INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Assessing Phrase Break of ESL Speech with Pre-trained Language Models and Large Language Models

Zhiyi Wang[1,2*], *Shaoguang Mao*[2], *Wenshan Wu*[2], *Yan Xia*[2], *Yan Deng*[2], *Jonathan Tien*[2]

[1]Peking University, Beijing, China
[2]Microsoft Research Asia, Beijing, China

wangzhy@stu.pku.edu.cn, shaoguang.mao@microsoft.com, wenshan.wu@microsoft.com, yanxia@microsoft.com, yaden@microsoft.com, jtien@microsoft.com

Abstract

This work introduces approaches to assessing phrase breaks in ESL learners' speech using pre-trained language models (PLMs) and large language models (LLMs). There are two tasks: overall assessment of phrase break for a speech clip and fine-grained assessment of every possible phrase break position. To leverage NLP models, speech input is first force-aligned with texts, and then pre-processed into a token sequence, including words and phrase break information. To utilize PLMs, we propose a pre-training and fine-tuning pipeline with the processed tokens. This process includes pre-training with a replaced break token detection module and fine-tuning with text classification and sequence labeling. To employ LLMs, we design prompts for ChatGPT. The experiments show that with the PLMs, the dependence on labeled training data has been greatly reduced, and the performance has improved. Meanwhile, we verify that ChatGPT, a renowned LLM, has potential for further advancement in this area.

Index Terms: phrase break, computer-aided language learning, ESL speech, pre-trained language models, large language models

1. Introduction

Proper phrase break is crucial to oral performance [1] and is always a challenge for English as a Second Language (ESL) learners. There has been considerable research in computer-aided language learning (CALL) [2, 3, 4, 5]. As for the phrase break assessment, there are two main categories: 1) break feature extraction and modeling [6, 7]. 2) modeling against reference speech [8, 9]. For example, a method was proposed to evaluate break by computing similarity between the assessed speech with utterances from native speakers or Text-to-Speech (TTS) system [8].

Although modeling against reference speech [8, 9] is an effective approach to assessing speech performance, some errors unavoidably occur when it comes to handling diverse phrase break cases. As shown in Figure 1, the correct phrase break patterns for the same text are various, and thus it is not to say that the phrasing is incorrect if it is different with template audios. The previous work fails to consider this fact. Instead, they model the phrase break like a fixed pattern prediction. Meanwhile, a large scale of high-quality human-labeled data is required for traditional methods. However, the subjective labeling is costly and the labeling consistency is hard to satisfy [2, 10, 11]. How to construct robust models with small datasets is still under research.

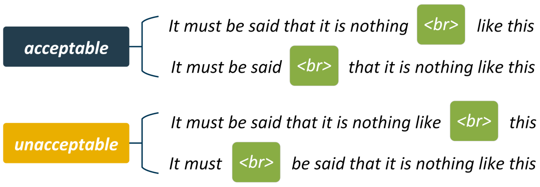

acceptable
It must be said that it is nothing <br> like this
It must be said <br> that it is nothing like this
unacceptable
It must be said that it is nothing like <br> this
It must <br> be said that it is nothing like this

Figure 1: *Example of diverse phrase break patterns(<br> represents a phrase break)*

Phrase break prediction is a traditional task in the TTS area [12, 13, 14, 15]. In Futamata's work [12], a phrase break prediction method is proposed that combines implicit features extracted from BERT [16] and explicit features extracted from Bidirectional Long Short-Term Memory (Bi-LSTM) with linguistic features. The goals of phrase break prediction and phrase break assessment are different, and the second one being much harder considering the diverse break facts. We can refer to the idea that the break information can be inferred from input text, and leverage the power of rising pre-trained language models (PLMs) [17, 18] and large language models (LLMs) [19].

