Duration-Aware Pause Insertion Using Pre-Trained Language Model for Multi-Speaker Text-to-Speech

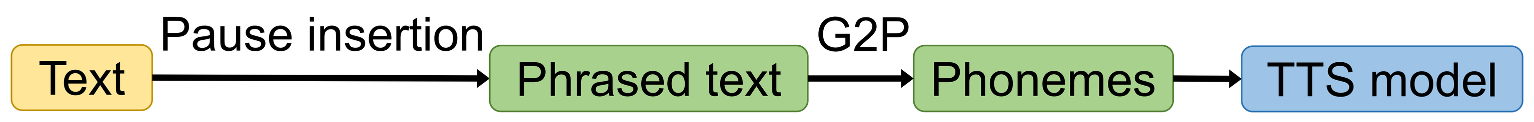
Dong Yang[1], Tomoki Koriyama[2], Yuki Saito[1], Takaaki Saeki[1], Detai Xin[1], Hiroshi Saruwatari[1]
([1]The University of Tokyo, Japan, [2]CyberAgent, Inc., Japan.)

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

- We propose two multi-speaker pause insertion models: the respiratory pause insertion (**RPI**) model and the categorized pause insertion (**CPI**) model.
- The **RPI** model is a **phrasing model** that performs **speaker-conditioned** position prediction of respiratory pauses.
- The **CPI** model is further designed for more natural multi-speaker TTS and predicts the **duration-aware pause marks**.

Background

- **Pause insertion** (a.k.a. phrase break prediction or **phrasing**)
  - Inserting proper silent pauses in TTS
  - Crucial for enhancing the rhythm of synthetic speech
  - Phrasing: position prediction of RPs



- Types of silent pauses
  - Respiratory pauses (**RPs**)
    - Inserted at word transitions without punctuation mark
  - Punctuation-indicated pauses (**PIPs**)
    - Inserted at punctuation marks

Lucy said: " (PIP) An Edgerunner will take me (RP) to the moon." (PIP)

- Conventional methods [1, 2, 3]
  - **Ignoring speaker's different styles** of inserting silent pauses in phrasing
    - ⇨ Performance declines when trained on a multi-speaker speech corpus
  - Treating all silent pauses as **one mark** during speech synthesis
    - ⇨ Duration of silent pauses in synthetic speech are not well differentiated
- Proposed methods
  - Injecting **speaker embeddings**
    - ⇨ Capturing various speaker characteristics in phrasing
    - ⇨ The phrasing model's predictive accuracy is improved significantly
  - Categorizing silent pauses by duration
  - Representing them with several marks
  - Inputting the **categorized pause marks** to the TTS model
    - ⇨ The synthetic speech has better rhythm
    - ⇨ More consistent with the speaker's feature

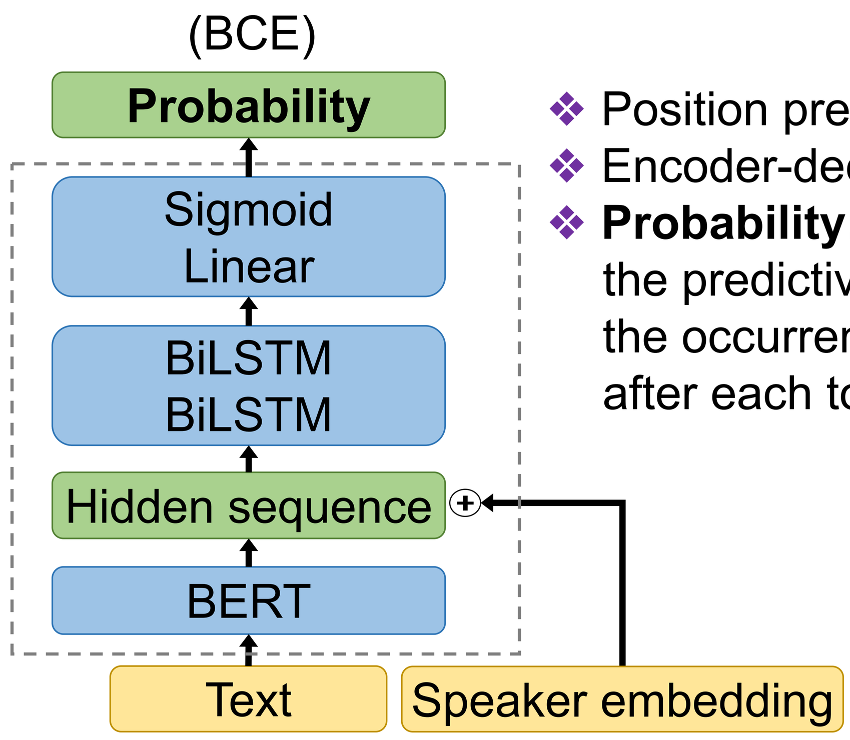
Dataset

- Corpus: LibriTTS
- Speech alignment & pause recognition
  - Aligner module of Montreal Forced Aligner
- Thresholds of categorization
  - Gaussian mixture model-based method [4]

Raw text	*Lucy said: "An Edgerunner will take me to the moon."*
Pre-processing	*lucy said : an edge ##runner will take me to the moon .*
Label (Position-RPs)	0 0 0 0 0 **1** 0 0 **1** 0 0 0 0
Label (Category-RPs)	0 0 0 0 0 **2** 0 0 **1** 0 0 0 0
Label (Position-PIPs)	0 0 **1** 0 0 0 0 0 0 0 0 0 **1**
Label (Category-PIPs)	0 0 **2** 0 0 0 0 0 0 0 0 0 **3**

