DEBATE：一个通过语音解析中文文本歧义的数据集

Haotian Guo*
haotianguo@hnu.edu.cn
Hunan University
Changsha, China

Jing Han*†
jh2298@cam.ac.uk
University of Cambridge
Cambridge, uk

Yongfeng Tu*
tuyongfeng@mlslabs.com.cn
Malanshan Audio & Video Laboratory
Changsha, China

Shihao Gao
shihaogao@hnu.edu.cn
Hunan University
Changsha, China

Shengfan Shen
shenshengfan@hnu.edu.cn
Hunan University
Changsha, China

Wulong Xiang
xiangwulong@hnu.edu.cn
Hunan University
Changsha, China

Weihao Gan†
ganweihao@mlslabs.com.cn
Malanshan Audio & Video Laboratory
Changsha, China

Zixing Zhang†
zixingzhang@hnu.edu.cn
Hunan University
Changsha, China

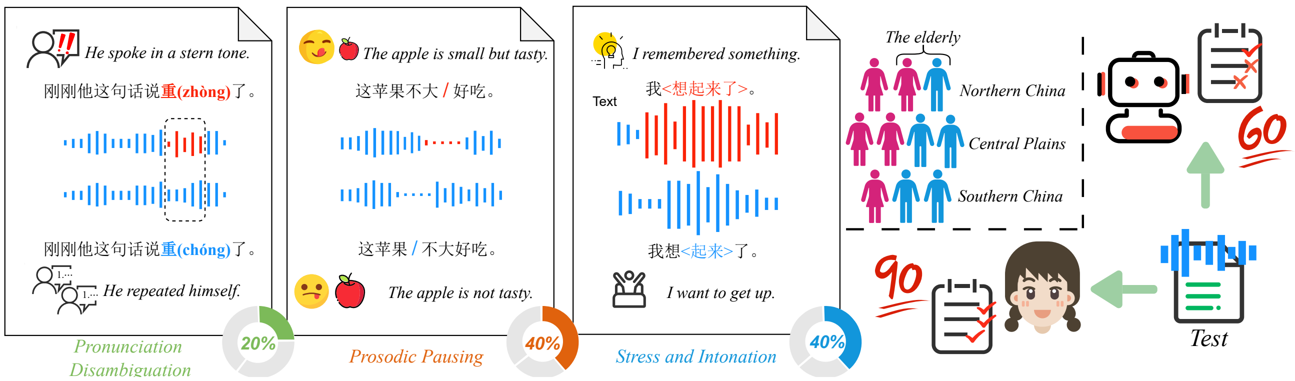

Figure 1: DEBATE 是第一个为基于语音的中文语义消歧设计的多模态数据集，重点研究声学线索如何帮助解决文本歧义。该数据集包含超过 10K 条音频录音，由 1,001 条具有歧义的文本提示构成，每条文本提示由 10 位母语者朗读，并带有已消除歧义的注释。左图展示了三个消歧场景，右图则展示了数据集的关键统计信息，包括说话人数、性别比例和场景分布。通过对 DEBATE 的评估，发现大型语音语言模型与人类之间存在显著的性能差距。

Abstract

尽管在文本和视觉消歧问题上已有大量研究，但通过语音进行消歧（DTS）仍然未被充分探索。这主要是由于缺乏高质量的数据集，这些数据集将语音句子与丰富的模糊文本配对。为了解决这一差距，我们推出了 DEBATE，这是一个独特的公共中文语音文本数据集，旨在研究语音线索和模式——发音、停顿、重音和语调——如何帮助解决文本模糊性并揭示说话者的真实意图。DEBATE 包含 1001 个精心挑选的模糊话语，每个话语由 10 名母语者录制，涵盖多样的语言模糊性及其通过语音进行的消歧。我们详细介绍了数据收集流程并提供了严格的质量分析。此外，我们对三个最先进的大型语音和音频语言模型进行了基准测试，展示了机器和人类在理解语音意图方面的明显和巨大的性能差距。DEBATE 代表了首次此类努力，并为跨语言和文化建立类似的 DTS 数据集提供了基础。数据集和相关代码可在以下网站获取：https://github.com/SmileHnu/DEBATE。

CCS Concepts

• **Information systems** → **Multimedia databases**; • **Computing methodologies** → **Lexical semantics**; **Speech recognition**.

*These authors contributed equally to this research.
†Jing Han, Weihao Gan, and Zixing Zhang are corresponding authors.

