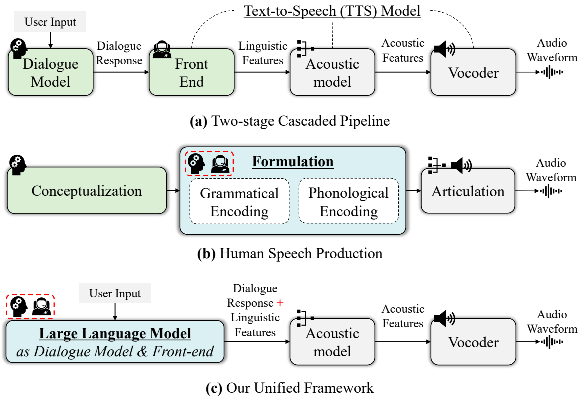
# Towards Joint Modeling of Dialogue Response and Speech Synthesis based on Large Language Model

**Xinyu Zhou**[1] **Delong Chen**[2] **Yudong Chen**[1]
[1]Communication University of China
[2]Hong Kong University of Science and Technology
{xinyuzhou, chenyd}@cuc.edu.cn, delong.chen@connect.ust.hk

## Abstract

This paper explores the potential of constructing an AI spoken dialogue system that "*thinks how to respond*" and "*thinks how to speak*" simultaneously, which more closely aligns with the human speech production process compared to the current cascade pipeline of independent chatbot and Text-to-Speech (TTS) modules. We hypothesize that Large Language Models (LLMs) with billions of parameters possess significant speech understanding capabilities and can jointly model dialogue responses and linguistic features. We conduct two sets of experiments: 1) Prosodic structure prediction, a typical front-end task in TTS, demonstrating the speech understanding ability of LLMs, and 2) Further integrating dialogue response and a wide array of linguistic features using a unified encoding format. Our results indicate that the LLM-based approach is a promising direction for building unified spoken dialogue systems.[1]

## 1 Introduction



**(a)** Two-stage Cascaded Pipeline

**(b)** Human Speech Production

**(c)** Our Unified Framework

Figure 1: **A high-level comparison of different speech production processes. As noted by the red dotted boxes, the novel LLM-based unified framework proposed in this study can *"think how to respond"* and *"think how to speak"* at the same time, which aligns better with the speech production process of humans.**

As we are developing more advanced AI systems, such as Large Language Model (LLM)-based chatbots like ChatGPT and GPT-4 (OpenAI, 2023), we also hope to establish natural, seamless, and efficient communication between humans and AI systems. In addition to typing and reading through the screen, the speech channel represents a valuable alternative for interpersonal exchange, given its convenience and capacity to convey richer information than text alone. Recently, researchers from both academia and the industry have made successful attempts to concatenate AI chatbots with off-the-shelf text-to-speech (TTS) (Tan et al., 2021) modules as in Figure 1 (a), representative applications include Siri, Xiaomi Xiaoai[2] and Call Annie[3].

However, the expressivity and interactivity of speech responses synthesized by these two-stage cascaded models are heavily limited. The reasons are two-fold. ***Firstly***, TTS modules are usually based on small language models (*e.g.,* BERT model with 0.1B parameters), which have limited capacity for understanding complex dialogue contexts. ***Secondly***, the dialogue response generation module (*i.e.,* the LLM Chatbot) and the TTS module work independently. During speech synthesizing, the TTS module can not access the information from the dialogue context, which is proven to be valuable for generating plausible and appropriate speech responses.

The current two-stage pipeline also has a fundamental difference with our understandings of the human speech production process (Levelt, 1993), where the *"grammatical encoding"* and *"phonological encoding"* are done in parallel within the "conceptualization-formulation-articulation" pro-

Xinyu Zhou and Delong Chen contributed equally.
This work is partially supported by National Social Science Fund of China (20&ZD295).

[1] Codes and datasets are publicly available at https://github.com/XinyuZhou2000/Spoken_Dialogue.

[2] https://xiaoai.mi.com

[3] https://callannie.ai

cess, as shown in Figure 1 (b). Inspired by this, we want to explore the possibility of building an AI speech dialogue system that *"thinks how to respond"* and *"thinks how to speak"* at the same time. In order to accomplish this goal, a model must possess a deep understanding of natural language and dialogue context, exhibit extensive world knowledge and commonsense, and demonstrate adequate learnability to handle text-speech joint modeling.

We hypothesize LLMs with (hundreds) billions of parameters (in comparison with BERT-based TTS front-ends (Chen et al., 2022) with only 0.3B parameters) are capable of achieving this goal. To verify this, in this paper, we provide two groups of experiments to demonstrate the possibility of building such LLM-based unified speech dialogue system.

***Firstly***, we get started with the prosodic structure prediction (Section 3), a typical task within the TTS text analysis front-end, to showcase the speech understanding ability of LLMs. Results show that both prompting-based ChatGPT and fine-tuning based ChatGLM (Zeng et al., 2022) model achieve competitive performance against traditional methods. We also show that LLM can utilize linguistic knowledge to improve prediction accuracy.

