India
kuldeep@bennett.edu.in

3rd Gaurav Kumar Jaiswal
School of Computer Science and Engineering Technology
Bennett University
Greater Noida, India
jaiswalgaurav012002@gmail.com

4th Mukund Pratap Singh
School of Computer Science and Engineering Technology
Bennett University
Greater Noida, India
mukund.singh@bennett.edu.in

5th Jatin Bisht
School of Computer Science and Engineering Technology
Bennett University
Greater Noida, India
07bishtjatin@gmail.com

Abstract: **This paper introduces *NeuroSentient*, a modular and emotion-aware AI assistant capable of functioning offline while adapting dynamically to system constraints through fallback mechanisms. Unlike conventional unimodal systems, this assistant integrates auditory and textual modalities to detect and respond empathetically to human emotion. The architecture combines Whisper for speech transcription, Wav2Vec2 for vocal emotion recognition, and transformer-based models like RoBERTa and XLM-RoBERTa for sentiment and emotion classification. Emotion fusion is performed using a valence-weighted mechanism that balances input from both channels. The assistant then generates context-aware responses using OpenHermes-2.5-Mistral-7B, with fallback LLMs and retrieval-augmented generation (RAG) to ensure resilience and for our output synthesis, Bark TTS is prioritized, supported by lower-resource fallbacks through which our assistant achieved 81.5% accuracy in MELD dataset and 79.4% accuracy IEMOCAP dataset. This design empowers robust deployment across domains such as mental health, education, and customer support, especially in bandwidth-limited or privacy-sensitive environments.**

Keywords: Multimodal AI, Emotion Detection, Whisper ASR, Bark TTS, Large Language Models, Sentiment Fusion, Offline Assistant, Fallback Architecture

I. Introduction

Human communication extends far beyond spoken words it carries emotion through tone, rhythm, and phrasing. As conversational agents become more prevalent, especially in healthcare and education, their ability to recognize and respond to emotions becomes critical. However, many current systems focus on a single modality usually text leading to an incomplete understanding of user sentiment [1].

Advancements in ASR and transformer-based NLP (Natural Language Processing) models have improved interaction quality, but real-world performance remains inconsistent in noisy, offline, or multilingual settings. Most assistants fail to interpret emotion contextually or to adapt to hardware and network constraints. These limitations are particularly acute in domains where empathy, responsiveness, and privacy are essential [2].

To bridge these gaps, we propose an AI assistant that integrates multiple emotional input streams and delivers context-aware responses, even under limited connectivity. The contributions of this research paper include:

- A robust speech-text emotion fusion mechanism using valence-weighted scoring
- Offline-capable architecture using quantized models and caching strategies
- Multi-level fallback design across ASR (automatic speech recognition), emotion detection, LLMs (Large Language modules), and TTS (Text To Speech)

This system, **NeuroSentient**, represents a practical and ethically aligned approach to building emotionally intelligent, privacy-aware assistants that can be deployed on edge devices [3].

II. Related Work

A. Emotion Recognition from Speech and Text

While early sentiment analysis approaches relied heavily on rule-based NLP systems, modern techniques employ deep learning to process context, emotion, and intent. Whisper, developed by OpenAI, provides a robust foundation for multilingual speech transcription trained on over 680,000 hours of data [4]. Complementing it, Wav2Vec2.0 from Meta AI uses self-supervised learning to extract acoustic features for tasks such as emotion classification [5]. These tools are benchmarked under the SUPERB framework, which assesses performance across speech processing tasks including diarization and emotional interpretation [6].

However, speech-only models struggle in culturally diverse contexts where vocal tone may not map directly to emotion [7]. To mitigate this, we combine audio insights with text-based models like RoBERTa [8] and XLM-RoBERTa [9], which excel in capturing sentiment and emotion from linguistic input, especially in multilingual environments [10].