Keywords

Speech-based Disambiguation, Chinese Speech Dataset, Large Speech-Language Models

1 介绍

"Words mean more than what is set down on paper.
It takes the human voice to infuse them with deeper meaning."
—Maya Angelou

"听其言，观其音，察其意。" —汉 · 班固

Despite advances in Large Language Models (LLMs), certain forms of ambiguity remain challenging to resolve [18, 21]. Ambiguity arises naturally in human language when words or phrases carry multiple possible interpretations. For example, Mandarin Chinese, an isolating (analytic) language, relies primarily on semantic context to convey grammatical relationships [20, 28]. This is different from synthetic languages, which leverage morphological inflections (e. g., changes in word form to indicate tense, number, or case) and use more complex grammatical conjugation [5, 30]. Therefore, while the structural flexibility of Mandarin enhances expressive richness, it also makes the language more prone to syntactic and semantic ambiguities, especially in written, text-only form [33, 38]. Also, in Mandarin, words can be formed from either single or multiple characters. Unlike English, Chinese sentences lack explicit word boundaries, such as spaces, and rely primarily on limited punctuation marks to complete full sentences. For instance, a string like "ABC" (e. g.,"地面积") can be interpreted as either "AB/C" (e. g.,"地面/积") or "A/BC"(e. g.,"地/面积"), both yielding valid but distinct meanings. This phenomenon is comparable to the classic English example "Godisnowhere", which can be understood as either "God is now here" or "God is nowhere".

While such ambiguity can sometimes be a source of humour or wordplay, most real-world communication demands clarity and precision to avoid misunderstanding [7, 22]. Disambiguation has long been a central research topic in Chinese NLP [4, 32, 41]. A variety of text-based corpora have been developed, and computational models have been proposed to resolve these ambiguities through techniques such as sentence rewriting or contextual reasoning [9, 17, 33, 38, 41]. However, most existing efforts focus solely on textual input, ignoring other communicative modalities that humans naturally rely on for clarity.

Meanwhile, recent advancements in speech and audio understanding and their integration with LLMs have enabled audio-based understanding to play a more significant role in multimodal AI systems [11, 13, 15, 16, 39]. In natural spoken communication, some of the ambiguities present in written Chinese can be effortlessly resolved. We categorise such speech-based disambiguation cues into three main types: (1) Pronunciation Disambiguation : Certain characters in Chinese have multiple pronunciations, which can convey different meanings. When spoken aloud, the sound naturally disambiguates the intended word. (2) Prosodic Pausing : Although written Chinese does not use spaces to separate words, native speakers tend to pause between semantically distinct word groups during speech, providing cues for proper segmentation and interpretation. (3) Stress and Intonation : Stress patterns are not represented in written text, but speakers may use stress and intonation to highlight intent and clarify meaning during oral communication.

Despite the growing availability of text-only ambiguity corpora and various speech-text paired datasets for ASR or question answering tasks, to the best of our knowledge, there is no existing dataset specifically designed to study how speech cues naturally disambiguate textual ambiguity in Mandarin. To address this gap, we introduce DEBATE : a novel speech-text corpus aimed at D isentangling T e xtual Am b iguity in M a ndarin t hrough Sp e ech. The DEBATE dataset is designed to evaluate and train models on their ability of disambiguation through speech (DTS). Our main contributions are as follows:

- 我们推出了 DEBATE，这是一个公开可用的普通话语音文本语料库，专门设计用于通过语音促进消歧研究。该语料库包括 1k+ 个文本条目，与 10k+ 个音频样本（共 9.66 小时）配对。
- 我们建立了一个结构化的数据收集框架，概述了构建数据库所涉及的每一个步骤，包括文本语料库的开发和音频录制设置，以确保多样性和质量。
- 我们对三个大型语音-语言模型进行了基准测试。我们的结果显示，这些模型在有效利用声学信息来理解真实意图方面仍然存在困难，其在 DTS 上的表现远不如人类参与者。

2 相关工作

In recent years, several corpora have been developed for solving disambiguation problems in Chinese text. Some focus on the character level, identifying commonly used morphemes with multiple interpretations and building corresponding vocabularies [33]. Others operate at the sentence level, crawling from the web data or integrating specialised corpora to create Word Sense Disambiguation (WSD) datasets [36, 38]. Similarly, many efforts have been made in other languages and cultures. Well-known examples of such efforts can be the multilingual WSD tasks from the SemEval evaluation series [23, 24]. These works provide valuable foundations for understanding the disambiguation issue and inspire the development of our own text corpus.