***Secondly***, we aim to further integrate a wide array of linguistic features into the model, and maintain LLM's dialogue capability at the same time (Section 4). To address the lack of a parallel dataset of dialogue response and linguistic annotations, we employ an automated dialogue context generation approach inspired by LongForm (Köksal et al., 2023), then train an LLM to produce both dialogue response speech features at the same time. Experiments show that LLM learns successfully.

## 2 Related Work

### 2.1 Human Speech Production

The process of human speech production is a longstanding research area. In 1993, Levelt (Levelt, 1993) proposed an encoding model for human speech production. First, concepts are generated, followed by the selection of appropriate vocabulary and the arrangement of these words according to grammatical rules. Then, the phonetics of the words are extracted in sequence, and motor programs are executed to initiate speech. The generation of spoken sentences is parallel and incremental, involving multiple stages of processing. Experiments (Schnur, 2011; Jaeger et al., 2012) prove that the phonetic planning of words begins as the grammatical structure of a sentence unfolds. Although there are many efforts to understand and explain human speech production process, TTS methods rarely take inspiration from these research results. To our best knowledge, this is the first study that attempts to build an AI system that imitates the simultaneous *"grammatical encoding"* and *"phonological encoding"* process of human speech production.

### 2.2 TTS Front-end and Expressivity

Typical TTS systems (Tan et al., 2021) usually consist of three main modules: front-end, acoustic model, and vocoder. The TTS Front-end models convert text into linguistic features, and are primarily BERT-based small language models, while the power of LLM is not well validated in this task yet. Hsu *et al.* (Hsu et al., 2021) and Stephenson *et al.* (Stephenson et al., 2022) have demonstrated that fine-tuning BERT can enhance the prosodic expression capabilities of TTS systems. Nevertheless, issues such as homograph ambiguity, ineffectiveness in stress, emotion and prosody still exist. Recent studies have explored the use of interactional resources (Chen, 2023), such as breathing (Székely et al., 2020), laughter (Xin et al., 2023), phonation type (Lameris et al., 2023), filled pauses and prolongations (Li et al., 2023), to improve the spontaneity and expressiveness. However, these studies have only focused on one single interactional resource, which limits their ability to capture rich and diverse subtle variations in natural conversation.

### 2.3 LLMs for Speech Processing

Understanding and generating speech signals are strongly related to natural language processing. With the recent explosion of LLM, many researchers in the field of speech processing also attempt to use LLMs to benefit speech or audio related tasks. AudioLM (Borsos et al., 2023) leverages a masked language model to capture the long-term structure and generate natural and coherent audio continuations given short prompts. SpeechGPT (Zhang et al., 2023), a multi-modal large language model, leverages its inherent capabilities to perceive and generate multi-modal content. PromptTTS (Guo et al., 2023) and PromptTTS2 (Leng et al., 2023) take prompts with both style and content descriptions as input to synthesize the corresponding speech.

