



Can AI Understand Mandarin Speech Prosody? A Framework and Benchmark Showcase

Zilong Wang¹, Xiaoxue Zhang², Xinyang Jiang¹, Kaitao Song¹, Jue Yu²

¹Microsoft Research Asia, Microsoft, China

²School of Foreign Studies, Tongji University, China

wangzilong@microsoft.com, Lexiel9836@163.com, xinyangjiang, kaitaosong@microsoft.com,
erinyu@tongji.edu.cn

Abstract

How to model and estimate speech prosody is considered as a challenging task in understanding and generating natural speech. We introduce the Mandarin Speech Prosody Benchmark (MSPB), a linguistically grounded dataset for evaluating Speech Large Language Models (Speech LLMs) in Mandarin. MSPB comprises eight tasks covering crucial prosodic features and their interactions with syntax, semantics, and pragmatics. All MSPB items, designed per Mandarin linguistic principles and validated by experts, were phonetically recorded and verified. We evaluated six Speech LLMs (GPT-4o, Gemini-1.5-Pro, Gemini-2-Flash, Qwen2-Audio-7B-Instruct, GLM-4-Voice, MiniCPM-o 2.6). Although some models perform well with context-rich cues (e.g., irony), they generally struggle with subtle prosodic variations (e.g., focus marking) and underperform humans. MSPB provides a valuable tool to assess and enhance prosodic comprehension, underscoring the need for improved prosodic integration in future research.

Index Terms: Speech LLMs, prosody, Mandarin, Benchmark

1. Introduction

Recent advancements in Large Language Models (LLMs) have driven significant progress in human-computer interaction, facilitating the development of sophisticated speech-based chatbot applications [1, 2, 3]. While traditional speech-based chatbot interfaces rely on a pipeline architecture consisting of “Automatic Speech Recognition (ASR) + LLM + Text-to-Speech (TTS),” recent research efforts have increasingly focused on developing end-to-end Speech Large Language Models (Speech LLMs) capable of directly processing audio input [4, 5].

This paradigm shift mitigates several limitations of the conventional approach, including information loss during modality conversion between speech and text, significant latency due to the pipeline’s inherent complexity, and the accumulation of errors across the three processing stages. Moreover, it potentially enhances the ability to capture crucial prosodic information embedded in speech signals, which is particularly relevant for nuanced communication. However, the ability of current Speech LLMs to effectively comprehend and utilize prosodic information remains under thorough examination and warrants careful scrutiny.

In fact, speech prosody, encompassing features such as tone, intonation, rhythm, stress, and pauses, plays a crucial role in human communication [6, 7, 8]. Prosody enables a wide range of communicative functions, including word segmentation [9, 10], grammatical relationship identification [11, 12], information structure marking [13, 14], and the conveyance of pragmatic nuances [15, 16]. For instance, rising intonation can signal a question, pauses can differentiate statements from lists

(e.g., “chocolate milk and apple” vs. “chocolate, milk, and apple”), and stress can drastically alter sentence implication (e.g., “I didn’t say SHE stole my money” vs. “I didn’t say she stole MY money”). Humans readily perceive these subtle acoustic variations [17, 18], leveraging prosody to ensure successful verbal communication. Therefore, robust comprehension of speech prosody is not merely desirable but a challenging and critical task for achieving natural and effective human-computer interaction.

Despite the evident importance of prosody, existing holistic audio benchmarks—such as AIR-Bench, and Audiobench [19, 20]—which cover a wide range of speech-related tasks, such as ASR, speaker number verification, language identification, speech QA, audio QA, as well as emotion, accent, and gender recognition, still fall short in offering a comprehensive evaluation of prosodic understanding. While several paralinguistic benchmark datasets, such as SD-Eval, StyleTalk, and SALMON, exist, their primary focus is on aspects like emotion, accent, age, environmental understanding, and speaking styles rather than prosody [21, 22, 23]. These benchmarks primarily focus on assessing linguistic content, paralinguistic features (e.g., emotion, speaker identity), and non-speech sounds. Crucially, they often neglect the pragmatic implications conveyed by prosodic cues, which are paramount for accurately discerning speaker intent, especially in interactive dialogues. Furthermore, many current benchmarks overlook language-specific prosodic profiles. This is a significant oversight, as prosodic realization varies considerably across languages. For example, Mandarin Chinese presents a particularly complex interplay between tone and intonation [24, 25], a characteristic that distinguishes it from many Western languages. These limitations in current evaluation methodologies hinder the accurate diagnosis of Speech LLM weaknesses, especially when processing speech in complex, real-world audio environments.