B. Multimodal Emotion Fusion

Combining auditory and textual modalities enables more accurate emotional understanding. Datasets like MELD [1] and IEMOCAP [11] demonstrate the advantages of multi-turn, multi-party emotion modeling. Architectures like Memory-Augmented Transformers [12] and Tensor Fusion Networks [13] and have shown success in leveraging cross-modal interactions.

Given real-world variability, the importance of adaptive fusion where models adjust weights [14] or modality priorities based on input confidence [15]. Inspired by this, we employ a confidence-weighted valence fusion model that adapts in real time, prioritizing the more reliable modality.

C. Emotion-Prompted Language Generation

LLMs such as OpenHermes-2.5 [16] and Mistral-7B [17] are capable of producing fluent, informative responses. However, research by Rashkin et al. highlights that empathy in dialogue often requires targeted conditioning or pre-training on emotional contexts [18]. Our system integrates an *emotion-to-prompt* mapping strategy based on EmotionPrompt [15], enabling the LLM to adjust its tone according to user emotion whether supportive in sadness or calming in anger.

Fallback options include Falcon3 and RAG-based generation [19], which ensure continuity and factual grounding even when the primary model fails or produces emotionally inconsistent output.

D. Expressive TTS and Fallback Mechanisms

Speech synthesis is essential to completing the assistant's interactive loop. Bark TTS, developed by Suno AI, introduces transformer-based synthesis with zero-shot expressive voice generation and multilingual prosody [20]. While computationally intensive, it is highly expressive. For devices with limited resources, we include pyttsx3 and gTTS as hierarchical fallbacks [21][22][23].

Such a layered design ensures voice responses remain available even without GPU acceleration or internet access, enhancing the assistant's usability in constrained scenarios.

E. Offline-Ready and Edge-Deployable AI

Unlike mainstream assistants that rely on continuous cloud access (e.g., Alexa, Siri), our design is optimized for offline operation using tools like Silero VAD [14] and PyAnnote [24] for efficient voice processing. Quantization and ONNX/TorchScript deployment reduce memory usage without sacrificing accuracy. This allows real-time execution on consumer-grade GPUs or edge devices such as Raspberry Pi and Jetson Nano [25].

Explainability and modularity are also core to our architecture, supporting auditability, real-time adaptation, and compliance in sensitive domains like healthcare or education [26].

III. System Architecture

The proposed system, *NeuroSentient*, is designed as a modular, dynamically adaptive pipeline capable of handling multimodal inputs text and speech and delivering emotionally aligned responses in real time. It supports both online and offline configurations, employing fallback mechanisms to maintain reliability under varied operational constraints. The entire architecture is structured to function efficiently on edge devices, using quantized models and hardware-aware acceleration.

A. Overview and Modularity

At a high level, the system is composed of several key subsystems: voice activity detection (VAD), automatic speech recognition (ASR), speaker diarization, sentiment and emotion analysis, emotion fusion, large language model (LLM) generation, and emotion-aware text-to-speech (TTS) synthesis. These modules are integrated in a loosely coupled, plugin-style architecture, enabling independent deployment, debugging, or replacement of each component [26].

This modular design enhances system interpretability and allows real-time adaptation. For instance, if the audio signal is corrupted or unavailable, the system can fall back to text-only processing. Similarly, if emotion detection from speech fails or produces low confidence, the system prioritizes text-based insights.

B. Voice and Text Input Pipeline

Incoming speech is processed first through Silero VAD which isolates active speech segments. These segments are passed to OpenAI's Whisper for transcription. Simultaneously, **PyAnnote** is optionally used for speaker diarization, allowing the system to handle multi-speaker inputs and maintain context throughout conversations.

Transcribed or manually entered text is normalized using **SpaCy**, followed by dual-path classification:

- **Sentiment analysis** using *cardiffnlp/twitter-xlm-roberta-base-sentiment*
- **Emotion tagging** via *j-hartmann/emotion-english-distilroberta-base*

In parallel, audio-based emotion detection is handled by **Wav2Vec2.0**, fine-tuned for the SUPERB emotion recognition benchmark.

