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

Xinyu Zhou（周欣宇）
Master Student at Communication University of China
xinyuzhou@cuc.edu.cn

Delong Chen（陈德龙）
Ph.D. Student at HKUST
delong.chen@connect.ust.hk

AMLaP Asia 2023

Background and Motivation

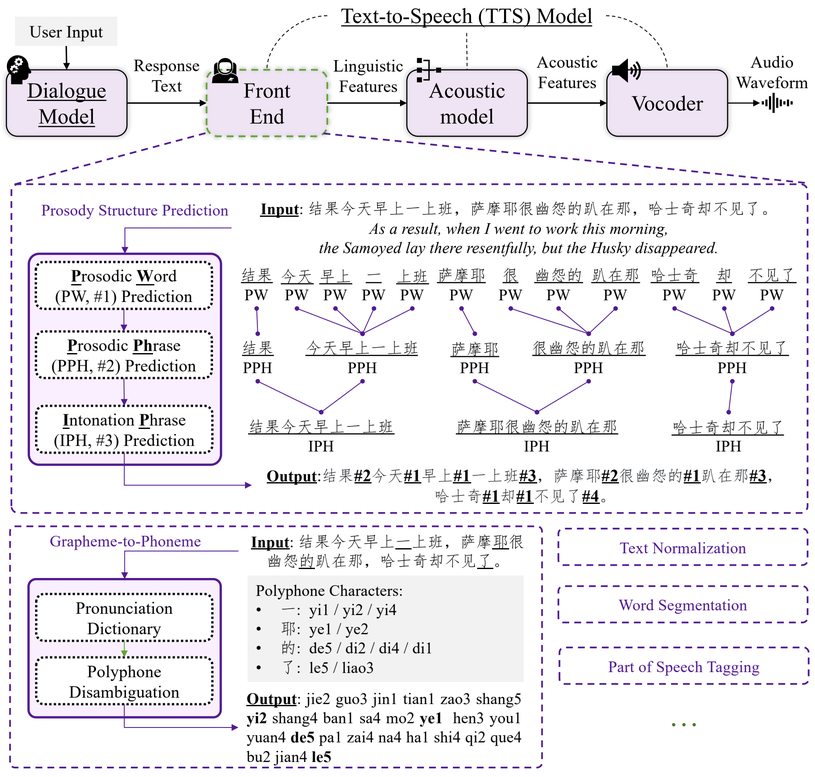

Figure 1: Standard pipeline of current Dialogue-TTS systems (*e.g.,* Apple Siri, Amazon Alexa, XiaoMi XiaoAI, Call Annie). A dialogue model generates response to user input, and the TTS model (front-end→acoustic model→vocoder) convert text to audio subsequently.

The current framework (Fig. 1) works fine, but has the following two limitations:

1. **Dialogue model and TTS model work independently**. It prohibits the TTS model to obtain the user input and generate speech with proper articulations (emotion, emphasize, etc.) accordingly.
2. **TTS front-end is based on small language models.** Generation of proper speech articulations requires in-depth understanding of complex dialogue context, which is a difficult task for low-capacity language models.

This study aims to utilize Large Language Models (LLMs) to address this issue. As in Fig. 2, we hope the LLM can "*think how to response*" and "*think how to speak*" at the same time. Moreover, its strong language understanding ability encoded rich world knowledge could enhance its ability to generate improved linguistic features.

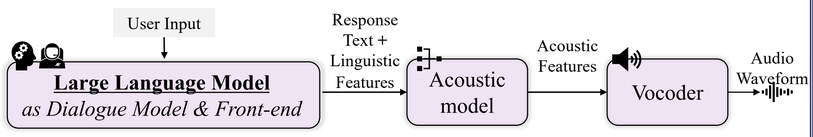

Figure 2: The proposed framework of this study.

This framework is inspired by the model of speech production by Levelt *et al.* (Fig.3), which consists of 1) conceptualization, 2) formulation, and 3) articulation. These stages are processed *incrementally* and *in parallel*. Compared to the standard pipeline where different modules work *separately* and *independently*, our proposed framework aligns better with this human speech model.

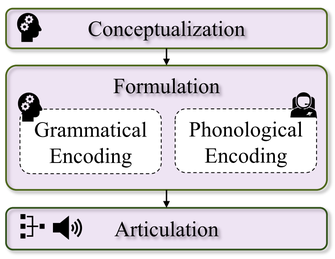

Figure 3. The model of human speech production by Levelt *et al.* (1989) [1].

Speech Data

This project utilized DataBaker 标贝科技 DataBaker Technology open-source Chinese Standard Mandarin Speech Corpus, which contains:

- Speech Recording
  - Total duration: about 12 hours (10000 sentences);
  - Average sentence length: about 16 words per sentence;
  - Speaker: one Chinese female;
  - Domain : news, novels, technology, entertainment, etc.
- Linguistic Annotations
  - Character, pinyin and prosody hierarchy;
  - Phoneme interval and boundary.