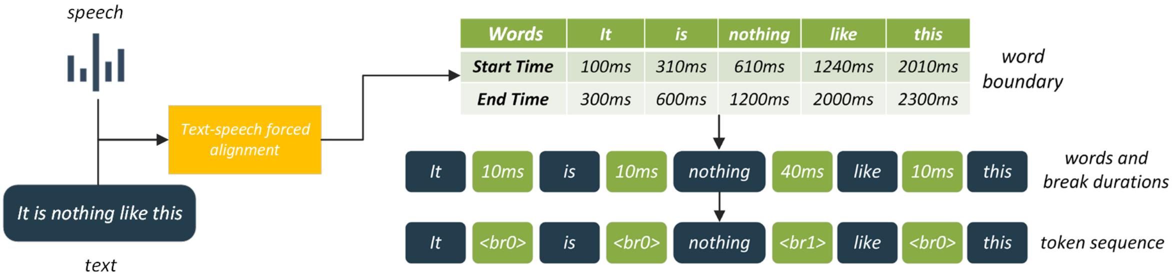
This paper presents approaches to assessing phrase break with PLMs and LLMs. In particular, there are two sub-tasks: assessment of phrase break for a speech and fine-grained assessment of each break position. To adopt those NLP models, each speech is processed into a token sequence with text-speech forced alignment [20, 21, 22], referencing Figure 2. The token sequence consists of words and associated phrase break tokens (break duration information for each between-words interval).

To adopt the pre-training models, a self-supervised replaced break token detection strategy is proposed. Each break token from the original sample has 15% chance of being replaced by other break tokens. Then, a discriminator is trained with augmented data riding on BERT [16] to identify whether the token sequence is edited. In the fine-tuning stage, the overall assessment and fine-grained assessment are fine-tuned with text classification and token classification, respectively. Additionally, by providing suitable prompts, LLMs can perform well on many NLP tasks and adapt for specific use-cases with just a few task examples [19, 23, 24]. Therefore, we design prompts and investigate the zero-shot and few-shot learning [25, 26, 27] setups with ChatGPT [28].

The main contributions are: first, this is pioneering work to explore the use of PLMs and LLMs for speech assessment. The experimental results demonstrate the possibility of using language models to perform speech evaluation in specific tasks. Second, this work takes diverse phrasing patterns into consideration to construct a more precise assessment.

[*] Work performed as an intern in Microsoft Research Asia

10.21437/Interspeech.2023-910



Words	It	is	nothing	like	this
Start Time	100ms	310ms	610ms	1240ms	2010ms
End Time	300ms	600ms	1200ms	2000ms	2300ms

Figure 2: *Overview of the data pre-processing. The speech-text forced alignment tool recognizes word boundaries and the duration between adjacent words. Then, the token sequence is obtained by converting the duration to break tokens with the mapping in Table 1 .*

2. Data Pre-processing

2.1. Task definition

We use two tasks to demonstrate how PLMs and LLMs can be applied into assessing phrase break.

One is predicting a rank r for a test speech to indicate its overall performance on phrase break. The other one is that given a speech S, consisting n words, predict a rank r_i for each interval b_i between two words on whether the phrase break is appropriate, including whether an existing break is appropriate and whether an expected break is missed.

2.2. Pre-processing

To leverage the power of PLMs and LLMs, the speech clips are first converted to a token sequence with speech-text forced alignment.

As shown in Figure 2, speech-text forced alignment is used to recognize the phrase break and duration between every pair of adjacent words w_i and w_{i+1}. Based upon the statistical information and linguists' assessment, the phrase breaks are categorized into four types, as shown in Table 1. A speech utterance is then tokenized into a token sequence $T : \{w_0, b_0, w_1, ..., w_i, b_i, w_{i+1}, ..., w_n\}$, including words and phrase break tokens.

Table 1: *The definition of break tokens*

Type	Duration	Comment
br0	(0, 10ms]	No break
br1	(10ms, 50ms]	Slight / Optional break
br2	(50ms, 200ms]	Break
br3	(200ms, $+\infty$)	Long break

3. Approach

3.1. Pre-trained Language Models for Break Assessment

Replaced Break Token Detection We introduce a pre-training approach named replaced break token detection. As shown in Figure 3, speech recordings by native speakers from TTS corpus are collected as original samples because TTS recordings have good performance in phrase break. Then, each sample is randomly corrupted several times with the strategy that each break token has 15% chance to be replaced with other kinds of break tokens. 15% is a hyper-parameter settled by pre-experiments. The proportions of different types of breaks after the random corruptions with a 15% change and the real speech by non-native speakers are very similar.

The pre-training is from BERT as it is trained on a large scale of texts and learns contextual relations between words (or sub-words) [16]. A discriminator is trained with cross-entropy loss on the augmented data to predict whether the input sequence has been corrupted. The trained model is called Break-BERT for convenience.

