Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation

Qiongqiong Wang[1], *Hardik B. Sailor*[1], *Tianchi Liu*[1], *Ai Ti Aw*[1]

[1]Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research (A⋆STAR), Singapore

{wang_qiongqiong, sailor_hardik_bhupendra, liu_tianchi, aaiti}@i2r.a-star.edu.sg

Abstract

Current speech-LLMs exhibit limited capability in contextual reasoning alongside paralinguistic understanding, primarily due to the lack of Question-Answer (QA) datasets that cover both aspects. We propose a novel framework for dataset generation from in-the-wild speech data, that integrates contextual reasoning with paralinguistic information. It consists of a pseudo paralinguistic label-based data condensation of in-the-wild speech and LLM-based Contextual Paralinguistic QA (CPQA) generation. The effectiveness is validated by a strong correlation in evaluations of the Qwen2-Audio-7B-Instruct model on a dataset created by our framework and human-generated CPQA dataset. The results also reveal the speech-LLM's limitations in handling empathetic reasoning tasks, highlighting the need for such datasets and more robust models. The proposed framework is first of its kind and has potential in training more robust speech-LLMs with paralinguistic reasoning capabilities.

Index Terms: Speech-LLM, Paralinguistic, empathetic, data condensation, Spoken QA generation, Contextual reasoning

1. Introduction

Rapid advancements in large language models (LLMs) have sparked significant interest in multimodal models that integrate LLMs with speech modalities. Recent speech-LLMs, such as GPT-4 [1], Qwen-audio [2, 3], SALMONN [4], and MERaLiON-AudioLLM [5, 6], have demonstrated remarkable performance in handling speech-based tasks. Some speech-LLMs, in particular, focus on contextual reasoning properties derived from speech [7–11].

Several studies have attempted to train models to understand emotions in speech and respond empathetically [7, 10, 12, 13]. Notable approaches are presented in [7] and [13], where the authors introduced training strategies to enhance QA performance by incorporating paralinguistic information from existing speech emotion datasets. These models, however, exhibit limited capabilities in contextual reasoning alongside paralinguistic understanding, primarily due to the lack of QA datasets that cover both aspects. To incorporate paralinguistic cues, QA generation must extend beyond linguistic features. We refer to such QA as contextual paralinguistic QA (CPQA).

Creating contextual paralinguistic data presents two major challenges and significantly hinders progress in this area. First, the availability of relevant metadata is limited. Publicly accessible speech datasets with paralinguistic labels are typically small and task-specific. Emotion-labeled data is even rarer than attributes like speaker identities or gender. While vast amounts of speech data exist, obtaining well-annotated data with reliable paralinguistic metadata remains difficult. Unlike textual data, speech annotation is more complex, as it requires both transcriptions and precise labeling of paralinguistic features. Some paralinguistic labels, such as speaker or gender, are relatively objective, while others, such as emotion, are particularly challenging due to their subjective nature. Emotion labeling requires multiple annotators and majority voting, with low-agreement samples often discarded.

Second, CPQA generation is non-trivial. High-quality QA requires that questions comprehensively cover all relevant aspects of performance and that the reference answers are both accurate and unbiased. Existing benchmarks, such as AudioBench [14] and Dynamic-Superb [15], AIR-Bench [16], OpenASQA [17] and MMAU [18] evaluate not only speech understanding but also paralinguistic tasks, including emotion recognition. However, the QA in these benchmarks primarily derives from speech emotion datasets, such as IEMOCAP [19] and MELD [20], which focus on isolated emotion-related tasks. These QA datasets typically frame QA pairs in a direct manner (e.g., explicitly asking for an emotion label) without incorporating contextual reasoning.

To address these challenges, we propose a novel framework for dataset generation from in-the-wild speech data. The framework consists of data condensation and automated CPQA generation, specifically focusing on emotion aspect. To the best of our knowledge, this work is the first to integrate contextual reasoning with paralinguistic cues. The proposed framework supports both training and evaluation data creation for speech-LLMs. In this paper, we validate the framework by generating an evaluation dataset and comparing it to human-generated QA sets in LLM evaluations. Our contributions are as follows:

- Speech Data Condensation: We address the challenge of limited and noisy paralinguistic labels by integrating categorical and dimensional emotion recognition models, resulting in more accurate pseudo emotion labels.
- CPQA Generation via LLM: We automate the generation of QA pairs using LLMs to capture both paralinguistic cues and contextual reasoning. These generated pairs are compared with human-generated sets in speech-LLM evaluations. This approach has the potential to generate large-scale training datasets for speech-LLMs.
- Open-Source Evaluation Dataset: We provide an open-source benchmark comprising 480 speech samples paired with CPQA pairs.[1]

2. Proposed data generation framework

We propose a novel framework for CPQA dataset generation from in-the-wild speech data. The framework consists of data

[1]https://huggingface.co/datasets/MERaLiON/CPQA-Evaluation-Set

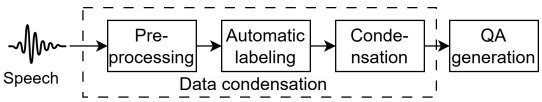

Figure 1: *Diagram of dataset creation framework.*

condensation and automated CPQA generation (see Figure 1).

2.1. Data condensation pipeline

Emotion annotation is challenging due to its subjective nature [21, 22], requiring multiple annotators and majority voting for reliability, which makes it resource-intensive. Additionally, most in-the-wild speech data predominantly exhibit a neutral emotional tone. An unbalanced distribution skewed toward the "neutral" category would limit the development of an effective dataset for empathetic speech-LLMs, as it would fail to provide meaningful insights for emotion-related tasks.

To address these challenges, we propose a data condensation pipeline. First, we employ automatic speech emotion recognition (SER) tools to avoid the high cost of human annotation. The limitations of these tools in SER accuracy of spontaneous speech [23] are mitigated through a condensation technique. This technique filters out low-confidence samples and ensures a balanced distribution across emotion categories, which is crucial for maintaining fairness, robustness, and generalizability within the dataset.

2.1.1. Pre-processing

Speech samples often contain silence or noise and vary in length, requiring pre-processing for SER. We use voice activity detection to remove non-speech, retaining segments $S = \{s_1, s_2, \ldots, s_N\}$. Each s_i is then segmented into sub-segments $s_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,M_i}\}$ of $t+2\Delta t$ windows with $2\Delta t$ overlaps. As shown in Figure 2, SER is applied on $s_{i,j}$ and assign an emotion label $e_{i,j}$ to the middle t of $s_{i,j}$, considering the contexts of Δt before and afterward.

2.1.2. Automatic emotion label annotation

Emotion recognition follows two paradigms: discrete emotion categories (e.g., happy, angry, sad, neutral) and dimensional representations(valence, arousal, dominance) [24]. Valence is closely related to sentiment. while discrete emotions are intuitive, they struggle with mixed emotions and data scarcity for certain classes. Dimensional SER models provide a broader view but lack of interpretability. A hybrid approach combining both paradigms improves SER performance by improving prediction confidence and enabling inference beyond predefined categories, resulting in a more flexible and robust system. For simplicity, we focus on the valence and, inspired by the sentiment dictionary [25], define a mapping rule between sentiment classes and the emotion categories based on valence values.

For categorical SER, we employ the state-of-the-art Emotion2Vec pipeline [26] with model ensembling for emotion labeling. It classifies emotions into nine categories: *angry, disgusted, fearful, happy, neutral, other, sad, surprised*, and *unknown*. The pipeline has shown strong performance across multiple public datasets[2]. For dimensional SER, we utilize a model[3] [27] that was created by fine-tuning the pre-trained

[2]https://github.com/ddlBoJack/emotion2vec

[3]Model: https://doi.org/10.5281/zenodo.6221127

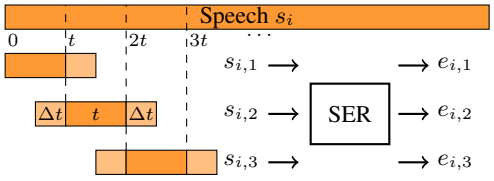

Figure 2: *Illustration of speech segmentation. SER is performed on segments $s_{i,j}$ of length $t + 2\Delta t$ and creates emotion labels $e_{i,j}$ for segments of length t.*