Moreover, various speech and multimodal datasets have been proposed to advance spoken language understanding, particularly for uncovering nuanced speaker intent beyond the literal meaning of words [3, 25, 34, 40]. For example, some datasets focus on identifying the intended recipient of an utterance in multi-party conversations [1, 29, 31], while others focus on detecting humour and sarcasm [3, 10, 37]. These studies highlighted that prosodic features such as stress and intonation play a vital role in disambiguating speech and inferring the true intent in real-life communication [8]. Additionally, dialogue datasets have been constructed to build models in enabling more natural and effective communication [19, 27].

While these resources have largely advanced research in speech-based understanding, none have explicitly tackled the challenge of DTS, especially for Mandarin. The absence of a high-quality dataset has thus hindered progress in this area. To bridge this gap, our work introduces a novel dataset specifically designed to investigate how spoken cues, such as pronunciation, pauses, and stress, can aid in disambiguating semantically ambiguous text in Mandarin Chinese.

3 DEBATE 数据集

3.1 数据集概述

The issue of ambiguity in Chinese has long existed, drawing widespread attention in linguistic research and posing persistent challenges for natural language processing systems. Depending on the modality through which information is conveyed, such ambiguities can be broadly divided into two categories: those that arise in the spoken modality but can be resolved using textual cues, and those that exist in text but require speech-based signals—such as pronunciation, intonation, and pauses—for disambiguation. The DEBATE dataset we constructed focuses primarily on the latter, where surface-level textual content is ambiguous, yet the intended meaning can be clarified through prosodic and phonetic information through speech.

Specifically, DEBATE covers three representative types of speech-assisted disambiguation scenarios. The three scenarios are categorised based on the primary speech cues required to resolve ambiguity, as outlined below.

First, polyphonic character ambiguity is a common phenomenon in Chinese, where a single character may have multiple pronunciations and corresponding meanings. This often results in textual expressions that are inherently ambiguous or even appear as puns. For instance, in the left example of Figure 1, the character "重" could be interpreted as indicating either a serious tone or repeated speech, depending on its pronunciation. In written form without sufficient context, determining the intended meaning is difficult. However, the speech signal clearly reveals the correct pronunciation, allowing the listener (and ideally a multimodal model) to resolve the ambiguity accurately.

Second, structural ambiguity arises due to the absence of explicit word boundaries and punctuation in Chinese text. Different syntactic segmentations may lead to entirely different meanings. As shown in the middle example of Figure 1, the same sentence can be parsed to suggest that the apple is tasty or not, depending on how it is segmented. In spoken language, the speaker' s natural prosodic pauses provide crucial structural cues that indicate sentence boundaries. By leveraging features such as the position and duration of pauses, one can infer sentence structure and reduce ambiguity at the syntactic level.

Third, ambiguity can stem from differences in semantic focus, which are often marked by stress or intonation ambiguity . In Chinese, shifting the position of stress within a sentence can significantly alter its meaning. For example, in the third pair of sentences shown in Figure 1, when the stress falls on "想起来了", it indicates that the speaker has recalled something. In contrast, when the stress is placed on "起来" , it suggests an intention to physically get up. Such semantic ambiguities caused by stress placement are often difficult to detect in text, but in spoken communication, speakers typically clarify the intended meaning through variations in pitch or intensity.

In summary, DEBATE targets ambiguity types that are textually obscure yet resolvable through speech.

3.2 数据生成流水线

The overall pipeline is illustrated in Figure 2, comprising three main stages. We depict each stage in the following.

3.2.1 原始文本收集和注释 .

We first constructed a Chinese text dataset tailored for research on speech-based disambiguation. The data was collected from three primary sources: (1) open-source corpora, (2) social media platforms, and (3) standardised examination question banks.

Specifically, we extracted representative ambiguous samples directly from publicly available Chinese ambiguity corpora hosted on GitHub. For social media data, we retrieved user-generated posts exhibiting semantic ambiguity by querying with keywords such as "ambiguity", "phrase segmentation", and "stress". Additionally, we selected structurally complex and ambiguous examples from sentence analysis tasks in the verbal reasoning sections of Chinese civil service examinations.

After collecting the text, we performed systematic manual annotation to classify the data into three distinct types of ambiguity: polyphonic character ambiguity , structural ambiguity , and focus ambiguity . Samples exhibiting other types of semantic ambiguity outside these categories were excluded. To further enrich the dataset in terms of coverage and syntactic diversity, we employed large language models (LLMs) such as GPT and DeepSeek to generate additional sentences following the structure and style of the manually collected examples. All machine-generated samples underwent human verification to ensure grammaticality, logical consistency, and alignment with the target ambiguity type.