Downstream Tasks The overall assessment task is treated as a sequence classification task. The model predicts a rank r for a token sequence, $r \in R$. The model consists of the head of the pre-trained model and a classifier on top, and is trained with cross-entropy loss.

The fine-grained assessment is treated as a sequence labeling task. An $r_i \in R$ is expected to be assigned to b_i. There is a token classification head on top of the hidden-state output from pre-trained model. It is also trained with a cross-entropy loss function.

3.2. Large Language Models for Break Assessment

We investigate the potential of ChatGPT for phrase break assessment in zero-shot and few-shot scenarios.

Prompts By taking into account the crucial impact that prompting has on output from generative models, we clarify our problem according to section 2.1, input scoring rubric that annotators adopt, and then standardize input and output formats for our task.

The input is formatted as T in section 2.2. The output is formatted as rank $r, r \in R$ and inappropriate break position set $P : \{p_1, ..., p_j, ...p_n\}, p_j = w_i b_i w_{i+1}$ represents that the phrase break b_i between word w_i and word w_{i+1} is inappropriate.

Zero Shot or Few Shot Learning We try zero-shot and few-shot learning to explore the potential of LLMs in speech phrase break assessment. For the zero-shot scenario, ChatGPT responses without specifically training on any data. A zero-shot example of question prompt and response is shown in Figure 4. For the few-shot learning, a few examples are provided as context information to enable the model be adapted for annotated cases. Details for case selection in few-shot learning will be discussed in the next section.

4. Experiments

4.1. Corpora

We collected 800 audio samples from different Chinese ESL learners. Then, two linguists are invited to assess them with overall performance on phrase break and each individual phrase

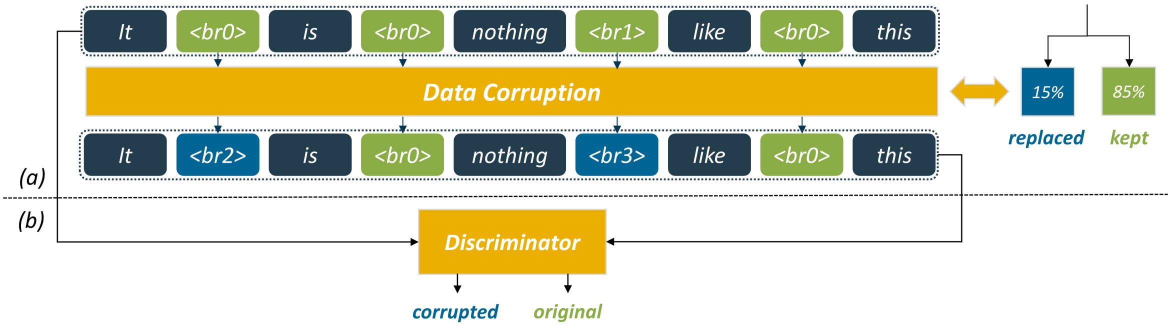
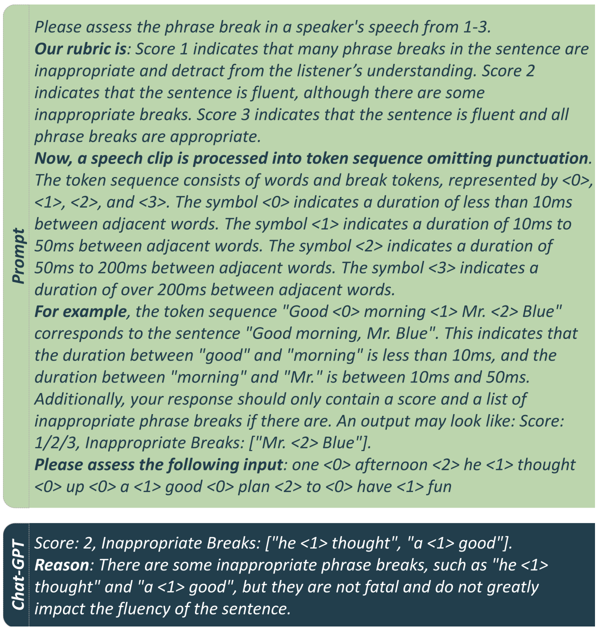