To address this critical gap in evaluation resources, we introduce the Mandarin Speech Prosody Benchmark (MSPB). MSPB is a linguistically grounded benchmark specifically designed to rigorously evaluate the ability of Speech LLMs to comprehend various facets of Mandarin prosody. By systematically examining the strengths and limitations of Speech LLMs in processing these intricate prosodic features in Mandarin Chinese, MSPB aims to provide a more comprehensive and nuanced evaluation framework for the field. We believe that MSPB will drive the development of advanced speech interaction systems that are capable of better interpreting speech with complex prosodic features, ultimately leading to more natural, efficient, and human-centered machine interactions.

Table 1: Prosodic tasks and examples

Type	Task	Number	Example
Intonation	Identify statements from interrogatives by pitch variations	40	<i>a. ni xiang chi xiangjiao</i> ↘ 1 ↗ 2 [1: You want to eat bananas.] vs. [2: Do you want to eat bananas?]
Prosodic disambiguation	Resolve syntactic ambiguities using prosodic cues	40	<i>b. hua_{#1} sheng_{#2} zhang zai wuhou de yuanzi li</i> [1: Flowers grow in the yard behind the house.] vs. [2: Peanuts grow in the yard behind the house.]
Prosodic focus marking	Identify information structure using prosodic cues	24	<i>c. Zhang Zhongbin₁ xingqitian₂ xiushouyinji.</i> [1: Zhang Zhongbin repaired the radio on Sunday.] vs. [2: Zhang Zhongbin repaired the radio on Sunday .]
Focus operator	Map prosodic structure to syntactic focus operators	14	<i>d. wo mingtian (shi₁) chenghuoche (shi₂) qu guangzhou.</i> [1: It is by train that I will go to Guangzhou tomorrow.] vs. [2: It is Guangzhou that I'm going to by train tomorrow.]
Scalar	Map prosodic cues to scalar implication	10	<i>e. yixie₁ xiaogou zai chi gutou₂.</i> [1: It is some puppies that are eating bones.] vs. [2: It is bones that some puppies are eating.]
Irony	Identify ironic implication using both prosodic cues and contextual information	10	<i>f. wo zhen gaoxing ni neng baimang zhong choukong lai jian women.</i> [Xiaolin promised to bring a salad to Xiaodong's dinner party. But she arrived two hours late. Xiaodong said, "I'm so glad that you were able to fit us into your busy schedule."]
Emotional prosody with context	Identify emotional states using both prosodic cues and contextual information	8	<i>g. Wo bugan xiangxin zhe shi zhen de.</i> [Happiness : I am the first in this exam?! I can't believe this is true.] vs. [Sadness : I failed the exam again. I can't believe this is true.]
Emotional prosody without context	Identify emotional states using prosodic cues	32	<i>h. Songqing zhu zai daqing shi.</i> [Songqing lives in Daqing city.]

2. Benchmark Dataset

2.1. Overview of Mandarin Speech Prosody Benchmark

The Mandarin speech prosody benchmark (MSPB) altogether comprises eight prosodic tasks: intonation, prosodic disambiguation, prosodic focus marking, focus operator, scalar, irony and emotional prosody with or without context. These tasks were selected to cover key aspects of speech prosody that interact with syntax, semantics, and pragmatics in Mandarin Chinese. The details of each task are displayed in Table 1.

Each task requires the models to accurately identify or map prosodic cues to other linguistic cues (such as context, syntactic structure, information structure) as well as paralinguistic cues (such as emotion) when interpreting the utterance. Intonation task use 2-alternative forced choice, while others use 4-alternative forced choice paradigm.

These test items were designed by the authors based on previous studies[26, 27, 28], and in accordance with Mandarin linguistic conventions and principles. They were further reviewed for linguistic validity and clarity through expert consultation and interviews with 20 native Mandarin speakers.

2.2. Audio Dataset Construction

As for the audio stimuli, a female Mandarin phonetician, seated in front of a laptop, with a Sennheiser head-worn microphone situated approximately 2 cm from her mouth, was instructed to produce each test item 3 times accurately and clearly. All recordings were collected under the control of a custom MATLAB program in a sound booth at the phonetics lab and were recorded directly onto a hard disk of the laptop with a 16-bit quantization rate and 44.1 kHz sampling rate. The audio stimuli were then phonetically checked using Praat[29] to ensure that each test item was sufficiently realized in term of prosodic cues and the most valid one was chosen. Finally, altogether 178

test stimuli were confirmed.