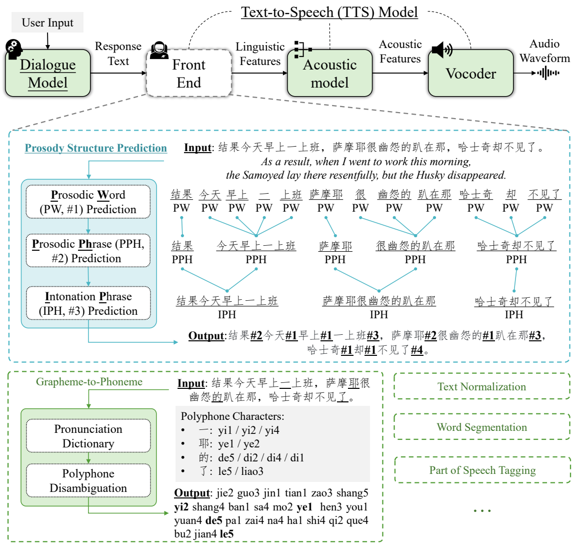

Text-to-Speech (TTS) Model
User Input
Dialogue Model
Response Text
Front End
Linguistic Features
Acoustic model
Acoustic Features
Vocoder
Audio Waveform
Prosody Structure Prediction
Input: 结果今天早上一上班，萨摩耶很幽怨的趴在那，哈士奇却不见了。
As a result, when I went to work this morning,
the Samoyed lay there resentfully, but the Husky disappeared.
Prosodic Word (PW, #1) Prediction
Prosodic Phrase (PPH, #2) Prediction
Intonation Phrase (IPH, #3) Prediction
结果 PW 今天 PW 早上 PW 一 PW 上班 PW 萨摩耶 PW 很 PW 幽怨的 PW 趴在那 PW 哈士奇 PW 却 PW 不见了 PW
结果 PPH 今天早上一上班 PPH 萨摩耶 PPH 很幽怨的趴在那 PPH 哈士奇却不见了 PPH
结果今天早上一上班 IPH 萨摩耶很幽怨的趴在那 IPH 哈士奇却不见了 IPH
Output:结果#2今天#1早上#1一上班#3，萨摩耶#2很幽怨的#1趴在那#3，哈士奇#1却#1不见了#4。
Grapheme-to-Phoneme
Pronunciation Dictionary
Polyphone Disambiguation
Input: 结果今天早上一上班，萨摩耶很幽怨的趴在那，哈士奇却不见了。
Polyphone Characters:
一：yi1 / yi2 / yi4
耶：ye1 / ye2
的：de5 / di2 / di4 / di1
了：le5 / liao3
Output: jie2 guo3 jin1 tian1 zao3 shang5 yi2 shang4 ban1 sa4 mo2 ye1 hen3 you1 yuan4 de5 pa1 zai4 na4 ha1 shi4 qi2 que4 bu2 jian4 le5
Text Normalization
Word Segmentation
Part of Speech Tagging
...

Figure 2: **Standard pipeline of current spoken dialog systems.** A dialogue model generates a response to user input, and the TTS model (front-end→acoustic model→vocoder) converts text to audio subsequently.

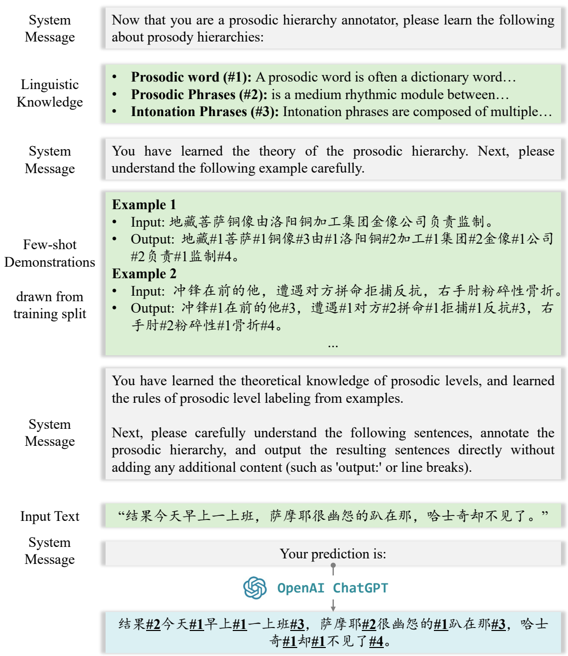
# 3 Prosodic Structure Prediction based on Large Language Model

Prosodic Structure Prediction (PSP) is a typical task in the Chinese TTS front-end (Chen et al., 2022), among others like grapheme-to-phoneme prediction, text-normalization, word segmentation, part-of-speech tagging, etc. As illustrated in Figure 2, a PSP model needs to identify multiple levels of prosody hierarchy, including Prosodic Word (PW), Prosodic Phrase (PPH), and Intonation Phrase (IPH), which can be denoted as #1, #2, and #3 respectively in the output sentence.

Prosodic structure is one of the most important linguistic features in Chinese TTS, and it is strongly related to the syntax of the sentence. In this section, we want to validate whether the LLMs, which have been well-proved to have superior semantic understanding abilities, but are trained on the text-only corpus, can handle this speech-related task. In the following, we present two methods for adapting LLM to the PSP task: prompting (Section 3.1), and fine-tuning (Section 3.2).

## 3.1 Prompting LLM for Prosodic Structure Prediction

Prompting is the most convenient way to adapt an instruction-following LLM to new tasks. In Figure 3, we present an overview of our proposed prompt structure for PSP on LLM, which consists of linguistic knowledge of Chinese prosodic structure, few-shot demonstrations as in-context learning examples, input sentence, and interleaved system messages to explain each part to the LLM.



Now that you are a prosodic hierarchy annotator, please learn the following about prosody hierarchies:

- **Prosodic word (#1):** A prosodic word is often a dictionary word…
- **Prosodic Phrases (#2):** is a medium rhythmic module between…
- **Intonation Phrases (#3):** Intonation phrases are composed of multiple…

You have learned the theory of the prosodic hierarchy. Next, please understand the following example carefully.

**Example 1**
- Input: 地藏菩萨铜像由洛阳铜加工集团金像公司负责监制。
- Output: 地藏#1菩萨#1铜像#3由#1洛阳铜#2加工#1集团#2金像#1公司#2负责#1监制#4。

**Example 2**
- Input: 冲锋在前的他，遭遇对方拼命拒捕反抗，右手肘粉碎性骨折。
- Output: 冲锋#1在前的他#3，遭遇#1对方#2拼命#1拒捕#1反抗#3，右手肘#2粉碎性#1骨折#4。

...