To maintain ethical standards, all sentences containing inappropriate or sensitive content were removed to ensure the dataset's compliance and public usability. As a result of this process, we curated a total of 200 samples of polyphonic character ambiguity, 401 ones with structural ambiguity, and 400 ones of focus ambiguity.

After collecting and preliminarily filtering the ambiguous text, we further introduced three prosodic-level strategies to systematically disambiguate the selected sentences. These strategies specifically target three common types of ambiguity: polyphonic character ambiguity, structural ambiguity, and focus-related ambiguity.

For polyphonic character ambiguity , we explicitly annotated the pronunciation of each character to eliminate semantic confusion caused by identical written forms with different pronunciations.

For structural ambiguity , we incorporated prosodic annotation by using the "/" symbol to mark prosodic boundaries within sentences. These boundaries indicate appropriate short pauses in speech, which help clarify syntactic structures and enhance sentence intelligibility.

For focus ambiguity , we marked key words requiring emphasis with stress symbols or pitch-rise indicators. This approach reinforces the pragmatic focus of the sentence and assists listeners in accurately identifying the speaker' s intended meaning and communicative emphasis.

To further enhance the usability of this dataset in semantic understanding and downstream evaluation tasks, we generated semantic annotations for each sentence. These annotations were produced using LLMs, generating multiple candidate explanations per sentence. We then conducted human review to evaluate the fluency, semantic accuracy, and logical consistency of each candidate. Only the most optimal explanation per disambiguated sentence was retained.

3.2.2 语音数据录制 .

To ensure both diversity and quality control of the speech data, we recruited ten volunteers to participate in the audio recording task, including eight young adults and two elderly individuals, with a balanced gender ratio (5:5) to ensure demographic representativeness in terms of both age and gender. All participants received detailed instructions prior to recording, including specific guidance on prosodic features such as stress placement, intonation patterns, and appropriate pause locations based on different task types.

The recordings were conducted using participants' own devices, primarily consisting of smartphone microphones and built-in laptop microphones. This flexible device setup was intentionally adopted to closely mimic real-world deployment scenarios of speech models, thereby enhancing the ecological validity and generalisability of the dataset. To control recording quality, each participant was required to submit ten test recordings prior to the formal session. These samples were manually reviewed to evaluate recording equipment performance and environmental conditions, such as

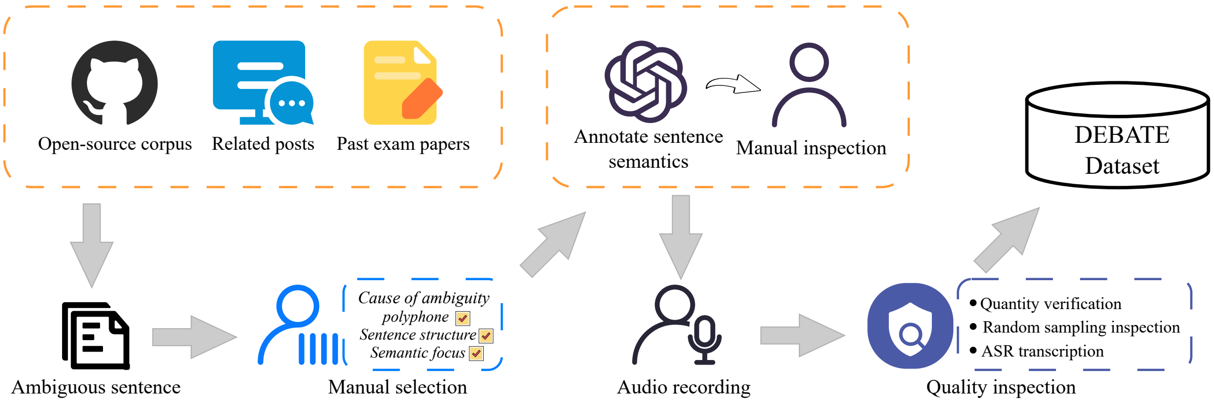

Figure 2: 创建 DEBATE 的流程——这是第一个通过语音设计用于中文消歧的语音文本数据集。模糊文本最初从多样化的来源收集，并经过人工检查以识别合适的候选句子。然后通过大型语言模型扩展这些句子，并进行第二轮人工验证。随后，选定的句子由 10 名母语人士朗读，并经过严格的质量控制。最终的 DEBATE 数据集包含了超过 1 万条经过精心策划的语音录音，并附有丰富的注释。

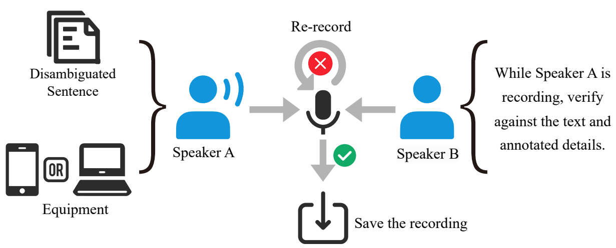

Figure 3: 音频录制过程的说明。每个会话涉及两个参与者：一个专门的说话者和一个被动的听者，同时录制以最大化录音质量。

background noise and reverberation levels. Participants who met the quality standards were allowed to proceed to the full recording task.

