More differentiated pause insertion for phoneme-based multi-speaker TTS models *

☆ Dong Yang[1], Tomoki Koriyama[2], Yuki Saito[1], Takaaki Saeki[1], Detai Xin[1], Hiroshi Saruwatari[1] ([1]The University of Tokyo, [2]CyberAgent, Inc.)

1 Introduction

Pause insertion is an essential part of text-to-speech (TTS) systems because proper pauses with natural duration significantly enhance the rhythm of synthetic speech. The position prediction of pauses is called phrase break prediction or phrasing, which is mainly solved by using neural networks with word representations nowadays [1, 2, 3]. However, conventional phrasing models ignore various speakers' different styles of inserting silent pauses, which degrades the performance of the model trained on a multi-speaker speech corpus. Besides, most mainstream TTS models (e.g., FastSpeech 2 [4]) mainly use phonemes as input. They treat all silent pauses as one phoneme, which leads to the duration of all silent pauses in synthetic speech following the same distribution and not being sufficiently differentiated. Therefore, we propose a pause insertion framework for multi-speaker phoneme-based TTS models, in which silent pauses are categorized by duration. In our previous work, bidirectional encoder representations from transformers (BERT) [5] pre-trained on a large-scale text corpus performed very well in predicting English phrase breaks [6]. Our approach in this study improves on the previous model by adding multi-task learning and injecting speaker embeddings to capture various speaker features.

2 Proposed Method

We first categorize the silent pauses by duration through the Gaussian mixture model-based method in [7]. We specify the pauses as three categories: brief (< 300 ms), medium (300–700 ms), and long (> 700 ms). Besides, there are two main types of silent pause: respiratory pauses (RPs) [8, 1] and punctuation-indicated pauses (PIPs). The former is inserted at word transitions without punctuation to utter long sentences fluently, and the latter is inserted at punctuation marks. Since their position, frequency, and duration distribution are different, we design the categorized pause insertion (CPI) model with a multi-task learning framework shown in Fig. 1. We use the encoder-decoder structure and take BERT as the encoder. Two sets of BiLSTM layers decode the output of the last layer (hidden sequence) in BERT and speaker embeddings (initialized randomly) that correspond to the predictions of RPs and PIPs. **Probability** and **Category** in Fig. 1 represent position prediction and category prediction, respectively. In practice, the model first predicts **Probability** that represents the occurrence of pauses and then outputs **Category** with the highest probability among the three categories. These predicted pauses are input into TTS models as phonemes, the three categories of which can be denoted as “sp1”, “sp2”, and “sp3” whereas we usually denote silent pauses as ”sp” in TTS tasks.

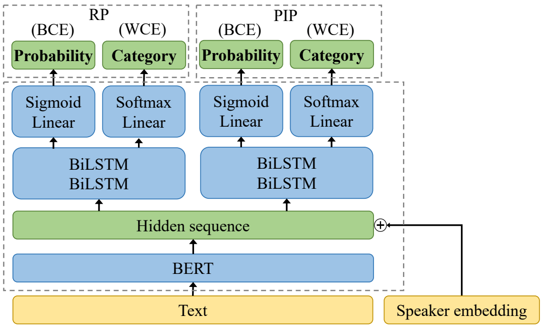

Fig. 1 Architecture of proposed method.

3 Experiments

3.1 Dataset

We constructed the dataset from LibriTTS [9], a multi-speaker English corpus derived from audiobooks. LibriTTS includes plenty of long-form sentences containing multiple silent pauses uttered by more than 2,000 speakers, and thus fits our purpose of evaluating the performance of multi-speaker pause prediction. We used the Montreal Forced Aligner (MFA) [10] to align text and speech and obtain pause durations. Because MFA recognizes the silence at word transitions over 30 ms as silent pauses, we regarded silent pauses over 30 ms at punctuation marks as PIPs and those over 50 ms at word transitions without punctuation as RPs.

3.2 Baseline

Klimkov et al. provided the mainstream framework of a phrasing model based on English audiobooks [1], which was implemented as our baseline. In the baseline model, the input word2vec embeddings were processed by BiLSTM projection with peephole connection (BiLSTMP) [11] layers and splicing windows. The splicing window stacked together seven frames before and after and fed them to the next layer. The model output the probability of the occurrence of a pause after each token.

3.3 Experimental Configuration

For the baseline model, the hidden size and projection size for each BiLSTMP layer were 512 and

* 音素ベース複数話者 TTS モデルにおけるより差別化されたポーズ挿入, 楊 棟 (東大院・情報理工) , 郡山 知樹 (サイバーエージェント), 齋藤 佑樹, 佐伯 高明, 辛 徳泰, 猿渡 洋 (東大院・情報理工)

Table 1 Results of position prediction of RPs.

	Precision	Recall	F_β
Baseline	0.393	0.187	$\mathrm{F}_{0.5} = 0.322$
CPI	0.575	0.261	$\mathrm{F}_{0.5} = 0.463$

Table 2 Confusion matrix of category prediction.