You have learned the theoretical knowledge of prosodic levels, and learned the rules of prosodic level labeling from examples.

Next, please carefully understand the following sentences, annotate the prosodic hierarchy, and output the resulting sentences directly without adding any additional content (such as 'output:' or line breaks).

“结果今天早上一上班，萨摩耶很幽怨的趴在那，哈士奇却不见了。”

Your prediction is:

结果**#2**今天**#1**早上**#1**一上班**#3**，萨摩耶**#2**很幽怨的**#1**趴在那**#3**，哈士奇**#1**却**#1**不见了**#4**。

Figure 3: **Our proposed prompt structure for LLM (ChatGPT)-based prosodic structure prediction.** We incorporate expert linguistic knowledge and few-shot demonstrations to enable LLMs to perform the prosodic structure prediction task.

***Linguistic Knowledge*** contains formal definitions of Chinese prosodic structure summarized from recognized research literature (Cao, 2003). It describes distinct characteristics and positions within sentences and phrases of three levels of prosodic structure in Chinese.

***Few-shot Demonstration*** provides input-output pairs to LLM for in-context learning. Examples are either randomly drawn from the training split or carefully selected based on the assessment of their representativity and quality from the linguistic perspective. The maximum number of few-shot demonstrations is 16, as more examples would exceed the context window length of LLM.

## 3.2 Fine-tuning LLM for Prosodic Structure Prediction

Context window length is a crucial limit for prompting-based methods, as it prohibits the LLM from learning from more (than 16) training examples. Furthermore, all model parameters remain fixed and unlearnable, resulting in limited learning capacity. To address these constraints, we propose the fine-tuning of a Large Language Model (LLM) to enhance Prosodic Structure Prediction learning from a substantially larger number of training examples, up to 8,000.

It has been proved that using a Pretrained Language Model (PLM) such as BERT (Devlin et al., 2018) to be the initialization of the PSP model is beneficial, such as SpanPSP (Chen et al., 2022), J-TranPSP (Shen et al., 2022), and MLC-PSP (Chen et al., 2023), our methodology of fine-tuning LLM has some difference from them. Despite the difference in model scale (0.1B vs 6B), previous BERT-based methods regard PSP as a *token classification* problem, where the model needs to determine whether there is a prosodic boundary after each character and what level is it. In contrast, here we formalize PSP as a sequence-to-sequence (Seq2seq) prediction task, where input $x$ is the raw sentence and the output $y$ is a string of character sequence with "#$n$" ($n \in \{1, 2, 3\}$) notation of prosodic structure.

We apply standard cross-entropy loss for auto-regressive language modeling as the learning objective, and we only calculate the loss on output tokens. We add a prefix $c$ of "Please perform prosodic prediction on the given sentence:" into the input for better initialization. The following is the loss function $\mathcal{L}(\theta)$ of the LLM $\theta$, where $N$ is the number of training samples: $\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p_\theta(y_i|x_i, c_i)$.

### 3.3 Experiment Setup

**Dataset**. We utilize the DataBaker open-source Chinese Standard Mandarin Speech Corpus[4], which contains 10-hour speech recordings of 10,000 sentences with an average length of around 16 words per sentence. It was articulated by a single Chinese female speaker. The corpus also encompasses diverse domains, including news, novels, technology, entertainment, etc.

Furthermore, the dataset is enriched with various linguistic annotations, including character, pinyin, and prosodic hierarchy information, as well as phoneme level interval and boundary data. Annotations for prosodic hierarchy comprise PW (#1), PPH(#2), IPH (#3), and the end of a sentence (#4). We discard the #4 annotations as every sample is a single sentence and only has a "#4" in the end. The remaining labels collectively form a hierarchical prosodic tree structure with three distinct layers.

We split 10k samples with an 8:1:1 ratio for training, validation, and testing. Few-shot demonstrations are drawn from the 8k training split. For the "random" selection setting, we sample demonstrations randomly three times and report the averaged performance.

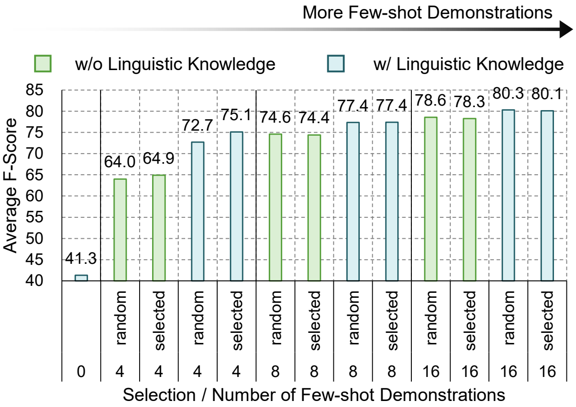


Figure 4: **Ablation study of prompting ChatGPT based PSP**. We compared different numbers of few-shot demonstrations, selection of few-shot demonstrations, and variants of with (w/) or without (w/o) linguistic knowledge.