During the formal recording sessions, we implemented a two-person collaborative mechanism: one individual performed the recording while the other monitored in real time, listening for issues such as mispronunciations, omissions, or semantic inconsistencies (see Fig. 3). If any issues were noticed, immediate feedback was provided and the utterance was re-recorded. This approach helped improve both the accuracy and consistency of the recordings, effectively preventing the accumulation of low-quality data.

As a result, we constructed a high-quality, multi-speaker speech dataset that maintains both naturalness and precision.

3.2.3 质量控制 .

Once the recording phase was completed, we conducted a systematic post-review of all recorded materials to ensure corpus completeness and audio quality, minimising issues such as missing files, duplicates, or speech defects. First, we verified the one-to-one correspondence between text entries and audio files, checking for potential omissions or redundant recordings to guarantee that each text sample was matched with a unique audio file per speaker.

Table 1: ASR 模型在 DEBATE 上的字符错误率（%）。

Models	T_{Proun}	T_{Pause}	T_{Stres}	Average
Whisper-large-v3 turbo	9.66	7.14	5.28	7.37
SenseVoice-Small	4.75	2.82	1.94	3.48

Building on this, we performed a quality assessment of each speaker’ s recordings across different tasks. Specifically, ten audio samples were randomly selected from each task and manually evaluated for pronunciation clarity, sentence completeness, naturalness of speech rate, and adherence to prosodic norms such as stress placement and pausing. This strategy enabled efficient identification of potential quality concerns and allowed for timely corrective actions. The sampling results indicated that all recordings met the expected standards, demonstrating overall high quality.

To further quantify audio quality, we exploited two advanced automatic speech recognition (ASR) systems, i. e., Whisper-large-v3-turbo [26] and SenseVoice-small [2] to transcribe all recordings and compute Character Error Rate (CER) as a metric of alignment between audio and reference transcripts. The evaluation results are presented in Table 1. Notably, sensevoice-small achieves CER of 2.96 % and 3.80 % on the two high-quality Mandarin speech recognition corpora AISHELL-1 [6] test and AISHELL-2 [14] test_ios, respectively, while attaining CERs of 2.82 % and 1.94 % on the T_{Pause} and T_{Stres} datasets. This performance demonstrates the high consistency between audio and text in our DEBATE dataset.

3.3 数据集统计

In total, the DEBATE dataset comprises approximately 9.66 hours of high-quality speech data spanning a variety of disambiguation task scenarios. Detailed descriptions of dataset structure, task distribution, and quality evaluation metrics can be found in Table 2. In particular, individual recordings range from 1.15 seconds to 11.80 seconds in duration, and the corpus exhibits strong diversity and

Table 2: DEBATE 数据集的详细统计。

Tasks	# Samples	Hours	Mean Duration (s)	Duration Range (s)
T_{Proun}	2 000	1.64	2.94	1.15-5.80
T_{Pause}	4 010	4.28	3.84	1.60-11.80
T_{Stres}	4 000	3.74	3.37	1.43-8.51
Total	10 010	9.66	3.47	1.15-11.80

representativeness in content. These characteristics make DEBATE a robust and valuable asset for future research and applications.

4 基准测试和结果

As a showcase of speech-based disambiguation using the DEBATE dataset, we benchmarked several representative Large Speech Language Models (LSLMs). The primary objective is to assess how well these models can interpret semantically ambiguous sentences by leveraging speech cues. Specifically, we examine their ability to utilise salient acoustic features—such as pronunciation differences, prosodic boundaries, speech rate and rhythm, and stress patterns—to resolve ambiguities stemming from polyphonic characters, syntactic structures, and unclear semantic focus within text.

4.1 实验设置

For this aim, we selected three widely-used LSLMs for evaluation, i. e., Qwen2-audio [11], Qwen2.5-omni [35], and Gemini 2.0 Flash [12]. These models have demonstrated excellent performance in speech understanding tasks. All evaluations were conducted under a zero-shot inference setting to examine whether the models can perform semantic disambiguation.