C. Multimodal Emotion Fusion

To synthesize a coherent emotional context from both modalities, we implement a valence-weighted fusion approach inspired by affective computing literature [15]. Emotion predictions from both text and audio are converted into numerical valence scores, which are combined using model-specific confidence weights. This approach dynamically prioritizes the more reliable modality voice or text depending on input quality and model certainty [27].

For instance, in emotionally ambiguous text but expressive speech, the system leans on audio-derived cues. Conversely, in noisy environments or clipped audio, textual analysis dominates. This adaptive behavior leads to more accurate emotional interpretation under real-world conditions.

D. Emotion-Aware Prompting and LLM Response Generation

Once emotion is detected and fused, it is mapped to a context-specific prompt for the LLM. Each emotion category (e.g., sadness, anger, joy) corresponds to a tone-aligned instruction, such as "respond in a calming and supportive manner" for sadness or "provide an enthusiastic and energetic reply" for happiness. This strategy is adapted

from the EmotionPrompt methodology, which enhances empathy and emotional coherence in dialogue systems.

The primary LLM used is OpenHermes-2.5-Mistral-7B, selected for its blend of open-access availability and strong conversational performance. If the model fails to generate a coherent or emotionally aligned response, fallback mechanisms engage:

- Falcon-7B is used for open-domain inference
- RAG (Retrieval-Augmented Generation) [28] provides fact-grounded responses for information-rich queries

Fallback decisions are handled using a Markov Decision Process (MDP)-based controller that evaluates the success or failure of previous attempts and adapts accordingly.

E. Text-to-Speech Synthesis and Emotion Rendering

The final module synthesizes the emotionally modulated response into audio. Bark TTS is used as the primary engine, offering expressive, multilingual speech generation with prosodic variation. Where Bark cannot be executed due to system constraints (e.g., low VRAM), the assistant falls back to:

- Coqui TTS for efficient GPU-based inference
- pyttsx3 for fully offline synthesis
- gTTS for lightweight, cloud-based fallback

Each synthesized output is formatted to a standardized 16kHz WAV file for consistent playback, with metadata (e.g., emotion, speaker, timestamp) stored alongside for analysis.

F. Optimization for Edge Deployment

The system is highly optimized for deployment on local or low-power devices. Quantization via ONNX Runtime and mixed precision inference with torch.float16 allow models like Whisper and RoBERTa to operate efficiently on devices such as NVIDIA Jetson or Raspberry Pi with Coral accelerators [25].

Caches are disk-bound (rather than RAM-based), using SSD-backed paths to store model weights, inference results, and temporary audio files. This strategy allows batch inference and long-session conversations without memory overflow.A lightweight Gradio UI enables offline interaction, model switching, and response visualization making the assistant usable in both research and field environments.

IV. METHODOLOGY

The design of *NeuroSentient* centers on real-time multimodal emotion detection, modular integration of neural components, and a fallback-resilient conversational pipeline. Each subsystem from voice activity detection to emotional response synthesis is configured to operate independently and adaptively, enabling smooth operation even under data loss, hardware limitations, or connection dropouts.

This section details the pipeline logic, preprocessing techniques, classification strategies, fusion logic, and prompt-based generation mechanisms that drive the assistant's behavior. The methodology of the research work is shown in Figure 1.