**Implementation Details**. For the prompting-based method, we test the OpenAI's `text-davinci-002` API (ChatGPT) and the ChatGLM2-6B model. For the fine-tuning-based method, we only unitize the ChatGLM2-6B model due to the limitation of computational resources. We apply P-tuning-v2 (Liu et al., 2022) for parameter-efficient fine-tuning using the official codebase [5]. We used a single NVIDIA A100 GPU for both training and testing.

### 3.4 Ablation Study

We first provide ablations for the prompting-based approach. Following previous works on PSP tasks, we use F-Score as the evaluation metric. As it can be seen from Figure 4, the number of few-shot demonstrations makes a significant impact. Four in-context examples lead to +22.7% improvements to zero-shot setting (41.3%→64.0%), while incorporating linguistic knowledge brings another around +8.7% improvements (→72.7%), and further, when swapping random demonstrations to carefully selected high-quality demonstrations, we receive another +2.4% performance gain (→75.1%).

We further ablate different levels of linguistic knowledge in Table 2. It shows that linguistic expert knowledge plays a crucial role in the prediction of Prosodic Phrase (#2) and Intonational Phrase (#3). We hypothesize it is caused by different difficulties of #1 to #3 predictions – #1 usually appears at word boundaries, while identifying #2 and #3 is

[4] `https://www.data-baker.com/data/index/TNtts`

[5] `https://github.com/THUDM/ChatGLM2-6B/tree/main/ptuning`

Table 1: **Benchmarking of PSP Models**. We compared the F-Score of the traditional BERT-based method SpanPSP and our newly proposed LLM-based methods (prompting or fine-tuning) using two LLMs with different scales (ChatGPT and ChatGLM).

| Model (#Parameters) | Variation | PW #1 | PPH #2 | IPH #3 | Average |
|---|---|---|---|---|---|
| SpanPSP (0.1B) | Databaker Pretrained | **96.35** | 69.34 | 65.64 | 77.11 |
| | PeopleDaily Pretrained | 89.20 | 71.08 | 79.12 | 79.80 |
| ChatGPT (175B) | Knowledge Only | 61.87 | 27.27 | 34.78 | 41.31 |
| | 16 Random Examples | 88.51 | 69.40 | 77.91 | 78.61 |
| | Knowledge + 16 Selected Examples | 90.12 | 69.40 | **80.85** | 80.12 |
| ChatGLM2-6B | Knowledge + 16 Selected Examples | N/A | N/A | N/A | N/A |
| | Fine-tuned | 93.86 | **73.28** | 80.00 | **82.38** |

Table 2: **Ablations of removing each level of linguistic knowledge.** Expert knowledge is especially useful for higher levels of prosodic structure prediction (*i.e.,* PPH and IPH).

| Knowledge | PW #1 | PPH #2 | IPH #3 | Average |
|---|---|---|---|---|
| Ablation | F-Score | F-Score | F-Score | F-Score |
| w/o #1 | **88.54** | 64.66 | 78.30 | 77.17 |
| w/o #2 | 87.57 | **61.63** | 79.09 | 76.10 |
| w/o #3 | 87.72 | 64.69 | **78.14** | 76.85 |
| Default (all) | **88.14** | **65.03** | **79.52** | **77.56** |

not that straightforward.

### 3.5 Benchmarking LLM-based PSP

**Baseline.** SpanPSP (Chen et al., 2022) is a classical character-level BERT-based model for the PSP task, which is based on a relatively small language model `bert-base-chinese`[6] with only 0.1B parameters. We use their official checkpoints and codebase[7] for evaluation.

We provide benchmarking results in Table 1. It reveals that carefully crafted linguistic knowledge and selected examples (*i.e.,* "Knowledge + 16 Selected Examples" variation) enable ChatGPT to outperform the traditional method SpanPSP (80.12% vs. 79.80%), but such a prompting-based learning strategy failed (N/A) at smaller open-source LLM (ChatGLM) due to its limited instruction-following ability. However, it shows that fine-tuning smaller LLM can outperform prompting larger LLM (82.38% vs. 80.12%), as it can access more training samples (8k training set vs. the maximum of 16 in-context examples).

## 4 Joint Prediction of Dialogue Response and Linguistic Features

In the last section, we have shown some positive results proving LLMs are competitive at a typical front-end task in Chinese TTS. Here in this section, we want to go beyond just a single task in TTS, and validate the possibility of building a LLM-based system that can handle versatile tasks in Chinese TTS front-end (Figure 2), and also maintain its dialogue capability of generating coherent responses to user queries simultaneously. By implementing such a model that jointly predicts dialogue response (*i.e., "think how to respond"*) and linguistic features (*i.e., "think how to speak"*) at the same time, we could have an AI system that communicates with humans in a way that is more similar to the human speech production process (Levelt, 1993), where the "grammatical encoding" and "phonological encoding" are done in parallel.