For task construction, speech samples were used as input queries, paired with carefully designed text prompts to specify the semantic discrimination objective. Each test item was formatted as a single-choice question: among multiple plausible interpretations of an ambiguous sentence, two were manually selected as answer options, only one of which matched the intended meaning conveyed in the audio. The model was required to select the correct interpretation based on the provided audio. The detailed prompt template used for constructing these tasks is provided in Table 3.

4.2 结果与分析

We report performance in terms of accuracy and macro-F1. Table 4 summarises the performance of three selected LSLMs across different types of semantic disambiguation tasks. Overall, Gemini-2.0-Flash and Qwen2.5-Omni demonstrated comparatively strong performance, exhibiting competitive capabilities across tasks, whereas Qwen2-Audio showed relatively weaker results. This performance gap can be partially attributed to the more powerful underlying LLM architectures and the larger-scale training corpora used by the former two models.

In terms of specific task performance, all three models demonstrated generally decent results in polyphone disambiguation (T_{Proun}) and sentence structure understanding tasks (T_{Pause}). Notably, in the sentence structure task, the model outputs were more stable

Table 3: 用于模型推理任务的提示构建模板。括号中的文本仅用于说明目的，不包括在实际提示中。在每次推理运行中，仅选择并使用三个可选文本块（用背景颜色突出显示）中的一个。请注意，原始提示是用中文书写的，这里翻译成英文只是为了清楚和展示。

Prompt Template
As a model equipped with professional-level audio semantic understanding capabilities, you are expected to accurately identify the precise meaning conveyed in the given audio segment.
(T_{Proun} only) A core challenge lies in handling polyphonic characters. You must correctly identify the actual pronunciation of each polyphonic word within the audio and, by integrating contextual information, accurately infer its intended meaning—an essential step for overall semantic interpretation.
(T_{Pause} only) A core aspect of this task lies in analysing pause-related features—such as the duration of silence and prosodic interruptions—which serve as primary cues for determining the hierarchical structure of the sentence. You must incorporate this structural information to accurately interpret the overall meaning and logical relationships within the sentence.
(T_{Stress} only) A key challenge lies in capturing prosodic variations in the audio—such as pitch rises or falls, changes in speech rate, and stress patterns—which are essential cues for inferring the speaker's intended emphasis.
I will provide two possible meanings for you to choose from: A. [Meaning 1] B. [Meaning 2] Please choose the option you consider most appropriate. No other output text/explanation. Only provide the option.

and consistent, indicating that current LSLMs possess the foundational capabilities to detect intra-sentence pauses. This suggests that these models can, to a certain extent, perceive hierarchical sentence structures and infer content that aligns more closely with the intended semantic meaning.

However, in the stress understanding task, all models performed poorly, with Qwen2-Audio in particular producing results that were nearly random. This highlights a clear limitation in the models' ability to capture finer prosodic features in speech, such as stress patterns and pitch variations. Compared to pauses, which are relatively explicit and have clear temporal boundaries, stress involves more complex acoustic variations within the speech signal. Accurate recognition of speaker emphasis at the semantic level thus requires a higher degree of prosodic modelling capability. Moreover, current pre-trained LSLMs are predominantly trained on large-scale, automatically transcribed corpora, in which prosodic features like stress are rarely annotated. This further constrains the models' ability to learn and generalise such nuanced characteristics.

Comparing Human Performance and LSLMs. We randomly selected 50 samples per subtask of the DEBATE dataset to construct a small-scale evaluation set, ensuring that each speaker contributed five unique samples. This set was used for manual evaluation, and we also assessed the performance of LSLMs on the same set to highlight the gap between model and human capabilities in speech

Table 4: 在 DEBATE 数据集上评估大规模语言模型（LSLMs）的准确性（%）和宏平均 F1 分数（%）。结果报告了三个模型：Qwen2-Audio、Qwen2.5-Omni 和 Gemini 2.0 Flash。对于每个任务，我们展示了 10 个说话者的平均值和方差。每列中表现最好的结果已突出显示。

Models	T_{Proun}		T_{Pause}		T_{Stres}	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Qwen2-Audio	58.60 ± 3.43	55.40 ± 4.03	55.64 ± 1.85	51.10 ± 1.66	51.57 ± 1.01	47.60 ± 1.43
Qwen2.5-Omni	65.65 ± 2.30	65.20 ± 2.62	68.08 ± 1.47	67.90 ± 1.60	55.02 ± 2.16	55.20 ± 2.15
Gemini 2.0 Flash	61.15 ± 3.95	60.80 ± 4.16	68.00 ± 1.69	67.90 ± 1.79	58.83 ± 4.06	58.00 ± 4.47