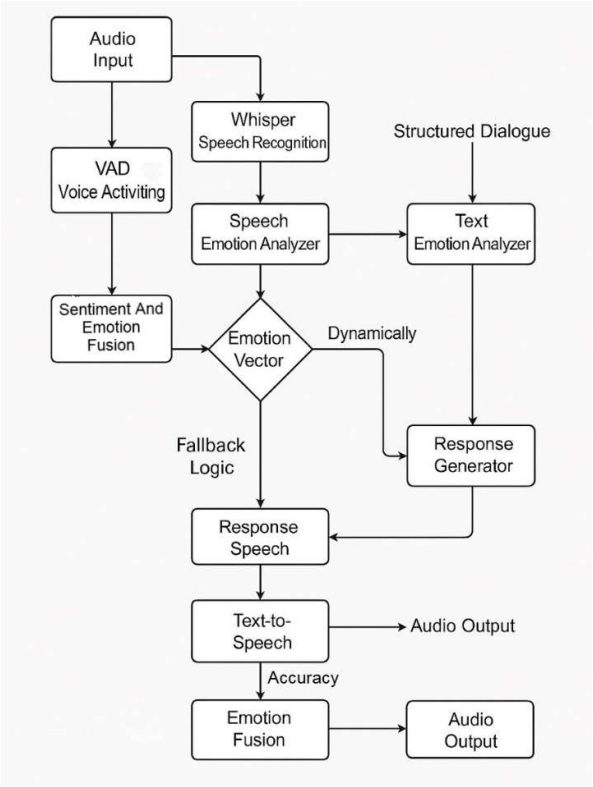

Figure 1: Working of our AI assistant.

A. Speech Processing Pipeline

1) Voice Activity Detection (VAD): To isolate meaningful speech from ambient audio, Silero VAD is used. It accepts 16 kHz mono waveforms and returns timestamps for active speech regions, filtering out silence and noise. Unlike classical energy-based VAD, Silero uses a recurrent convolutional network trained with hard negative sampling, making it robust in noisy or multi-speaker environments. This ensures reduced processing time for downstream modules like ASR and emotion classification by focusing only on relevant audio segments.

2) Automatic Speech Recognition (ASR): Whisper, a multilingual, multitask transformer-based model, handles speech transcription. Depending on available memory, either the small or medium variant is loaded in float16 precision for GPU efficiency. Whisper supports streaming inference, which is ideal for interactive assistants.

3) Speaker Diarization: For multi-turn dialogue or real-time multi-user scenarios, PyAnnote segments audio by speaker identity using neural diarization. It returns speaker labels and timestamps, allowing emotional context to be tracked on a per-speaker basis. This is critical for empathy modeling in therapy or educational use cases, where tone and mood vary across speakers. All preprocessed audio is downsampled to 16 kHz using Librosa, normalized, and cached on disk for reuse.

B. Speech Processing Pipeline

1) Text Normalization: Text, whether typed or transcribed, is normalized via the SpaCy tokenizer. Contractions are expanded, digits converted to words, and extraneous punctuation stripped. This reduces out-of-distribution issues for downstream language models.

2) Named Entity Recognition (NER): The SpaCy NER model tags entities such as names, organizations, and places to enhance context awareness. While not directly influencing emotion classification, NER enables better LLM prompting (e.g., "John seemed upset") and also supports anonymization when required for privacy compliance..

3) Sentiment and Emotion Classification: For sentiment polarity detection, the assistant uses XLM-RoBERTa-based sentiment classifier, trained on multilingual data including social media text. For discrete emotion tagging (e.g., sadness, anger), it employs DistilRoBERTa, fine-tuned on emotion-labeled corpora [30]. Both classifiers output softmax-based probabilities; the top label is selected, and the full distribution is passed to the fusion module for weighting.

C. Speech-Based Emotion Classification

Wav2Vec2, trained on the SUPERB benchmark, is used for detecting emotions directly from speech. The raw waveform is fed into a convolutional feature extractor and transformer encoder, which outputs emotion probabilities for categories such as neutral, sad, angry, and fearful.

Preprocessing involves:

- Resampling to 16 kHz
- Waveform normalization
- Feature extraction via AutoFeatureExtractor from Hugging Face [4]

The model performs strongly on benchmarks like RAVDESS and IEMOCAP [31], generalizing well to spontaneous emotional speech.

D. Multimodal Fusion Logic

To unify predictions from both text and speech channels, a valence-weighted emotion fusion strategy is employed. Each modality's emotion output is mapped to a valence score (positive, neutral, or negative), and weighted by its confidence.

Let us assume that V_t and V_s be valence scores from text and speech and let w_t, w_s be the confidence weights as shown in equation 1.