Algorithm 1 Filters for Data Condensation

Input: $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$: N speech samples
$\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}, \mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$
  where $\mathcal{C}_i = \{c_{i,1}, \ldots, c_{i,M_i}\}$ *(categorical labels)*
  and $\mathcal{D}_i = \{d_{i,1}, \ldots, d_{i,M_i}\}$ *(dimensional values)*
$\mathcal{Z}$: set of emotion categories
X_E, X_O: filter conditions (consistency, occurrence)
Output: $\widetilde{\mathcal{S}} \subseteq \mathcal{S}$: condensed set with reliable labels
$\widetilde{\mathcal{S}} \leftarrow \emptyset$
for $i \leftarrow 1$ to N **do**
  for $j \leftarrow 1$ to M_i **do** ▷ 1) Sub-segment Consistency
    if $(c_{i,j}, d_{i,j})$ is consistent w.r.t. X_E **then**
      $c_{i,j} \leftarrow c_{i,j}$
    else
      $c_{i,j} \leftarrow$ "unknown"
    end if
  end for
  for all $Z_k \in \mathcal{Z}$ **do** ▷ 2) Label Assignment
    if $|\{c_{i,j} == Z_k\}| \in X_O(Z_k)$ **then**
      label $\mathbf{s}_i$ as Z_k
      $\widetilde{\mathcal{S}} \leftarrow \widetilde{\mathcal{S}} \cup \{\mathbf{s}_i\}$
    end if
  end for
end for
return $\widetilde{\mathcal{S}}$

wav2vec2-large-robust model[4] on MSP-Podcast (v1.7) [28]. To map categorical emotions to dimensional emotion, we group the seven categories from Emotion2Vec (excluding the "other" and "unknown") into four sentiment classes: *positive* Y_{pos}, *negative* Y_{neg}, *neutral* Y_{neu}, and *ambiguous* Y_{amb}:

$$Y_{\text{pos}} = \{\textit{happy}\} \quad Y_{\text{neu}} = \{\textit{neutral}\} \quad Y_{\text{amb}} = \{\textit{surprised}\}$$
$$Y_{\text{neg}} = \{\textit{angry, disgusted, fearful, sad}\} \tag{1}$$

To be noted, these mappings are not exhaustive, and there is potential to explore additional emotion categories and dimensions. In addition, we also annotated speaker gender to include in our QA generation task using WavLM-ECAPA[5] model [29,30] fine-tuned on the VoxCeleb2 dataset [31].

2.1.3. Data condensation

To ensure meaningful CPQA pairs with sufficient reasoning context, we first filter the dataset based on audio length, discarding speech segments shorter than a pre-determined threshold τ. Next, we condense speech data using automatically estimated discrete emotion classes and valence values in emotional dimensions. The condensation process involves SER consistency

[4]https://huggingface.co/facebook/wav2vec2-large-robust

[5]https://github.com/wenet-e2e/wespeaker

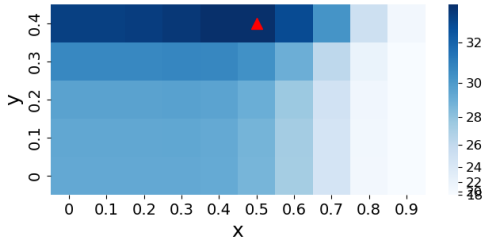

Figure 3: *Heatmap of UWA(%) performance on the SG TV/movie dataset for varying factors x and y. The triangle marks the best performance.*

filtering and occurrence filtering (see Algorithm 1), applied under the following conditions:

- SER consistency condition X_E: Ensures consistency between sentiment class mapped from discrete emotion categories and valence values. A sub-segment $s_{i,j}$ satisfies X_E, a set of the following:

$$c_{i,j} \in Y_{\text{pos}} \quad \wedge \quad v_{i,j} \geq v_{\text{pos,min}} \tag{2}$$

$$c_{i,j} \in Y_{\text{neg}} \quad \wedge \quad v_{i,j} \leq v_{\text{neg,max}} \tag{3}$$

$$c_{i,j} \in Y_{\text{neu}} \quad \wedge \quad v_{\text{neu,min}} \leq v_{i,j} \leq v_{\text{neu,max}} \tag{4}$$

$$c_{i,j} \in Y_{\text{amb}} \tag{5}$$

  Here, $c_{i,j}$ and $v_{i,j}$ denote the estimated emotion category and valence for sub-segment $s_{i,j}$. Thresholds $\{v_{\text{neu,min}}, v_{\text{neu,max}}, v_{\text{pos,min}}, v_{\text{neg,max}}\}$ are set empirically or optimized based on labeled datasets.
- Occurrence condition X_{O}: Filters a subset that has greater confidence in speech s_i belongs to a specific emotion category by applying a threshold α on the number of labels of that emotion within s_i. Due to the class imbalance, α varies across emotions. When $\alpha = 1$, the filter is disabled, retaining all segments.