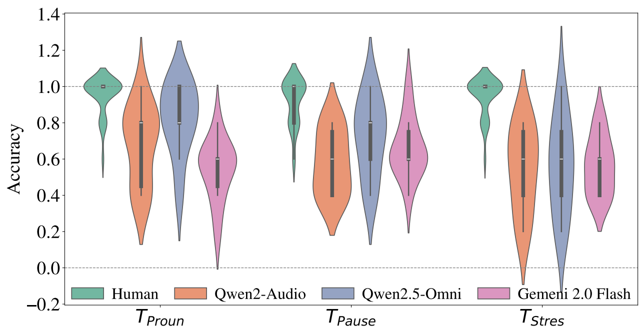

Figure 4: 使用小提琴图比较人在小规模测试集上的表现与模型的表现。"人"分数是由三名志愿者评估的综合结果。

understanding. The evaluation procedure was consistent with the model inference process and was independently conducted by three volunteers who were not involved in this study. The violin plots in Fig. 4 illustrate that LSLMs not only underperform compared to human listeners across all three tasks, but also exhibit greater performance dispersion. This suggests that all tested LSLMs struggle both in accuracy and in consistency across speakers and sentences.

In summary, while current LSLMs show potentials in perceiving certain speech cues, there remains substantial room for improvement. This highlights the critical need for future research to enhance models' ability to interpret the rich prosodic and acoustic cues in speech—an essential step toward achieving deeper semantic understanding and more accurate user intent interpretation.

5 结论

We introduce DEBATE, a publicly available Mandarin speech-text corpus specifically designed for semantic disambiguation tasks. Using DEBATE, we conducted a systematic assessment of current large-scale speech-language models (LSLMs), including representative models from the Qwen and Gemini families. Our results suggest that current LSLMs exhibit a modest ability to perform speech-based disambiguation, with best accuracy about 58-68 % across all tasks and models—still well below human performance. This limitation hinders their ability to accurately resolve complex semantic ambiguities, highlighting the need for further improvements in acoustic modelling within LSLMs. In addition to serving as a benchmark for semantic understanding, DEBATE also holds promise as a fine-tuning resource for text-to-speech (TTS) systems—particularly for improving pronunciation disambiguation of homographs and the natural rendering of prosodic stress. These remain key challenges in most current TTS systems, and the DEBATE corpus offers targeted data to address them. Moreover, the dataset contains implicit regional phonetic variation, making it a valuable resource for modelling linguistic variants. This opens up possibilities for research in sociolinguistics and language adaptation, especially in studying the regional adaptability and standardisation of Mandarin pronunciation across different dialectal backgrounds.

Limitations The current version of DEBATE covers only Mandarin Chinese. Besides, we evaluated only three existing models, which, while representative, may not reflect the full spectrum of approaches in audio-language understanding. Future work should explore broader model families and training paradigms.

Ethics Statement All data collection was conducted with informed consent from native speakers. Individuals cannot be identified or re-identified from the data to protect user privacy. The dataset is intended strictly for academic and research purposes. Care was taken to ensure diversity in audio/text samples and to avoid any content that could be considered harmful or offensive.

Acknowledgments

This work was mainly supported by the Malanshan Audio & Video Laboratory, China, and was partially supported by the Guangdong Basic and Applied Basic Research Foundation under Grant Nr. 2024A1515010112 and the Changsha Science and Technology Bureau Foundation under Grant Nr. kq2402082. We would like to thank all the native speakers who participated in the data collection and human inference process for their time and effort. Their support was essential in creating the DEBATE dataset.

References

[1] Oleg Akhtiamov, Dmitrii Ubskii, Evgeniia Feldina, Aleksei Pugachev, Alexey Karpov, and Wolfgang Minker. 2017. Are you addressing me? Multimodal addressee detection in human-human-computer conversations. In *Proc. 19th International Conference on Speech and Computer (SPECOM)*. Hatfield, UK, 152–161.

[2] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, et al. 2024. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *arXiv preprint arXiv:2407.04051* (2024).

[3] Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1363–1375.

[4] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*. Montreal, Canada, 4330–4338.

[5] Balthasar Bickel and Johanna Nichols. 2007. Inflectional morphology. *Language Typology and Syntactic Description* 3, 2 (2007), 169–240.

[6] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proc. 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. Seoul, Korea, 1–5.

[7] Chiara Bucaria. 2004. Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. *Humor* 17, 3 (2004), 279–309.

[8] Henry S Cheang and Marc D Pell. 2008. The sound of sarcasm. *Speech Communication* 50, 5 (2008), 366–381.

[9] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, 1025–1035.