	Prediction of RPs			Prediction of PIPs		
Label	1	2	3	1	2	3
1	**2,565**	885	0	**6,155**	1,766	2,058
2	300	**513**	0	2,258	**3,186**	2,509
3	14	20	**0**	335	352	**2,735**

128, respectively. The dimension of word2vec embeddings was 300. For the proposed model, we configured the BERT as $\mathrm{BERT}_{\mathrm{BASE}}$[1] and set the hidden size of each BiLSTM layer to 512. The dimensions of the hidden sequence and speaker embeddings were 768. Word2vec embeddings and $\mathrm{BERT}_{\mathrm{BASE}}$ were pre-trained on BookCorpus [11]. The loss function was binary cross-entropy (BCE) loss for **Probability** and weighted cross-entropy (WCE) loss for **Category**. The weights of categories 1–3 in WCE loss function equaled the total number of non-pause tokens divided by their total number in the training set. As an exception, we set the weight of category 3 of RPs to 1.0 due to its sparseness.

During training, each mini-batch had 32 sentences. We used the Adam optimizer and set the initial value of the learning rate to 5×10^{-5}. The learning rate dropped by 0.2 times when the model's performance on the validation set did not improve within 5000 iterations. We took 200,000 as the maximum iterations (about 16 epochs). The models that performed best on the validation set during these iterations were saved to make comparisons.

3.4 Objective Evaluations

For phrasing models, their performance is usually measured by position prediction of RPs because position prediction of PIPs is a relatively simple task. The results are shown in Table. 1, in which our proposed CPI model performed much better than the baseline. Word representations from pre-trained BERT and speaker embedding bring a huge boost to the model. Table 2 shows the results of category prediction of CPI model. In category prediction, the accuracy of category 2 was lower than that of the other two categories, which suggests that there is still some room for improvement in the choice of thresholds for categorization.

3.5 Subjective Evaluations

To explore the performance of our proposed models in multi-speaker TTS, especially to show the improvement of inputting categorized pause phonemes, we performed AB preference tests using FastSpeech 2 as our TTS model with HiFi-GAN [12] as the vocoder. We trained two TTS models with silent pause phonemes, one with non-categorized pauses, and one with categorized pauses. We selected 16 speakers (eight males, eight females) with good synthetic sound quality and 123 long sentences with more than 1 RPs from the test set. Each sentence corresponds to several speakers, and then 277 text-speaker pairs were used in our listening tests. We asked native listeners from Amazon Mechanical Turk to participate in the tests. Every test was completed by 30 listeners, each of whom listened to ten pairs of synthetic speech and was asked to choose the one with better rhythm.

Table 3 Subjective performance of CPI.

Method A	Score	Method B
CPI(RPs)	**0.560 vs. 0.440**	FastSpeech2
CPI(RPs)	0.537 vs. 0.463	Baseline
CPI	**0.557 vs. 0.443**	Baseline
CPI(RPs)	0.488 vs. 0.512	CPI(Position)
CPI(RPs)	**0.460 vs. 0.540**	CPI

The results are shown in Table. 3, in which Baseline and CPI(PRs) only predicted the position of RPs, and CPI(Position) predicted the position of RPs and PIPs. Test scores with p-values below 0.05 are bold-faced. From the results, listeners perceived the difference between CPI(RPs) and Baseline to be insignificant, which shows people are insensitive to the position of RPs. People only became aware when listening to a long sentence without a pause, that was why CPI(RPs) performed a little better than Baseline. Listeners were more likely to recognize that CPI performed better than Baseline and CPI(RPs). Coupled with the inability of the listeners to distinguish significantly between CPI(RPs) and CPI (Position), we can conclude that inputting categorized pause phonemes to phoneme-based TTS models makes the rhythm of synthetic speech better.

4 Conclusion

In this paper, we proposed a categorized pause insertion model. The results of objective evaluations showed that word representations from pre-trained BERT and speaker embeddings can bring a large improvement to phrasing models trained with a multi-speaker dataset. The results of subjective evaluations showed that by inserting categorized pauses, the synthetic speech had better rhythm.

References

[1] V. Klimkov et al., in *INTERSPEECH*, 2017.
[2] K. Futamata et al., in *INTERSPEECH*, 2021.
[3] A. Abbas et al., in *INTERSPEECH*, 2022.
[4] Y. Ren et al., in *ICLR*, 2021.
[5] J. Devlin et al., in *NAACL-HLT*, 2019.
[6] D. Yang et al., in *ASJ*, 2022.
[7] E. Campione, J. Véronis, in *Speech Prosody*, 2002.
[8] G. Bailly, C. Gouvernayre, in *INTERSPEECH*, 2012.
[9] H. Zen et al., in *INTERSPEECH*, 2019.
[10] M. McAuliffe et al., in *INTERSPEECH*, 2017.
[11] Y. Zhu et al., in *ICCV*, 2015.
[12] J. Kong et al., in *NeurIPS*, 2020.

[1] `https://huggingface.co/bert-base-uncased`