3. Experiments and Results

3.1. Experiments on Human Participants

3.1.1. Experimental Setup

In total 27 human participants (26F/1M, 18-53 years-old) were enrolled in current study. The experiment was conducted in a quiet room using e-prime v.3.0, and participants wore closed-ear headphones. At the beginning of the test, they were instructed to read the written instructions carefully on the computer screen until they understand their task, after which they could press any button to begin. During the test, participants first heard an auditorily-presented utterance once while focusing on a fixation point in the middle of the screen. They were then shown the multiple-choice options and asked to select the one that best matched their own comprehension by pressing the corresponding button. There was no time limit for the test, as the main goal was to measure response accuracy. The whole experiment consisted of 5 blocks (1: intonation; 2: prosodic disambiguation; 3: prosodic focus marking, focus operator and scalar; 4: irony; 5: emotional prosody with or without context) with a compulsory break of at least 10 seconds between blocks. Both the order of blocks and the order of stimuli within each block were randomized. The entire experiment lasted approximately 25 minutes.

3.1.2. Results of Human Participants

The general accuracy performance of human participants is presented in Figure 1 across different tasks. It is evident that on average, humans face little difficulty in identifying or mapping prosodic cues to other linguistic information, as their accuracy (81.01%-95.67%) is significantly higher than that of most

Speech LLMs. Notably, most participants achieved near-perfect performance in the emotional prosody with context (95.67%) and intonation (95.29%) tasks, with a clear ceiling effect. This may be due to the abundant cues (both contextual and prosodic) available in the first task, while the second one benefits from the robustness of intonation in everyday communication. On the other hand, they showed relatively lower accuracy (81.01%) in the emotional prosody without context task, suggesting that prosody alone may not be sufficient for humans to identify emotional states. However, this result was still far better than most models, showing that human participants were highly successful in interpreting various prosodic cues.

3.2. Benchmark on Speech LLMs

3.2.1. Experimental Setup

Models: We assessed six Mandarin-supporting Speech LLMs, including GPT-4o-Audio (OpenAI), Gemini-1.5-pro and Gemini-2-flash (Google), Qwen2-Audio-7B-Instruct (Alibaba Cloud), GLM-4-Voice (Zhipu AI), and MiniCPM-o 2.6 (OpenBMB)[30, 31, 32, 33, 34]. The selected Speech LLMs encompass diverse multimodal architectures and application scenarios, spanning proprietary cloud-based solutions (GPT-4o, Gemini series) to open-source models capable of local deployment (Qwen2, GLM-4, MiniCPM). All models support Mandarin audio and text processing.

Prompting Method: We used a zero-shot multiple-choice question answering (MCQA) prompting paradigm. Models received multimodal inputs: text prompts (instructions, context, and options) and audio stimuli. Responses were generated independently for each trial. All models used default parameters, with five trials per condition for consistency. GLM-4-Voice, handling one modality at a time (per official documents), received text then audio sequentially.

Automatic Evaluation: To objectively assess content understanding while neutralizing format biases, we implemented an LLM-based semantic mapping for automatic accuracy evaluation. Specifically, we used GPT-4o to map free-text responses to corresponding choice labels (from 'A' to 'D'). If no suitable choice was identified, the label 'X' was assigned. accuracy was calculated as $\frac{\text{Correct Choices}}{\text{Total Trials}} \times 100\%$, with mean values and standard deviations computed across five independent runs.

3.2.2. Results of Speech LLMs

Overview of Speech LLMs Performance: Our evaluation of six Mandarin-supporting Speech LLMs on the MSPB revealed considerable performance differences across models and tasks (Figure 2). Proprietary models, specifically GPT-4o and the Gemini series generally outperformed the open-source alternatives. GPT-4o achieved the highest average accuracy(59.70%), followed by Gemini-2-flash (59.28%) and Gemini-1.5-pro (56.05%). Among the open-source models, MiniCPM-o 2.6 demonstrated competitive performance with an average accuracy of 50.44%, surpassing Qwen2-Audio-7B-Instruct (46.41%) and GLM-4-Voice (47.66%). The superior performance of proprietary models is likely due to their larger size, more extensive training data, and improved multimodal integration. However, MiniCPM-o 2.6 shows that smaller models can still be effective for prosodic understanding.