### 4.1 Methodology

**Dialogue Context Generation**. Our objective is to investigate the feasibility of constructing a unified model capable of simultaneously generating coherent responses to user queries in dialogues and diverse fine-grained linguistic features for TTS. Unfortunately, the DataBaker dataset only comprises isolated sentence recordings, and there are not any other datasets having dialogue context and parallel speech recordings or annotations. Drawing inspiration from the LongForm approach (Köksal et al., 2023), we prompt ChatGPT to anticipate the dialogue context and transform it into a dataset of single-turn dialogues:

**Prompt for Dialogue Context Generation**

`### System Message:`

Please generate the most likely sentence spoken by A based on B's response.

`### User:`

A:

B: *"When I went to work this morning, the Samoyed lay there resentfully, but the Husky disappeared."*

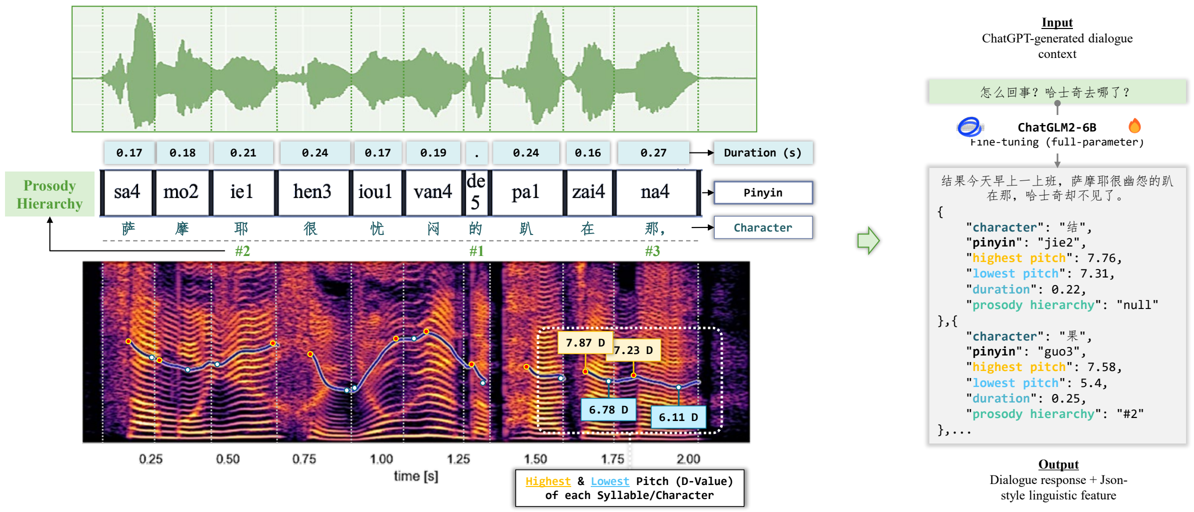
`### ChatGPT:`

A: *"What's going on? Where did the Husky go?"*

**Linguistic Feature Extraction**. As shown in Figure 5 left, we automatically extract the follow-

[6] https://huggingface.co/bert-base-chinese

[7] https://github.com/thuhcsi/SpanPSP



```
结果今天早上一上班，萨摩耶很幽怨的趴
在那，哈士奇却不见了。
{
    "character": "结",
    "pinyin": "jie2",
    "highest pitch": 7.76,
    "lowest pitch": 7.31,
    "duration": 0.22,
    "prosody hierarchy": "null"
},{
    "character": "果",
    "pinyin": "guo3",
    "highest pitch": 7.58,
    "lowest pitch": 5.4,
    "duration": 0.25,
    "prosody hierarchy": "#2"
},...
```

Figure 5: **Left:** an overview of linguistic feature extraction. We automatically extract a wide array of linguistic features, including character, duration, pinyin, prosody hierarchy, highest pitch and lowest pitch ($\mathcal{D}$-Value). **Right:** the illustration of data formatting. We encode extracted linguistic features into JSON-formatted strings, such that they can be fed to LLM directly as learning targets.

ing four categories of linguistic attributes: characters, their corresponding duration, pinyin (phonetic transcription representing character pronunciation), prosodic hierarchy, and the highest and lowest pitch values ($\mathcal{D}$-Value). The use of $\mathcal{D}$-value is inspired by Shen Jiong's theory (Shen, 1985): the D-value is a logarithmic scale used to describe pitch and quantifies the relationship between a pitch ($F$) in Hertz (Hz) and a reference frequency ($F_0$). It provides a measure of pitch variation, which is especially useful for observing pitch contours in speech. The formal definition of ($\mathcal{D}$-Value) is: $\mathcal{D} = 5 \times \log_2(F/F_0)$.