[10] Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W Schuller. 2025. Towards multimodal prediction of spontaneous humor: A novel dataset and first results. *IEEE Transactions on Affective Computing* 16, 2 (2025), 844–860.

[11] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, et al. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919* (2023).

[12] Google DeepMindg. 2024. *Introducing Gemini 2.0: our new AI model for the agentic era.* Retrieved May 30, 2025 from https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

[13] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. New Orleans, LA, 18090–18108.

[14] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: Transforming mandarin ASR research into industrial scale. *arXiv preprint arXiv:1808.10583* (2018).

[15] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, et al. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Miami, FL, 6288–6313.

[16] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James R Glass. 2024. Listen, think, and understand. In *Proc. 12th International Conference on Learning Representations (ICLR)*. Vienna, Austria, 30 pages.

[17] Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised Chinese word sense disambiguation method based on Hownet. In *Proc. 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain, 1752–1757.

[18] Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics* 12 (2024), 738–754.

[19] Bo Li, Huan Zhao, and Zixing Zhang. 2023. Diversifying emotional dialogue generation via selective adversarial training. *Sensors* 23, 13, Article 5904 (2023), 16 pages.

[20] Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional Reference Grammar.* Univ of California Press.

[21] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, et al. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Singapore, 790–807.

[22] Esther Majambere. 2011. Clarity, precision and unambiguity: Aspects for effective legislative drafting. *Commonwealth Law Bulletin* 37, 3 (2011), 417–426.

[23] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. 9th International Workshop on Semantic Evaluation (SemEval)*. Denver, CO, 288–297.

[24] Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. 7th International Workshop on Semantic Evaluation (SemEval)*. Atlanta, GA, 222–231.

[25] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems* 6, 3, Article 25 (2016), 25 pages.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. International Conference on Machine Learning (ICML)*. Honolulu, HI, 28492–28518.

[27] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*. New York, NY, 8689–8696.

[28] Claudia Ross, Jing-heng Sheng Ma, Pei-Chia Chen, Baozhang He, and Meng Yeh. 2024. *Modern Mandarin Chinese Grammar: A Practical Guide.* Routledge.

[29] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, et al. 2017. The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proc. the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden, 3442–3446.

[30] Oswald Szemerényi and Oswald John Louis Szemerényi. 1999. *Introduction to Indo-European Linguistics.* OUP Oxford.

[31] Fiseha B Tesema, Jason Gu, Wei Song, Hong Wu, Shiqiang Zhu, Zheyuan Lin, et al. 2023. Addressee detection using facial and audio features in mixed human–human and human–robot settings: A deep learning framework. *IEEE Systems, Man, and Cybernetics Magazine* 9, 2 (2023), 25–38.

[32] Yue Wang, Qiliang Liang, Yaqi Yin, Hansi Wang, and Yang Liu. 2024. Disambiguate words like composing them: A morphology-informed approach to enhance Chinese word sense disambiguation. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Bangkok, Thailand, 15354–15365.

[33] Yue Wang, Hua Zheng, Yaqi Yin, Hansi Wang, Qiliang Liang, and Yang Liu. 2024. Morpheme Sense Disambiguation: A New Task Aiming for Understanding the Language at Character Level. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. Turin, Italy, 11605–11618.

[34] Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. 2021. Mumor: A multimodal dataset for humor detection in conversations. In *Proc. 10th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)*. Qingdao, China, 619–627.

[35] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215* (2025).

[36] Fukang Yan, Yue Zhang, and Zhenghua Li. 2023. Construction of a modern Chinese word sense dataset based on online dictionaries. In *Proc. 22nd Chinese National Conference on Computational Linguistics*. Harbin, China, 43–53.

[37] Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. SarcNet: A multilingual multimodal sarcasm detection dataset. In *Proc. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. Turin, Italy, 14325–14335.

[38] Qin Zhang, Sihan Cai, Jiaxu Zhao, Mykola Pechenizkiy, and Meng Fang. 2024. CHAmbi: A new benchmark on Chinese ambiguity challenges for large language models. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*. Miami, FL, 14883–14898.

[39] Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W Schuller. 2024. Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems* 11, 5 (2024), 6690–6704.

[40] Zixing Zhang, Weixiang Xu, Zhongren Dong, Kanglin Wang, Yimeng Wu, Jing Peng, et al. 2024. ParaLBench: A large-scale benchmark for computational paralinguistics over acoustic foundation models. *IEEE Transactions on Affective Computing* (2024). In press, 17 pages.

[41] Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for Chinese word sense disambiguation. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*. Punta Cana, Dominican Republic, 918–923.