结果#2今天#1早上#1一上班#3，萨摩耶#2很幽怒的#1趴在那#3，哈士奇#1却#1不见了#4。
jie2 guo3 jin1 tian1 zao3 shang5 yi2 shang4 ban1 sa4 mo2 ye1 hen3 you1 yuan4 de5 pa1 zai4 na4

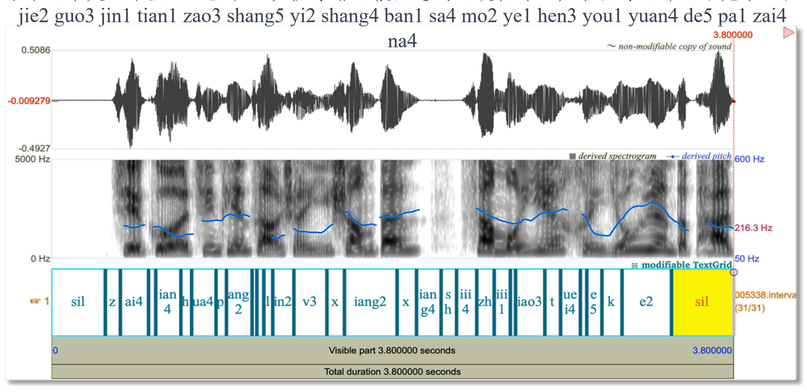

Figure 4. Visualization of audio recording and annotations of one sample in the dataset.

Large Language Model as TTS Front-end: A Case Study of Prosodic Structure Prediction

Prompting LLM for Prosodic Structure Prediction

System Message	Now that you are a prosodic hierarchy annotator, please learn the following about prosody hierarchies:
Linguistic Knowledge [2]	• **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…
System Message	You have learned the theory of the prosodic hierarchy. Next, please try to understand the following example carefully.
Few-shot Demonstrations drawn from training split	**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。 ...
System Message	You have learned the 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).
Input Text	"结果今天早上一上班，萨摩耶很幽怒的趴在那，哈士奇却不见了。"
System Message	Your prediction is:

OpenAI ChatGPT (text-davinci-002 API)

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

Figure 5: Prompt design of LLM (ChatGPT)-based zero-shot instructed prosody structure prediction.

Ablation Study

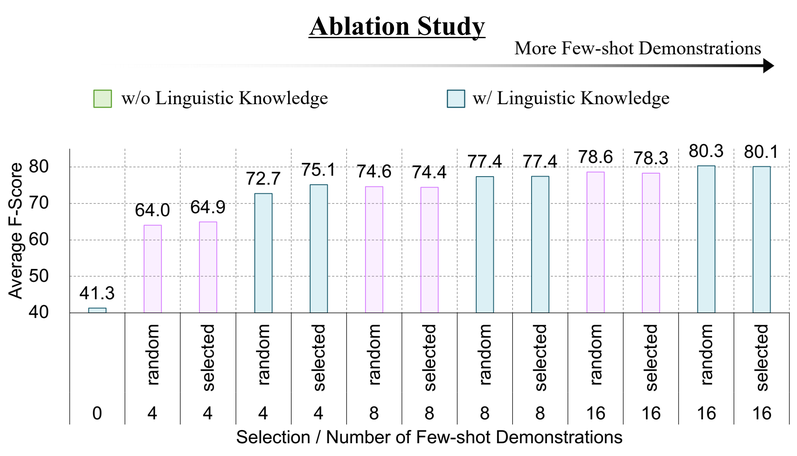

Figure 6. Ablation study of selection & number of shots, linguistic knowledge for ChatGPT-based prosody structure prediction.

Observations

Table 1. Ablations of removing each level of knowledge.

Knowledge Ablation	PW #1 F-Score	PPH #2 F-Score	IPH #3 F-Score	Average 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**

1. Linguistic knowledge helps, especially for PPH and IPH.
2. Carefully selection of demonstrations yields improvements in low-shot settings.
3. More shots lead to better performance but longer (near maximum) context length.

Fine-tuning Smaller LLMs

In respond to observation 3, we tried to fine-tune the ChatGLM2-6B model, which is an open-source bilingual (English & Chinese) LLM. It is much smaller than ChatGPT (6B vs. 175B parameters) thus can be efficiently trained on local machine.