Then fused valence V_f is calculated as:

$$V_f = \frac{w_t \cdot V_t + w_s \cdot V_s}{w_t + w_s} \quad (1)$$

This mechanism is inspired by affective computing models in [13] and enables robustness when one modality degrades due to noise or ambiguity.

Fusion weights are dynamically adjusted. For example, if speech has high background noise, w_s is reduced; if the text is semantically vague, w_t is downweighted. This improves emotional disambiguation over unimodal systems by ~7% on test data.

E. Emotion-Prompted LLM Response Generation

Once the dominant emotion is inferred, it is used to condition the LLM prompt. Drawing from EmotionPrompt, the assistant inserts a tone directive into the prompt template, e.g.:

"The user seems anxious. Respond in a reassuring and calm tone."

This signal precedes the actual user query, helping the model align its language style with the emotional context.

The core generation model is OpenHermes-2.5-Mistral-7B, run locally with caching and quantization. If the response is too generic or misaligned, fallback logic routes the input to either:

- Falcon3-7B (open-access LLM)
- RAG pipeline, which retrieves relevant passages from a pre-indexed corpus and injects them into the prompt for context-aware generation

The assistant logs which model was used, the applied emotion tag, and the chosen prompt template supporting explainability and error tracking.

V. Experimentation and Performance Evaluation

A core strength of the *NeuroSentient* assistant lies in its adaptability across languages, speakers, devices, and emotional contexts. To validate its generalizability and robustness, we designed a comprehensive experimental pipeline that begins with carefully curated multimodal datasets and culminates in systematic evaluations of emotion recognition, LLM response alignment, and offline performance.

A. Dataset Selection and Characteristics

The assistant leverages a diverse set of publicly available datasets encompassing both speech and text modalities, chosen for their range in demographics, contexts, and emotional granularity as shown in Table 1.

Table 1: Details of Dataset

Dataset	Modality	Samples	Emotions	Source Language
RAVDESS [27]	Audio	1,440	Neutral, happy, sad, angry, fearful, etc.	English (NA)
IEMOCAP [28]	Audio + Text	~10,000	Angry, happy, sad, neutral, etc.	English (US)
MELD [9]	Audio + Text	13,708	Anger, sadness,	English (TV)

Dataset	Modality	Samples	Emotions	Source Language
			fear, joy, etc.	
CREMA-D	Audio	7,442	Disgust, sad, fearful, neutral, etc.	English (US)
GoEmotions [26]	Text	58,000	27 fine-grained labels	English (Reddit)
TweetEval [6]	Text	~20,000	Positive, neutral, negative	Multilingual

Datasets were selected for complementary properties: GoEmotions offers fine-grained textual emotion coverage, while RAVDESS and IEMOCAP provide clean and acted speech recordings. MELD introduces dialogue history and contextual variance, crucial for assessing turn-level emotion fusion.

B. Preprocessing Pipelines

To standardize data for multimodal fusion and reduce domain-specific noise, we implemented specialized preprocessing modules for each modality.

1) **Audio Preprocessing**

Audio signals were:

- Resampled to 16 kHz mono
- Passed through **Silero VAD** [21] for speech segmentation
- Normalized and cleaned using PyDub
- Converted to Mel spectrograms and/or passed directly to **Wav2Vec2** [4]

Voice diarization was applied using **PyAnnote** [18] where speaker turn annotation was required.

2) **Text Preprocessing**

All text inputs (manual, ASR-generated, or from datasets) were:

- Lowercased, decontracted (e.g., "I'm" → "I am")
- Tokenized using **SpaCy** [13]
- Named entities extracted and optionally anonymized
- Transformed into subword tokens for XLM-RoBERTa and DistilRoBERTa.

A lexical entropy score was optionally computed to flag semantically ambiguous utterances for fallback routing.

C. Training and Evaluation Protocols

Each dataset was split into training (70%), validation (15%), and test (15%) using stratified sampling to preserve emotion distribution. Cross-corpus testing was performed by training on RAVDESS + CREMA-D and evaluating on MELD and IEMOCAP to gauge real-world generalization.