Performance Comparison Across Task Categories: Performance varied significantly across the eight prosodic tasks shown in Figure 1. Most Speech LLMs excelled in tasks with both prosodic and contextual cues(e.g., irony, focus operator).

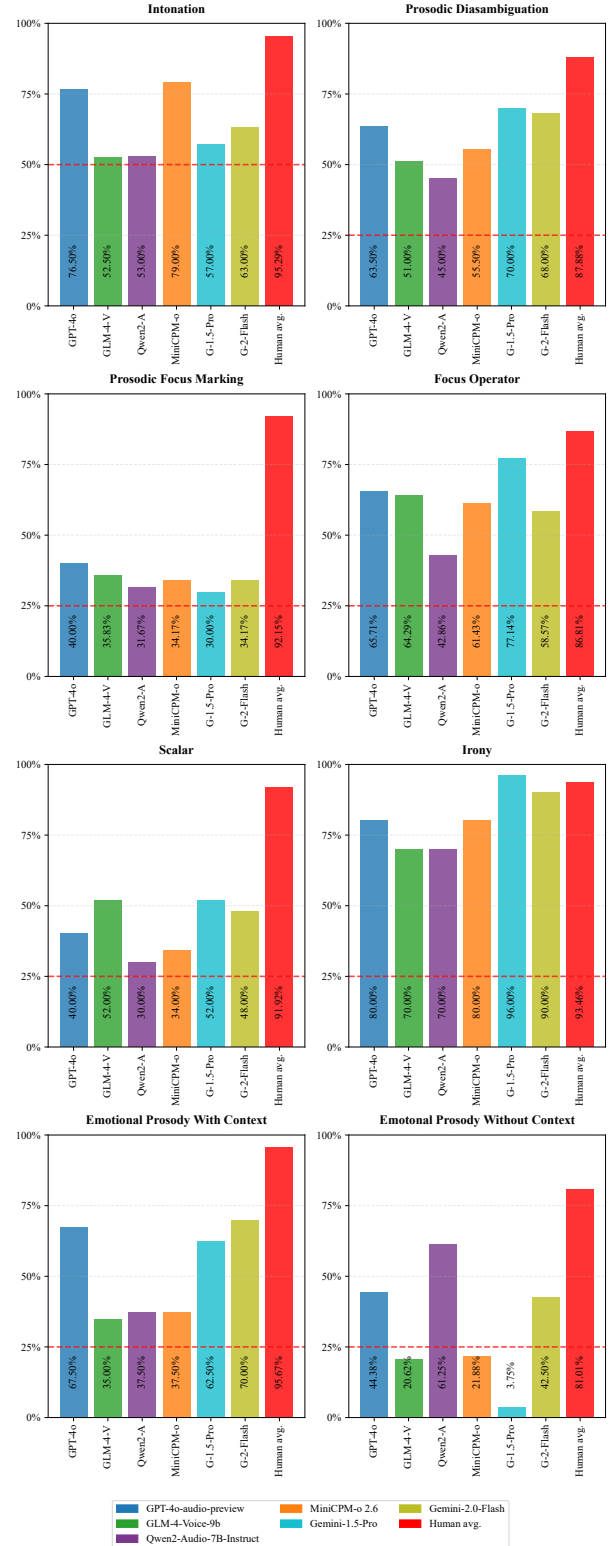


Figure 1: Performance comparison across various tasks between Speech LLMs and the average performance of human participants. The red dashed line represents the chance level.

Most models also showed improved performance on emotional



Figure 2: Radar chart comparing the performance of different Speech LLMs. **Tasks:** Intonation, ProsDisamb (Prosodic Disambiguation), ProsFocMark (Prosodic Focus Marking), FocusOp (Focus Operator), Scalar, Irony, Emo w/ Ctx (Emotional Prosody with Context), Emo w/o Ctx (Emotional Prosody without Context). The red solid line denotes human participants average performance, while the red dashed line represents chance level. The shaded regions illustrate the standard deviation.

prosody with context, highlighting their reliance on contextual information. Conversely, all models consistently struggled with tasks requiring strong reliance on speech prosody (e.g., prosodic focus, scalar meaning). Overall, the better performance on context-rich tasks compared to prosody-dependent tasks underscores a critical limitation of current Speech LLMs: they rely more on contextual cues than on subtle speech prosody variations. Addressing this gap presents a key area for future development. Notably, Gemini-1.5-Pro declined to respond to the emotional prosody without context tasks, which may be attributed to safety or compliance restrictions.