**Data Formatting**. As shown in the right side of Figure 5, we format extracted linguistic features into a string of JSON-style dictionaries, and concatenate it with the response text generated by ChatGPT, together serving as the learning target. Such implementation realizes joint learning in a seamless way and enjoys simplicity over the traditional method, where different types of outputs (response, various linguistic features) are usually produced by different models, or one model with different task-specific heads (Bai and Hu, 2021). Our approach also shares some similarities with recent advances in LLM research, such as RT-2 (Brohan et al., 2023) from Google DeepMind, where the LLM are trained to produce not only natural language output but also some continuous values.

## 4.2 Experiment Setup

**Training**. Empirically, we found that P-tuning (as used in fine-tuning-based PSP in Section 3.2) failed to learn how to generate dialogue response and JSON-style linguistic features. Therefore, for this section, we turn to use full-parameter fine-tuning to enable more learning capacity. We use 4-bit quantization to boost memory efficiency, as JSON-style encoding takes much longer context than that in the PSP task (maximum 1.6k tokens vs. 128 tokens).

Table 3: **Evaluation result of LLM produced linguistic features.** Most of the output JSON-style is incorrect grammar and parsable, and the majority of these parsable characters can be matched with ground truth. However, we can observe a notable train-test performance gap, meaning that the model suffers from overfitting.

| | Parsable Samples | Matched Characters | Matched Pinyin | Matched Prosody |
|---|---|---|---|---|
| Training Split | 95.90% | 86.88% | 98.79% | 97.75% |
| Testing Split | 89.70% | 69.26% | 86.29% | 77.70% |

**Testing**. Based on our data formatting (Figure 5), given a user utterance as input, the model will first give its dialogue response, then the JSON-style linguistic feature of each word in the response sentence subsequently. However, this poses a challenge for the evaluation of the linguistic feature, since for unseen testing quires, the LLM-outputted response would be different from the ground truth response, thus making them not comparable. To solve this issue, we use the ground truth response as a generation prefix, and then try to parse the generated dictionary and compare them with ground truth linguistic features.

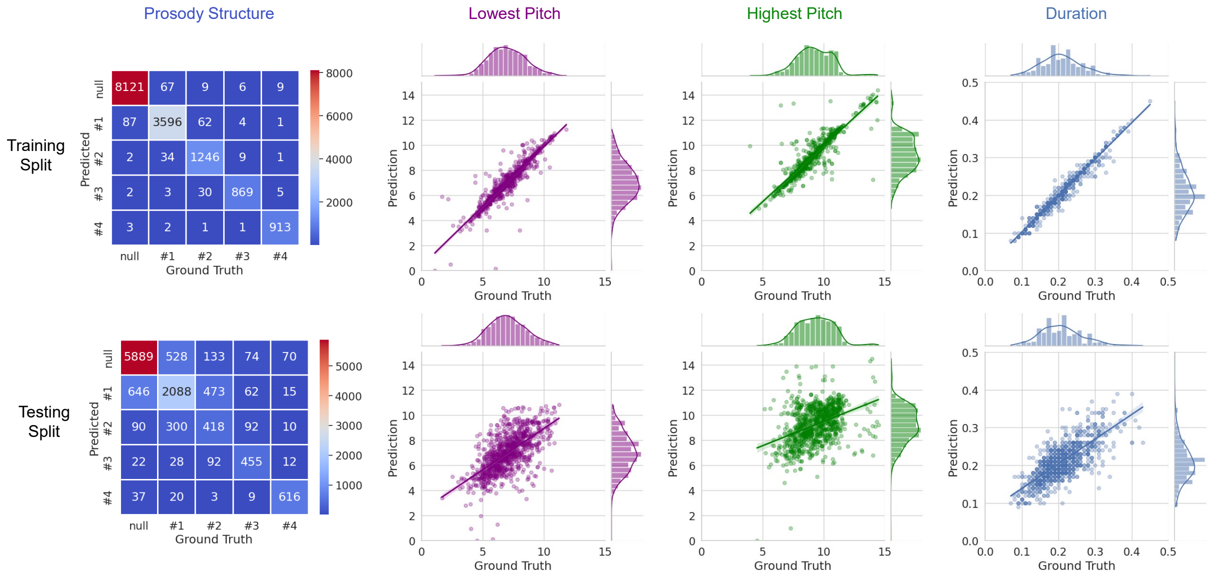


Figure 6: **Evaluations results of fine-tuning ChatGLM2-6B on joint dialogue response and linguistic features.** Visualizations show that the model fits the training data quite well, showing the feasibility of our proposed joint learning approach. But possibly due to insufficient dataset scale, the generalization ability of the model is somewhat weak.

### 4.3 Experiment Results

In Table 3, we provide the evaluation result of 1k testing samples and randomly sampled 1k training samples. As can be seen, the model performed quite well on the training set, achieving 95.90% parsable samples, 86.88% matched characters, and 98.79% matched Pinyin, showing that it successfully fit the JSON-format data. When tested on unseen samples, the model successfully generated Json-style linguistic features with an 89.70% success rate. However, due to the limited capacity of the small LLM, missing characters were frequently observed, resulting in only 69.26% of characters from the ground truth being found in the generated results. Within those matched characters, the model achieved an 86.29% success rate in producing matched Pinyin, and a 77.70% success rate of matched prosody structure annotation.