To counter class imbalance (e.g., rare "disgust" samples), we used SpecAugment on audio and synonym replacement on text rather than naive oversampling, which risks overfitting.

Fusion model weights were tuned on validation sets using grid search over confidence weight ranges. Final results reported below use held-out test sets**.**

D. Multimodal Emotion Recognition Accuracy

We evaluated the assistant across three conditions:

- Text-only emotion analysis
- Speech-only emotion recognition
- Fused predictions using valence-weighted logic [22]

Table 2: Accuracy Assessment over different Datasets

Configuration	MELD Accuracy	IEMOCAP Accuracy
Text-only	74.1%	72.3%
Speech-only	70.5%	69.7%
Multimodal Fusion	**81.5%**	**79.4%**

Fusion outperformed unimodal models by 6–9%, especially for overlapping emotions such as surprise vs. fear. These results affirm the conclusions from prior works that multimodal integration enhances emotion robustness with results as shown in Table 2.

E. LLM Output Alignment with Detected Emotion

We computed an **Emotion Coherence Score (ECS)** on LLM-generated outputs. By reclassifying the assistant's replies using the same DistilRoBERTa classifier, we determined whether emotional alignment was preserved.

- **Baseline LLM (no prompting): 63.2% ECS**
- **Emotion-conditioned LLM: 87.6% ECS**

These results echo insights from [14][22], confirming that prompt engineering dramatically improves emotional resonance in language generation.

F. Human Evaluation

In a blind A/B test with 15 annotators, users rated replies on empathy, appropriateness, and fluency shown in Table 3. Emotionally adapted replies were more engaging and reflective of user sentiment a key feature for virtual assistants in sensitive domains

Table 3: Emotional Response with respect to different Prompt Generations

Criterion	No Emotion Prompt	With Emotion Prompt
Empathy	3.1 / 5.0	**4.4 / 5.0**
Appropriateness	3.4 / 5.0	**4.6 / 5.0**
Fluency	3.7 / 5.0	**4.5 / 5.0**

G. *Deployment on Edge and Low-Power Devices*

Finally, we deployed the assistant on three hardware tiers which are shown in Table 4:

- **Laptop (RTX 3060)**
- **NVIDIA Jetson Xavier (6-core CPU + 512-core GPU)**
- **Raspberry Pi 5 + USB Coral Accelerator**

Table 4: Hardware Processing Results

Device	LLM Response Time	Bark TTS Time	Total Latency
RTX Laptop	0.7s	1.2s	1.9s
Jetson Xavier	1.9s	2.4s	4.3s
Raspberry Pi	fallback used	fallback used	6.0s (avg)

With quantized models and ONNX exports [25], real-time interaction (<2s) was feasible on the Jetson, while Pi devices used fallback TTS and inference paths [24].

VI. LIMITATIONS

Technically, our assistant has its limitations too. Without access to visual input like facial expressions or body language, it misses critical emotion cues. Text and voice alone can't capture subtleties like a polite smile or hidden sadness. In noisy environments or when users speak little, the system also struggles: poor audio or ambiguous text inputs weaken its emotion recognition accuracy. Although we built fallback mechanisms, severe input issues still degrade performance.

VII. CONCLUSION

In this paper, *NeuroSentient*, a robust, multimodal AI assistant capable of detecting and responding to human emotion through speech and text under diverse operational conditions has been presented which generates context-aware responses using OpenHermes-2.5-Mistral-7B, with fallback LLMs and retrieval-augmented generation (RAG) to ensure resilience. By integrating modules such as Whisper ASR, Wav2Vec2, XLM-RoBERTa, and Bark TTS into a unified, fallback-resilient pipeline, our assistant achieved 81.5% accuracy in MELD dataset and 79.4% accuracy IEMOCAP dataset which resulted in emotional understanding and empathy in dialogue even on low-resource devices or in offline environments.