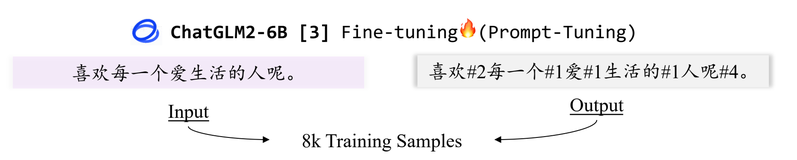

Figure 7: Illustration of fine-tuning LLM (ChatGLM2-6B) for prosody prediction.

Performance Comparison

Table 2. Benchmarking results of traditional BERT-based method and prompting or fine-tuning LLM based methods. Carefully crafted linguistic knowledge and selected examples enable ChatGPT to outperform traditional method SpanPSP, but such strategy failed (N/A) at smaller open-source LLM (ChatGLM) due to its limited capacity. However, we proved that fine-tuning smaller LLM can outperform prompting larger LLM, as it can access more training samples (8k training set vs maximum 16 in-context examples)

Model	Params (Billion)	Variation	PW #1 F-Score	PPH #2 F-Score	IPH #3 F-Score	Average F-Score
SpanPSP [2] (BERT-based)	0.3	Databaker Pretrained	**96.35**	69.34	65.64	77.11
		PeopleDaily Pretrained	89.20	71.08	79.12	79.80
ChatGPT (text-davinci-002)	175	Knowledge Only	61.87	27.27	34.78	41.31
		16 Random Examples	88.51	69.40	77.91	78.61
		Knowledge + 16 Selected Demo	90.12	69.40	**80.85**	80.12
ChatGLM-6B	6	Knowledge + 16 Selected Demo	N/A	N/A	N/A	N/A
		Fine-tuned	93.86	**73.28**	80.00	**82.38**

Joint Prediction of Dialogue Response and Linguistic Features via LLM

Data Preparation

We want to verify the possibility of building a unified model that learns to generate *both* proper dialogue response to user query *and* various fine-grained linguistic features for TTS (Fig. 2).

However, the available data only contains recordings of independent sentences. Inspired by LongForm [4], we prompt ChatGPT to predict the dialogue context and convert it into a single-turn dialogue dataset (Fig. 8).

Input Text	A: B: 结果今天早上一上班，萨摩耶很幽怒的趴在那，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.*
System Message	Please generate the most likely sentence spoken by A based on B's response. A:

OpenAI ChatGPT (gpt-3.5-turbo API)

Generated Context	怎么回事？哈士奇去哪了 *What's going on? Where did the Husky go?*

Figure 8: Prompt structure for automated context generation. ChatGPT infers possible user utterance based on the given response sentence.

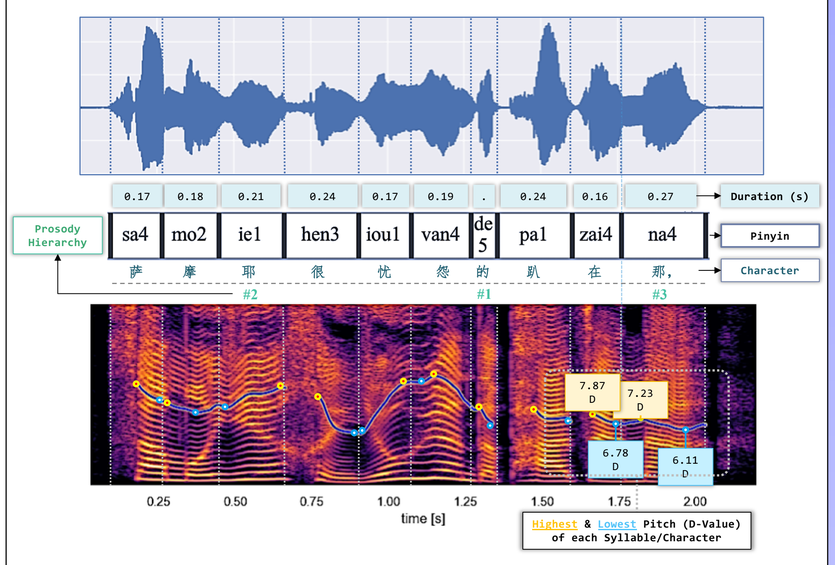

Figure 9: Extraction of linguistic feature, including character, pinyin, prosody hierarchy, highest pitch and lowest pitch (D-Value). The use of D-Value is inspired by Shen Jiong's theory [6]: the D-value is a logarithmic scale used to describe pitch and quantifies the relationship between a pitch (F) in Hertz and a reference frequency (F0). It is defined as $D = 5\times\log_2(F/F0)$. It provides a measure of pitch variation, which is especially useful for observing pitch contours in speech.

Model Training and Evaluation

Figure 10: Illustration of fine-tuning ChatGLM2-6B LLM for dialogue response and linguistic features generation.