2.2. Automated speech QA generation

To use text LLMs for QA generation, we need to provide time-aligned speech transcript and paralinguistic metadata information. We utilized WhisperX to generate word-level alignments [32]. To align the emotion and gender metadata with the word-level transcript, each word is matched to the corresponding paralinguistic segment based on its timestamp. The process includes temporal overlap-based matching: Each word's start and end time is compared with the time intervals of emotion and gender annotations. If a word falls within or overlaps with a paralinguistic segment, it inherits the corresponding emotion label and gender label. This alignment ensures that each word is enriched with affective and speaker identity cues, making it valuable for emotion-aware speech processing, speaker profiling, and multimodal sentiment analysis. After alignment, we used specific prompt shown below to generate QA pair. The QA generator LLM used in our evaluation is GPT4o API (2024-07-01-preview version)[6] from Azure. We use the prompt in Figure 4 for the QA generation.

3. Evaluation dataset creation

We construct a dataset by applying our proposed data creation framework to speech data collected from top Singaporean YouTube channels.

[6]https://learn.microsoft.com/en-us/azure/ai-services/openai/

Prompt for Generating QA Pairs from Audio Clips
Generate diverse paralinguistic, content-based, and contextual reasoning QA pairs from a given audio clip. These QA pairs will train multimodal LLMs to reason from audio and text without relying on metadata.
Guidelines:

1. **Focus Areas:**
   - Explore speaker attributes such as emotion, gender, speaker transitions, and emotional reasoning.
   - Include content-based questions, balancing simple and complex reasoning.
2. **Multimodal Integration:**
   - Use both emotion labels and transcriptions, refining questions when labels are inaccurate.
   - Ensure reliance on audio features without assuming metadata availability.
3. **Diversity and Depth:**
   - Avoid repetitive or overly narrow questions.
   - Encourage deeper multimodal reasoning.
4. **Output Format:**
   - Format each QA pair using `Q:` and `A:` tags.

Inputs: Utterance: `'{utterance}'`, Word-level paralinguistics data: {word_level_data}.

Figure 4: *Prompt for Generating QA Pairs from Audio Clips*

3.1. Preliminary study for framework parameters

To set framework parameters, we conduct a preliminary study on an internal emotion dataset, SG TV/Movie dataset, comprising 117k speech segments (120 hours) from Singaporean TV shows and movies, primarily in English with some Mandarin. The emotion labels are annotated by human annotators.

In pre-processing, we set $t = 2$ sec and $\Delta t = 1$ sec for automatic emotion labeling by SER. Additionally, gender labels were obtained using $t = 2$ sec and $\Delta t = 0.5$ sec in the same manner stated in Section 2.1.1. For emotion labeling, we evaluated the Emotion2vec models on the SG TV/Movie dataset, excluding *embarrassment*, *sarcasm*, and *worry* from the dataset's original ten emotion classes for compatibility. Zero-shot inference showed that the *emotion2vec base* model exhibited the lowest performance among four individual models. The best performance was achieved using an ensemble of three *emotion2vec+* models [26] by averaging their posteriors. It achieved accuracy of 51.10%, unweighted accuracy (UWA) of 29.25%, and F1-score of 49.56%. Thus, it was selected for categorical SER.

We apply length filtering with a minimum duration of $\tau = 30$ sec. For sub-segment SER consistency filtering, we explore UWA across different valence thresholds. Unlike accuracy, UWA mitigates selection bias toward dominant categories, addressing the varying difficulty of emotion categories in SER. Thresholds in X_E are set as $v_{\text{pos,min}} = x, v_{\text{neg, max}} = 1 - x, v_{\text{neu, min}} = y$, and $v_{\text{neu,max}} = 1 - y$. Figure 3 shows the highest UWA 33.65% is achieved at $x = 0.5$ and $y = 0.4$

Table 1: *Comparison of Human- and ChatGPT-generated QA.*

	Human	LLM
Total QA Pairs	2184	2647
Questions about Emotion	150	850
Questions about Speakers	484	228
Contextual Paralinguistic Reasoning Questions	1530	1390
Others (content, topics, etc.)	20	179

on the SG TV/movie dataset. For occurrence filtering, we set $\alpha = [10, 10, 4, 4, 2, 3]$ for the six non-neutral emotion categories: *angry, disgusted, fearful, happy, sad*, and *surprised*, to ensure sufficient and balanced samples are retained for each category in the condensed dataset.