3.3. Performance between Speech LLMs and Human Participants

A mixed-design ANOVA was conducted to examine the effects of testee (model vs. human), task category, and their interaction on accuracy. The analysis revealed significant main effects for both testee ($F(6, 49) = 189.77, p < 0.001$) and task category ($F(7, 343) = 116.86, p < 0.001$), as well as a significant interaction between method and task category ($F(42, 343) = 23.80, p < 0.001$). This result confirmed that both the type of testee (model or human) and the specific prosodic task contribute significantly to performance variation. The significant interaction effect underscores that these performance disparities are not uniform but are highly task-dependent.

Further pairwise comparisons were conducted using parametric or non-parametric tests (determined by data characteristics) with Bonferroni correction. These analyses compared model and human performance across task categories. Results showed significant differences between models and humans in most tasks, though several models achieved human-comparable performance in specific domains. In the Irony task, the performance of Gemini-1.5-Pro, Gemini-2.0-Flash, GPT-4o-audio-preview, and MiniCPM-o showed no significant difference from humans. Similarly, in the Focus Operator task, Gemini-1.5-Pro, GLM-4-Voice, GPT-4o, and MiniCPM-o achieved statistically indistinguishable performance from humans. Despite achieving human-level performance in specific tasks, systematic analysis revealed a persistent performance gap in key prosodic dimensions between Speech LLMs and humans. Human testee significantly outperformed models, especially when prosodic cues were the sole cues available, underscoring the considerable challenge these tasks pose for Speech LLMs. This indicates that despite progress in multimodal integration and task-specific cue utilization, current Speech LLMs still require substantial improvement to fully capture the nuanced subtleties of human speech prosody.

4. Conclusion

This study introduced the Mandarin Speech Prosody Benchmark (MSPB), a novel, linguistically-grounded benchmark for evaluating the prosodic comprehension of Speech LLMs in Mandarin. Our evaluation of six state-of-the-art models, including both proprietary (GPT-4o, Gemini series) and open-source (Qwen2, GLM-4, MiniCPM) options, revealed a significant performance gap between Speech LLMs and human participants across most prosodic tasks. While some models achieved human-comparable performance on tasks with rich contextual cues (e.g., irony, focus operator), they generally struggled with tasks relying heavily on subtle prosodic variations (e.g., prosodic focus marking, scalar meaning). This highlights a critical need for future research to better integrate fine-grained prosodic information into Speech LLMs.

Key contributions of this work include the MSPB itself, a comprehensive evaluation of leading Speech LLMs, and an in-depth analysis revealing model-specific strengths and weaknesses. It is noteworthy that this study focused on a zero-shot, single-round approach in Mandarin, and future work should explore multi-turn conversational settings and expand to a multilingual benchmark. Despite these limitations, our findings provide crucial insights into the current state of Speech LLMs and underscore the importance of prosodic understanding for achieving truly natural and effective human-machine interaction.

5. Acknowledgements

This study was funded by the National Social Science Foundation of China (Project No. 23BY192). We sincerely thank all participants for their support.

6. References

- [1] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu *et al.*, "Audiopt: Understanding and generating speech, music, sound, and talking head," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 802–23 804.
- [2] R. A. F. Dewatri, A. Z. Al Aqthar, H. Pradana, B. Anugerah,