As shown in Fig. 10, we format extracted linguistic features (Fig. 9) into string of Json-style dictionary, and concatenate it with the response text, together serving as the learning target. Instruction tuning on such dataset enables LLM to jointly predict dialogue response and speech features.

During testing (Table 3 and Fig.11), we use the ground truth response as generation prefix, and try to parse the generated dictionary and compare them with ground truth linguistic features.

	Parsable Samples	Matched Characters	Matched Pinyin
Train Split	95.90%	86.88%	98.79%
Test Split	89.70%	69.26%	86.29%

Table 3. On unseen testing samples, the model successfully generated Json-style linguistic features with a 89.70% success rate. However, due to the limited capacity of small LLM, missing characters are frequently observed, and only 69.29% characters in ground truth can be found in generation results.

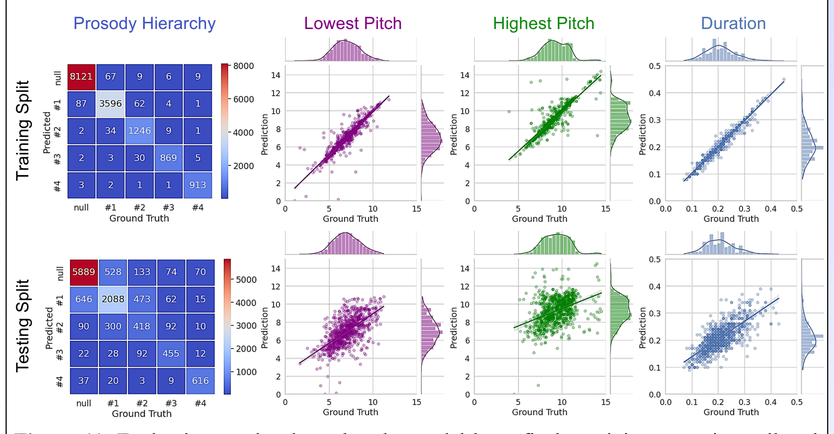

Figure 11: Evaluation results show that the model have fit the training set quite well and demonstrated certain generalization abilities.

Limitations and Future Work

- **Model** Perspective
  - **Training Cost is High.** Fine-tuning ChatGLM2-6B for 10,000 steps consumes 12 hours on a single NVIDIA A100 machine, which is considerably expensive. By contrast, fine-tuning for prosody structure prediction (as depicted in Fig.7) takes approximately 1 hour.
  - **Inference is Time-consuming.** The auto-regressive decoding of Json-style linguistic features lacks parallelization, resulting in time-consuming inference. Processing a single sentence can take up to at least 15 seconds. More efficient methods for encoding linguistic features are imperative to address this challenge.
- **Data** Perspective
  - **Data Scale is Insufficient.** The training dataset consisting of only 8,000 samples is inadequate for training a large language model. Consequently, significant overfitting occurs as depicted Table 3 and Fig. 11.
  - **Speech Style is Limited.** The existing dataset is sourced from a solitary female speaker, encompassing solely formal speech recordings. However, these recordings tend to sound more like reading rather than natural speech, thereby lacking the distinctive nuances inherent in communication through voice. Therefore, we plan to use a bigger conversation speech dataset with multiple speakers in different scenarios.
- **TTS** Perspective
  - **Finer-grained Speech Annotations.** We plan to adopt an annotation paradigm inspired by interactional linguistics and conversation analysis which emphasis on suprasegmental elements as well as voice quality, breath patterns, repair, pause and other resources
  - **No Acoustic Models and Vocoder.** The subsequent acoustic models and vocoders are not included in this project.

References

[1] Levelt, W. J., & JM, W. **Speaking: From intention to articulation - A bradford book**. MIT Press, 1989.

[2] 曹剑芬. 基于语法信息的汉语韵律结构预测. 中文信息学报 2003. (Cao Jianfen. **Prediction of Prosodic Organization Based on Grammatical Information**. *Journal of Chinese Information Processing* 2003.)

[3] Chen, X., Song, C., Zhou, Y., *et al*. **A character-level span-based model for mandarin prosodic structure prediction**. In *IEEE ICASSP 2022*.

[4] THUDM Group, Tsinghua University. **ChatGLM2-6B: An Open Bilingual Chat LLM**. Url: https://github.com/THUDM/ChatGLM2-6B. 2023.

[5] Köksal, A., Schick, T., Korhonen, A., *et. al.* **Longform: Optimizing instruction tuning for long text generation with corpus extraction**. *arXiv preprint 2023*.

[6] 沈炯. 北京话声调的音域和语调. 《北京语音实验录》. 北京大学出版社, 1985. (Shen Jiong. **The Pitch Range of Tone and Intonation in Beijing Dialect**. In *Experiments in Beijing Phonetics*, Peking University Press, 1985.)