3.2. CPQA Evaluation dataset

We randomly selected 80 samples for each emotion category from the condensed data to construct a balanced evaluation dataset. This resulted in a total of 480 audio samples, each ranging from 30 to 60 sec, corresponding to approximately 6.5 hours of data. From these samples, we generated 2,647 CPQA pairs with the ChatGPT LLM. For comparison, two human annotators created a CPQA set for the same data, with the objective of involving both paralinguistics and content for reasoning.

Table 1 shows statistics of the two QA sets. Questions about emotion include the speaker's emotional state or feelings, while questions about the speaker focus on attributes like gender and number. Questions regarding contextual paralinguistic reasoning explore the underlying causes of a speaker's emotions and feelings. Other questions include content, topics, relationship between speakers, etc. During a manual review of ChatGPT QA, we observed repetitive questions, for example, multiple variations asking about the reasons behind emotions or changes in emotional state to increase diversity. We also encountered some irrelevant questions that assumed the existence of a text transcript (e.g., "What is the content in the audio from the text transcript?"). These questions were removed by post-filtering for keywords such as "text" or "transcript".

Emotion-related question-answer pairs are notably more frequent, while contextual reasoning and content-related questions are slightly less common in the model-generated set compared to human-generated one. As a result, out of 850 QA pairs, many questions belong to the CPQA section or associated with other attributes, such as gender and speaker identity. The LLM also generates general questions about content and topics, etc. Since the human annotators are explicitly instructed to focus on paralinguistic aspects, only 20 non-paralinguistic questions in their set, compared to 179 when using ChatGPT. Although this diversity in QA is valuable, the LLM generated fewer speaker-related questions due to its inability to process audio with multiple speakers of the same gender. To address this, future work will incorporate speaker diarization.

4. Evaluation

We validate the ChatGPT-generated CPQA set by evaluating Qwen2-Audio-7B-Instruct[7] speech-LLM since it is the best performing open source model as shown in large scale MMAU evaluation [18]. To interpret the performance, we use Llama-3-70B-Instruct-AWQ[8] and ChatGPT-4o (2025-01-29 version)[9] as two judger LLMs. Qwen2 [3] and Llama-3 [33] are deployed using a single NVIDIA H100 GPU (80GB). Notably, the QA generation and performance judgers use text-LLMs, while the evaluation uses a speech-LLM (see Figure 5).

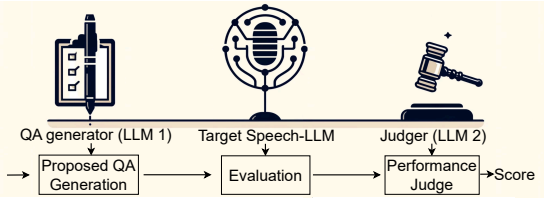

Figure 5: *The illustration of the evaluation pipeline.*

Table 2: *Zero-shot evaluation of CPQA sets.*

Judge	Prompt	LLM QA	Human QA
Llama 70B	Prompt 1	53.86	52.29
	Prompt 2	56.82	54.33
ChatGPT	Prompt 1	59.64	56.59
	Prompt 2	60.28	59.46

Two evaluation prompts are employed for judgment. one from AudioBench [14] that focuses on content accuracy and relevance and a refined version (Prompt 2). Prompt 2 incorporates both speech content and paralinguistic information and removes the penalty for brevity. Since Qwen2 processes audios up to 30 sec, we evaluate both the first and the last 30 sec of each segment and use the higher score for the QA pair. The final evaluation performance is the average score across all QAs.

The results in Table 2 indicate that the Qwen2 model demonstrates comparable performance on both LLM- and human-generated QA sets. This finding provides evidence, to some extent, that LLM-generated QA can serve as a viable tool for evaluating speech-LLMs. Furthermore, Prompt 2, which has been refined to align more closely with the contextual paralinguistic task, is expected to enhance the model's ability to assess evaluation performance more effectively. The consistent improvement is observed for both QA sets when using Prompt 2. Additionally, the observed correlation between the two QA sets further supports the validity of LLM-generated CPQA as a reasonable approach to evaluate speech-LLMs.