In Figure 6, we visualize the model predictions versus the ground truth of continuous values. Again, we observed that the model effectively fit the training set and demonstrated a certain level of generalization ability when applied to new data.

## 5 Discussion

In this study, we presented two groups of experiments to validate the possibility of building an LLM-based spoken dialog system that *"thinks how to respond"* and *"thinks how to speak"* at the same time. In the first group of experiments, we proved that LLM is a competitive prosodic structure predictor, which means that its rich world knowledge and semantic understanding ability acquired from text-only pretraining can transferred to benefit speech-related tasks. Based on this observation, we further involve many other linguistic features in our second group of experiments and further proved that it is possible for LLM to learn to generate dialogue response and speech features at the same time. However, there are still several noticeable limitations of this study, which are summarized from the following four perspectives:

**Model Perspective**. The training cost of LLM is high. Additionally, the auto-regressive decoding of Json-style linguistic features is quite time-consuming – processing a single sentence and its linguistic features takes at least 15 seconds (for long sentences, it could be 40+ seconds).

**Data Perspective**. The current training dataset consists of only 8k samples, which is insufficient and has led to a substantial over-fitting phenomenon. Speech style in the dataset is limited, as it was sourced from a single speaker, primarily containing formal read recordings, lacking the nuances inherent in natural conversations.

**Expressivity Perspective**. According to interactional linguistics studies (Couper-Kuhlen and Selting, 2017), finer-grained annotation system by comprehensively and meticulously annotating the collected speech dataset with interactional resources like voice quality, phonation type, breath patterns, repair, interjection, pause, prolongation, etc, will further increase the expressivity.

**System Perspective**. It is important to note that so far this study does not include subsequent acoustic models and vocoders and is not able to generate audio waveform, we only use speech linguistic features to represent speech information.

## References

Zilong Bai and Beibei Hu. 2021. A universal bert-based front-end model for mandarin text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Jianfen Cao. 2003. Prediction of prosodic organization based on grammatical infomation (in chinese). *Journal of Chinese Information Processing*, (41-46).

Jie Chen, Changhe Song, Deyi Tuo, Xixin Wu, Shiyin Kang, Zhiyong Wu, and Helen Meng. 2023. Improving mandarin prosodic structure prediction with multi-level contextual information. *arXiv preprint arXiv:2308.16577*.

Xueyuan Chen, Changhe Song, Yixuan Zhou, Zhiyong Wu, Changbin Chen, Zhongqin Wu, and Helen Meng. 2022. A character-level span-based model for mandarin prosodic structure prediction. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7602–7606. IEEE.

Yudong Chen. 2023. Studies of prosodic expressions in interactive language: A review (in chinese). *Contemporary Linguistics*, 25(197-222).

Elizabeth Couper-Kuhlen and Margret Selting. 2017. *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

T Florian Jaeger, Katrina Furth, and Caitlin Hilliard. 2012. Incremental phonological encoding during unscripted sentence production. *Frontiers in Psychology*, 3:481.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. Longform: Optimizing instruction tuning for long text generation with corpus extraction. *arXiv preprint arXiv:2304.08460*.

Harm Lameris, Shivam Mehta, Gustav Eje Henter, Joakim Gustafson, and Éva Székely. 2023. Prosody-controllable spontaneous tts with neural hmms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. 2023. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*.

Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

Weiqin Li, Shun Lei, Qiaochu Huang, Yixuan Zhou, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2023. Towards spontaneous style modeling with semi-supervised pre-training for conversational text-to-speech synthesis. *arXiv preprint arXiv:2308.16593*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Tatiana T Schnur. 2011. Phonological planning during sentence production: Beyond the verb. *Frontiers in Psychology*, 2:319.

Binbin Shen, Jian Luan, Shengyan Zhang, Quanbo Shen, and Yujun Wang. 2022. J-tranpsp: A joint transition-based model for prosodic structure prediction, word segmentation and pos tagging. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 280–284. IEEE.

Jiong Shen. 1985. Pitch range of tone and intonation in beijing dialect (in chinese). *Experiments on Beijing Dialect. Beijing: Peking University*, pages 73–130.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. Bert, can he predict contrastive focus? predicting and controlling prominence in neural tts using a language model. *arXiv preprint arXiv:2207.01718*.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2020. Breathing and speech planning in spontaneous speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7649–7653. IEEE.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Detai Xin, Shinnosuke Takamichi, Ai Morimatsu, and Hiroshi Saruwatari. 2023. Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus. *arXiv preprint arXiv:2305.12442*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pretrained model. *arXiv preprint arXiv:2210.02414*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.