Figure 3: *An overview of the replaced break token detection pre-training process. Part (a) describes an example of data corruption where each break token in the original token sequence has 15% chance of being replaced with other tokens. Part (b) is the pre-training stage where a discriminator is trained to distinguish original or corrupted data.*



Please assess the phrase break in a speaker's speech from 1-3.
Our rubric is*: Score 1 indicates that many phrase breaks in the sentence are inappropriate and detract from the listener's understanding. Score 2 indicates that the sentence is fluent, although there are some inappropriate breaks. Score 3 indicates that the sentence is fluent and all phrase breaks are appropriate.*
Now, a speech clip is processed into token sequence omitting punctuation*. The token sequence consists of words and break tokens, represented by <0>, <1>, <2>, and <3>. The symbol <0> indicates a duration of less than 10ms between adjacent words. The symbol <1> indicates a duration of 10ms to 50ms between adjacent words. The symbol <2> indicates a duration of 50ms to 200ms between adjacent words. The symbol <3> indicates a duration of over 200ms between adjacent words.*
For example*, the token sequence "Good <0> morning <1> Mr. <2> Blue" corresponds to the sentence "Good morning, Mr. Blue". This indicates that the duration between "good" and "morning" is less than 10ms, and the duration between "morning" and "Mr." is between 10ms and 50ms. Additionally, your response should only contain a score and a list of inappropriate phrase breaks if there are. An output may look like: Score: 1/2/3, Inappropriate Breaks: ["Mr. <2> Blue"].*
Please assess the following input*: one <0> afternoon <2> he <1> thought <0> up <0> a <1> good <0> plan <2> to <0> have <1> fun*

Score: 2, Inappropriate Breaks: ["he <1> thought", "a <1> good"].
Reason*: There are some inappropriate phrase breaks, such as "he <1> thought" and "a <1> good", but they are not fatal and do not greatly impact the fluency of the sentence.*

Figure 4: *An example of prompt and response from ChatGPT.*

break ranging from 1 (Poor), 2 (Fair), and 3 (Great). If two experts' opinions are inconsistent, an extra linguist will intervene and do the final scoring. The statistics of collected corpus are listed in Table 2. All data is publicly available for research on `https://github.com/Chris0King/phrasing-break-assessment`.

Table 2: *Statistics of downstream datasets.*

Dataset	Poor	Fair	Great	Total
Overall	21	136	643	800
Fine-grained	129	644	10797	11570

4.2. Pre-training Setups

The data for the downstream task was obtained from recordings of 800 Chinese students during ”read-after-me” exercises. Thus, we utilized the TTS corpus, a speech dataset from reading scenarios, for pre-training purposes. The pre-training was performed using the LJ Speech dataset [29], which is commonly known for its good phrase breaks and diverse break patterns. There are 22.5 hours of speech recordings in the training set and 1 hour in the test set, containing 192k words and 8k words respectively.

In data corruption, the ratio between the corrupted samples and the original samples is 3:1, i.e. for each original sample, three random corrupted samples are augmented. The pre-training begins from BERT$_{\mathbf{BASE}}$, and a simple linear classifier is added on the top. It is trained with a batch size of 64 for 3 epochs over the dataset. The maximum sequence length is set to 128. We used back propagation and Adam optimizer with a learning rate of 1e-4. After the pre-training, the binary classification results of Break-BERT achieve 83.9% in accuracy and 89.7% in f-score.

4.3. Experimental Setup

Baselines Bi-LSTM+Linear Layer and Bi-LSTM+CRF (Conditional Random Field) [30] are set as baselines for overall and fine-grained assessment, separately. We apply Bi-LSTM as a backbone considering it still works well in a relatively small dataset. The hidden layer size is set to 1024. Meanwhile, a direct fine-tuning with downstream data on BERT is conducted to verify the validity of the proposed pre-training process. The baseline models take the identical token sequences by the BERT tokenizer.

We also adopt the Against-TTS method [8] as a baseline and tag the output break similarity score [0, 0.3), [0.3, 0.7), [0.7, 1.0] as poor, fair, great, separately. The adopted TTS system is from Microsoft Cognitive Service en-US-AriaNeural [1].