5. Summary

We propose a novel framework for generating dataset with contextual paralinguistic QA (CPQA) pairs from in-the-wild speech data, addressing the scarcity of data available for developing empathetic speech-LLMs. Our framework consists of pseudo paralinguistic label-based data condensation and LLM-based CPQA generation. We release a benchmark dataset comprising 480 speech samples. Evaluation using the Qwen2-Audio-7B-Instruct model, alongside comparisons with a human-generated set, demonstrates both the effectiveness of our dataset and the limitations of current speech-LLMs in empathetic reasoning. Our dataset provides a benchmark for evaluating speech-LLMs' contextual paralinguistic reasoning capabilities. Future work will explore more comprehensive assessments of paralinguistic contextual reasoning in speech-LLMs.

[7] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[8] https://huggingface.co/casperhansen/llama-3-70b-instruct-awq

[9] https://help.openai.com/en/articles/9624314-model-release-notes#

6. Acknowledgement

This research/project is supported by the National Research Foundation, Singapore, under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

7. References

[1] J. Achiama, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[3] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[4] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[5] Y. He, Z. Liu, S. Sun, B. Wang, W. Zhang, X. Zou, N. F. Chen, and A. T. Aw, "MERaLiON-AudioLLM: Technical report," *arXiv preprint arXiv:2412.09818*, 2024.

[6] M. Huzaifah, T. Liu, H. B. Sailor, K. M. Tan, T. K. Vangani, Q. Wang, J. H. Wong, N. F. Chen, and A. T. Aw, "Towards a speech foundation model for singapore and beyond," *arXiv preprint arXiv:2412.11538*, 2024.

[7] C. Wang, M. Liao, Z. Huang, J. Wu, C. Zong, and J. Zhang, "BLSP-Emo: Towards empathetic large speech-language models," *arXiv preprint arXiv:2406.03872*, 2024.

[8] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, and other, "AudioPaLM: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[9] Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng *et al.*, "LauraGPT: Listen, attend, understand, and regenerate audio with GPT," *arXiv preprint arXiv:2310.04673*, 2023.

[10] G.-T. Lin, C.-H. Chiang, and H.-Y. Lee, "Advancing large language models to capture varied speaking styles and respond properly in spoken conversations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 6626–6642.

[11] C. Wang, M. Liao, Z. Huang, J. Lu, J. Wu, Y. Liu, C. Zong, and J. Zhang, "BLSP: Bootstrapping language-speech pre-training via behavior alignment of continuation writing," *arXiv preprint arXiv:2309.00916*, 2024.

[12] H. Kim, S. Seo, K. Jeong, O. Kwon, S. Kim, J. Kim, J. Lee, E. Song, M. Oh, J.-W. Ha *et al.*, "Paralinguistics-aware speech-empowered large language models for natural conversation," in *Neural Information Processing Systems (NeurIPS)*, 2024.

[13] W. Kang, J. Jia, C. Wu, W. Zhou, E. Lakomkin, Y. Gaur, L. Sari, S. Kim, K. Li, J. Mahadeokar *et al.*, "Frozen large language models can perceive paralinguistic aspects of speech," *arXiv preprint arXiv:2410.01162*, 2024.

[14] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "AudioBench: A universal benchmark for audio large language models," *NAACL*, 2025.

[15] C.-Y. Huang, K.-H. Lu, S.-H. Wang, C.-Y. Hsiao, C.-Y. Kuan, H. Wu, S. Arora, K.-W. Chang *et al.*, "Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 136–12 140.

[16] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou, and J. Zhou, "AIR-bench: Benchmarking large audio-language models via generative comprehension," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 1979–1998.

[17] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," in *International Conference on Learning Representations (ICLR)*, 2024.

[18] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *International Conference on Learning Representations (ICLR)*, 2025.

[19] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[20] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[21] A. Ortony and T. Turner, "What's basic about basic emotions?" *Psychological review*, pp. 315–331, 1990.

[22] R. Plutchik and H. Kellerman, "Theories of emotion," *Academic Press*, 2013.

[23] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Proc. Interspeech*, 2024, pp. 1580–1584.

[24] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[25] D. Demszky, D. Movshovitz-Attias, J. Ko, A. S. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Annual Meeting of the Association for Computational Linguistics*, 2020.

[26] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[27] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[28] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrievingemotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[32] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in *Proc. Interspeech*, 2023.

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The Llama 3 herd of models," *CoRR*, 2024.