- and W. H. Nurchahyo, "Potential tools to support learning: Openai and elevenlabs integration," *Southeast Asian Journal on Open and Distance Learning*, vol. 1, no. 02, 2023.
- [3] Z. Wang, N. Chen, L. K. Qiu, L. Yue, G. Guo, Y. Ou, S. Jiang, Y. Yang, and L. Qiu, "The potential and value of ai chatbot in personalized cognitive training," *arXiv preprint arXiv:2410.19733*, 2024.
 - [4] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, Y. Guo, and I. King, "Recent advances in speech language models: A survey," *arXiv preprint arXiv:2410.03751*, 2024.
 - [5] J. Peng, Y. Wang, Y. Xi, X. Li, X. Zhang, and K. Yu, "A survey on speech large language models," *arXiv preprint arXiv:2410.18908*, 2024.
 - [6] D. Crystal, "The analysis of intonation in young children," *Communication and cognitive abilities: Early behavioral assessment*, pp. 257–271, 1978.
 - [7] A. Cutler and D. R. Ladd, "Prosody in understanding speech: A review," *Language and Cognitive Processes*, vol. 16, no. 5, pp. 631–671, 2001.
 - [8] C. Gussenhoven and A. Chen, *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford, UK: Oxford University Press, 2020.
 - [9] D. Norris, A. Cutler, J. M. McQueen, and S. Butterfield, "Phonological and conceptual activation in speech comprehension," *Cognitive Psychology*, vol. 53, no. 2, pp. 146–193, 2006.
 - [10] L. D. Sanders and H. J. Neville, "Lexical, syntactic, and stress-pattern cues for speech segmentation," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 6, pp. 1301–1321, 2000.
 - [11] M. H. K. Ip and A. Cutler, "Juncture prosody across languages: Similar production but dissimilar perception," *Laboratory Phonology*, vol. 13, no. 1, 2022.
 - [12] A. J. Schafer, S. R. Speer, P. Warren, and S. D. White, "International disambiguation in sentence production and comprehension," *Journal of psycholinguistic research*, vol. 29, pp. 169–182, 2000.
 - [13] M. Yan, P. Warren, and S. Calhoun, "Focus interpretation in l1 and l2: The role of prosodic prominence and clefting," *Applied Psycholinguistics*, vol. 43, no. 6, pp. 1275–1303, 2022.
 - [14] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 1044–1098, 2010.
 - [15] B. Braun, N. Dehé, J. Neitsch, D. Wochner, and K. Zahner, "The prosody of rhetorical and information-seeking questions in german," *Language and Speech*, vol. 62, no. 4, pp. 779–807, 2019.
 - [16] S. Repp, "The prosody of wh-exclamatives and wh-questions in german: Speech act differences, information structure, and sex of speaker," *Language and speech*, vol. 63, no. 2, pp. 306–361, 2020.
 - [17] E. Pratt, "Prosody in sentence processing," *The handbook of psycholinguistics*, pp. 365–391, 2017.
 - [18] J. S. Chan, L. Wade-Woolley, L. Heggie, and J. R. Kirby, "Understanding prosody and morphology in school-age children's reading," *Reading and Writing*, vol. 33, no. 5, pp. 1295–1324, 2020.
 - [19] 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," *arXiv preprint arXiv:2406.16020*, 2024.
 - [20] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou *et al.*, "Air-bench: Benchmarking large audio-language models via generative comprehension," *arXiv preprint arXiv:2402.07729*, 2024.
 - [21] J. Ao, Y. Wang, X. Tian, D. Chen, J. Zhang, L. Lu, Y. Wang, H. Li, and Z. Wu, "Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words," *arXiv preprint arXiv:2406.13340*, 2024.
 - [22] G. Maimon, A. Roth, and Y. Adi, "A suite for acoustic language model evaluation," *arXiv preprint arXiv:2409.07437*, 2024.
 - [23] 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," *arXiv preprint arXiv:2402.12786*, 2024.
 - [24] Y. R. Chao, *Mandarin primer: An intensive course in spoken Chinese*. Harvard University Press, 1948.
 - [25] A. Li and M. Lin, "Studies on chinese prosody-professor wu zongji's thoughts on prosody and their far-reaching implications," in *Proceedings of the 8th Phonetics Conference of China*, 2008, pp. 1–17.
 - [26] J. Yu, K. Zhang, and S. Zhu, "Prosodic decoding profiles of chinese mandarin-speaking children under visual and audio modality," in *Proc. SpeechProsody 2024*, 2024, pp. 752–756.
 - [27] T. Wang and Y.-c. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in mandarin chinese?" *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. EL117–EL123, 2015.
 - [28] J. Shen, "Pitch range and intonation in beijing mandarin," *Beijing Phonetic Experiment Records / Peking University Press*, 1985.
 - [29] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glott International*, vol. 5, no. 9/10, pp. 341–347, 2001.
 - [30] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.
 - [31] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.
 - [32] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.
 - [33] A. Zeng, Z. Du, M. Liu, K. Wang, S. Jiang, L. Zhao, Y. Dong, and J. Tang, "Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot," *arXiv preprint arXiv:2412.02612*, 2024.
 - [34] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.