Meanwhile, ChatGPT was asked to assess the phrase break from 1-3 defined in 4.1 and list all inappropriate breaks. For few-shot learning, we evaluate each example in the test set by randomly selecting four samples from the corresponding training set as context information to maintain the similar distribution. The prompt and few-shot learning strategy are determined through preliminary experiments. All examples are tested on text-chatdavinci-002 (using OpenAI's playground). The temperature is set 0 to ensure a consistent prediction.

Cross-validation We apply five-fold cross-validation to avoid instability of sampling and report the mean and standard deviation of experiments.

[1] `https://learn.microsoft.com/en-us/azure/cognitive-services/`

Table 3: *Performance of overall and fine-grained assessment models. '#' stands for 'Fine-tune' and 'w/' stands for 'with'.*

Assessment Model		**Metric avg.(std)**		
		Acc.	F-Score(weighted)	F-Score(macro)
Overall	Bi-LSTM	80.2(6.4)	76.4(9.6)	39.2(8.1)
	Against-TTS	54.4(9.9)	61.1(7.1)	36.3(5.6)
	#BERT	80.4(6.5)	77.9(7.0)	40.9(7.1)
	#Break-BERT	**82.5(5.0)**	**81.7(5.7)**	**52.3(10.5)**
	ChatGPT w/ Zero-shot Learning	55.6(6.2)	61.6(4.9)	40.6(3.7)
	ChatGPT w/ Few-shot Learning	65.8(5.8)	70.5(4.8)	47.3(4.0)
Fine-grained	Bi-LSTM	92.5(3.9)	90.1(5.6)	39.9(3.7)
	Against-TTS	70.9(2.6)	78.6(4.0)	31.1(1.5)
	#BERT	91.8(4.1)	89.0(5.8)	39.5(4.1)
	#Break-BERT	**92.8(3.1)**	**91.6(4.0)**	**44.3(2.5)**

Table 4: *Comparing the performance of #Break-Bert and ChatGPT on fine-grained assessment, category 1 (Poor) and category 2 (Fair) are mapped to inappropriate break, and category 3 (Great) is mapped to appropriate break.*

	#Break-BERT		**ChatGPT w/ Zero-shot**		**ChatGPT w/ Few-shot**	
Category	Precision	Recall	Precision	Recall	Precision	Recall
Poor and Fair	**60.1%**	33.7%	26.5%	31.4%	26.9%	32.6%
Great	**94.8%**	**98.4%**	94.4%	94.4%	94.5%	93.1%

4.4. Results

Accuracy, weighted f-score and macro f-score are taken as metrics [10]. As shown in Table 3 and Table 4, compared with Bi-LSTM, Against-TTS, fine-tuning on BERT and ChatGPT, the proposed pre-training fine-tuning greatly improves all metrics. The knowledge learned from the pre-training stage efficiently enhances model performance. It is worth mentioning that the Against-TTS system performs much worse than the proposed approach and ChatGPT. More discussions are included in the next section.

5. Discussion

5.1. Influence of Pre-training

The pre-training process takes TTS human recordings as correct samples, where multiple phrase break patterns exist. After a series of random corruptions, the augmented samples are likely to be incorrect in phrasing. After the pre-training on original and constructed incorrect patterns, the discriminator has learned general linguistic patterns and phrase break information through self-supervised learning. The experiments verified the assumptions. The proposed model yields better results. The knowledge learned from pre-training benefits downstream tasks.

5.2. How Diverse Breaks are Handled

The experimental results verified Against-TTS approach's limits on handling multiple possible phrase breaks. As shown in Table 5, there are sharp drops of the recall of category 3 (Great) and the precision of category 1 (Poor), while the precision of category 3 (Great) and the recall of category 1 (Poor) are kept. For a test speech, if it shows a different phrase break pattern with reference audio, it tends to be classified as poor even if it is correct. When it shows a similar phrase break pattern to the template, it is highly possible to be a correct phrasing. This explains the high precision, low recall for category 3 (Great), as well as the high recall, low precision of category 1 (Poor).

Table 5: *Performance analysis on different categories.*

	Against-TTS		**#Break-BERT**	
Category	Precision	Recall	Precision	Recall
Poor	4.3%	**28.6%**	**50.0%**	14.3%
Fair	26.7%	46.3%	**49.2%**	**47.8%**
Great	87.3%	57.5%	**89.7%**	**92.4%**

5.3. LLMs for Break Assessment

According to experiments, we noticed that the ChatGPT model with zero-shot learning exhibits partial understanding of punctuation breaks. However, it tends to overlook the slight pauses between semantic groups, such as the initiation of a clause or phrase. Despite some improvement with few-shot learning setting, the ChatGPT model still struggles to adequately address the breaks between semantic groups and manifests unstable performance when attempting to rectify incorrect breaks.