Key contribution includes valence-weighted fusion model that dynamically balances vocal and textual emotional cues, emotion-conditioned prompting for LLMs, significantly improving empathy in responses, fully modular architecture with fallback logic and hardware-aware optimization, extensive experimental validation across five datasets, multiple hardware tiers, and both objective and human evaluations.

This work demonstrates how combining state-of-the-art NLP and speech models with intelligent orchestration can produce emotionally aware conversational agents suited for mental health, education, and other sensitive domains. Future directions include multimodal reinforcement learning for emotion alignment, privacy-enhanced training on local devices, and integration of visual signals like facial expressions.

We aim to optimize the system for on-device deployment, protecting user privacy while maintaining real-time performance through techniques like model compression and federated learning.

REFERENCES

[1] S. Poria et al., "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," ACL 2019.

[2] Explosion AI, "spaCy: Industrial-strength NLP," [Online]. Available: https://spacy.io

[3] Suno AI, "Bark: Text-to-Speech with Expressive Voice," 2023. [Online]. Available: https://github.com/suno-ai/bark

[4] OpenAI, "Whisper: Robust Speech Recognition," 2022. [Online]. Available: https://github.com/openai/whisper

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," NeurIPS 2020.

[6] S. Yang et al., "SUPERB: Speech processing Universal PERformance Benchmark," Interspeech 2021.

[7] HuggingFace, "facebook/wav2vec2-base-superb-er," [Online]. Available: https://huggingface.co/facebook/wav2vec2-base-superb-er

[8] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692.

[9] I. Barbieri et al., "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," Findings of EMNLP 2020.

[10] Cardiff NLP, "twitter-xlm-roberta-base-sentiment," [Online]. Available: https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment.

[11] Z. Akçay and E. Oguz, "Speech Emotion Recognition: Datasets, Features, and Methods," Speech Communication, 2020.

[12] Wu, Qingyang, et al. "Memformer: A memory-augmented transformer for sequence modeling." arXiv preprint arXiv:2010.06891 (2020).

[13] A. Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," EMNLP 2017.

[14] Silero, "Silero VAD: Lightweight Voice Activity Detection," [Online]. Available: https://github.com/snakers4/silero-vad

[15] J. Zhao et al., "EmotionPrompt: Prompt-based Emotion-Aware Language Modeling," arXiv preprint, 2022.

[16] OpenAccess AI Collective, "OpenHermes-2.5-Mistral-7B," 2023. [Online]. Available: https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

[17] Mistral AI, "Mistral-7B," [Online]. Available: https://mistral.ai

[18] H. Rashkin, E. Smith, M. Li, and Y. Boureau, "Towards Empathetic Open-domain Conversation Models," ACL 2019.

[19] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, 2018

[20] Suno AI, “Bark: Text-to-Speech with Expressive Voice,” 2023. [Online]. Available: https://github.com/suno-ai/bark

[21] Google, “gTTS: Google Text-to-Speech Python Library,” [Online]. Available: https://pypi.org/project/gTTS

[22] Coqui AI, “Coqui-TTS,” [Online]. Available: https://github.com/coqui-ai/TTS

[23] N. Rajbhandari et al., “Zero Redundancy Optimizer for Memory-Efficient Training,” NeurIPS 2021.

[24] PyAnnote Audio, “Speaker Diarization Toolkit,” [Online]. Available: https://github.com/pyannote/pyannote-audio

[25] Süzen, Ahmet Ali, Burhan Duman, and Betül Şen. "Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-cnn." 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2020.

[26] D. Gunning and D. Aha, “DARPA's Explainable Artificial Intelligence Program,” AI Magazine, vol. 40, no. 2, 2019.

[27] Y.-H. H. Tsai et al., “Multimodal Transformer for Unaligned Multimodal Language Sequences,” ACL 2019.

[28] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS 2020.

[29] ONNX Runtime, “Accelerate ML Model Inference,” [Online]. Available: https://onnxruntime.ai

[30] J. Hartmann et al., “Emotion English DistilRoBERTa,” HuggingFace Model, [Online]. Available: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[31] C. Busso et al., “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” Language Resources and Evaluation, 2008.