While not reaching the state-of-the-art performance, the potential of ChatGPT in phrase break assessment is noteworthy. After few-shot learning, all metrics improve significantly in overall assessment task. We believe that with further optimization in prompt design, ChatGPT has the potential to demonstrate greater power in speech assessment.

6. Conclusions

This work presents new approaches to tackling ESL speech phrase break assessment with pre-trained language models (PLMs) and large language models (LLMs). The introduction of PLMs greatly minimizes the requirements for collecting labeled data, and the proposed self-supervised learning can handle multiple possible phrase break patterns of the same text. Also, we verify that ChatGPT, a classical and renowned LLM, has potential for further advancement in this area. In the future, leveraging PLMs and LLMs to solve other prosody assessment tasks, like intonation and stress, is well worth researching.

7. References

[1] M. Fach, "A comparison between syntactic and prosodic phrasing." in *Eurospeech*, vol. 99. Citeseer, 1999, pp. 527–530.

[2] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. Soong, "Nn-based ordinal regression for assessing fluency of esl speech," in *Proc. ICASSP*. IEEE, 2019, pp. 7420–7424.

[3] B. Lin, L. Wang, H. Ding, and X. Feng, "Improving l2 english rhythm evaluation with automatic sentence stress detection," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 713–719.

[4] S. Mao, F. Soong, Y. Xia, and J. Tien, "A universal ordinal regression for assessing phoneme-level pronunciation," in *Proc. ICASSP*. IEEE, 2022, pp. 6807–6811.

[5] W. Hu, Y. Qian, F. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[6] K. Fu, S. Gao, X. Tian, W. Li, Z. Ma, and A. Bytedance, "Using fluency representation learned from sequential raw features for improving non-native fluency scoring," in *Proc. Interspeech*, 2022, pp. 4337–4341.

[7] K. Sabu and P. Rao, "Automatic assessment of children's oral reading using speech recognition and prosody modeling," *CSI Transactions on ICT*, vol. 6, no. 2, pp. 221–225, 2018.

[8] Y. Xiao and F. Soong, "Proficiency assessment of esl learner's sentence prosody with tts synthesized voice as reference." in *INTERSPEECH*, 2017, pp. 1755–1759.

[9] J. Proença, G. Raboshchuk, Costa, P. Lopez-Otero, and X. Anguera, "Teaching american english pronunciation using a tts service." in *SLaTE*, 2019, pp. 59–63.

[10] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[11] H. Meng, W. Lo, A. Harrison, P. Lee, K. Wong, W. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support chinese learners of english: The cuhk experience," in *Proc. APSIPA ASC*, 2010, pp. 811–820.

[12] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis," *arXiv preprint arXiv:2104.12395*, 2021.

[13] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using wav2vec 2.0," in *International Conference on Text, Speech, and Dialogue*. Springer, 2022, pp. 377–388.

[14] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, "Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 274–285, 2020.

[15] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *Proc. ICASSP*. IEEE, 2016, pp. 5655–5659.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] L. Dong, N. Yang, W. Wang, F. Wei, Y. Liu, X.and Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] V. Mathad, T. Mahr, N. Scherer, K. Chapman, K. Hustad, J. Liss, and V. Berisha, "The impact of forced-alignment errors on automatic pronunciation evaluation." in *Interspeech*, 2021, pp. 1922–1926.

[21] P. Moreno, C. Joerg, J. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments." in *ICSLP*, vol. 98, 1998, pp. 2711–2714.

[22] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. ICASSP*. IEEE, 2009, pp. 4869–4872.

[23] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[24] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[25] S. Ravi and H. Larochelle, "as a model for few-shot learning," in *International conference on learning representations*, 2017.

[26] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[27] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[28] OpenAI, "Chatgpt," https://chat.openai.com/, 2022.

[29] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[30] M. Rei, "Semi-supervised multitask learning for sequence labeling," *arXiv preprint arXiv:1704.07156*, 2017.