- An example of text pre-processing and label setting
  - Category 0: no pause (placeholder)
  - Category **1**: **brief** pause (< 300 ms)
  - Category **2**: **medium** pause (300 - 700 ms)
  - Category **3**: **long** pause (> 700 ms)

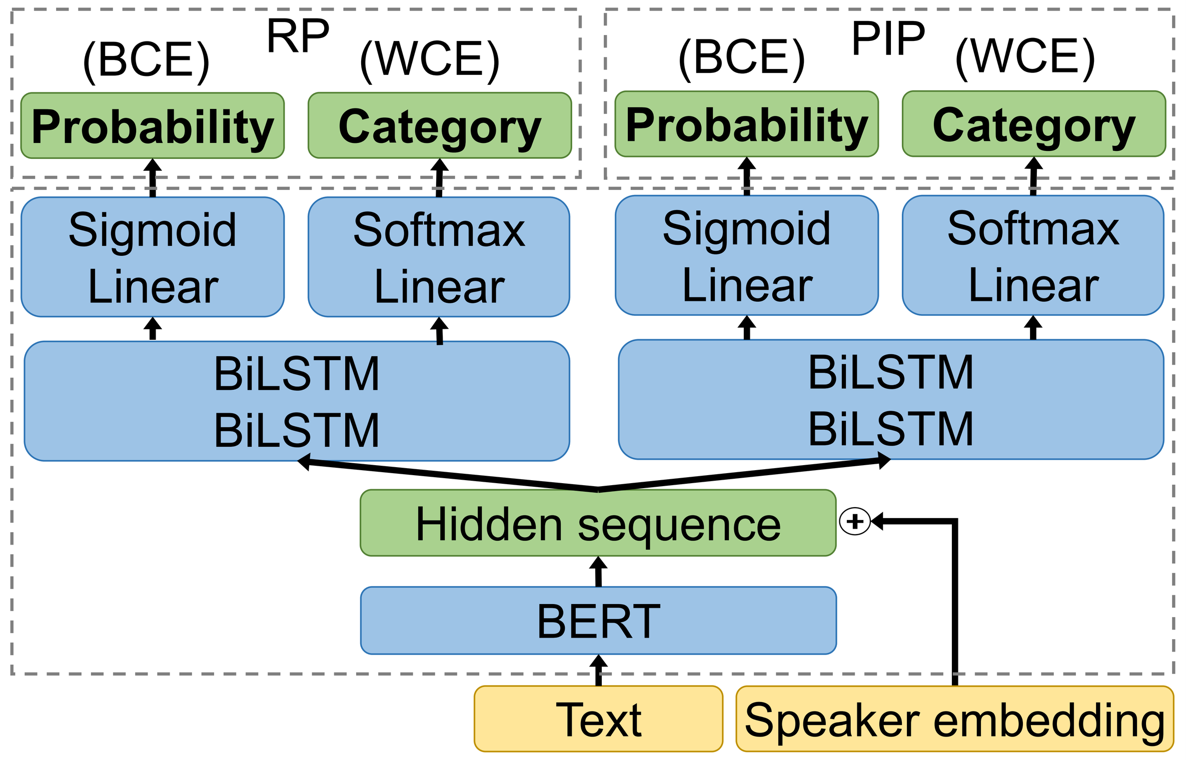
Proposed methods



- RPI model

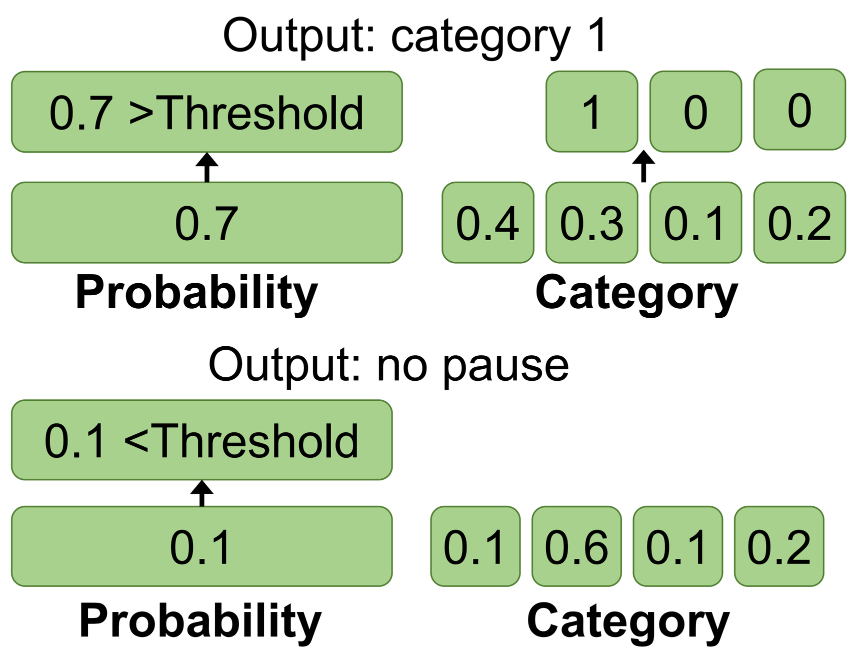
- Position prediction of RPs
- Encoder-decoder structure
- **Probability:** the predictive probability of the occurrence of a pause after each token.

BCE: Binary cross entropy loss; WCE: Weighted cross entropy loss



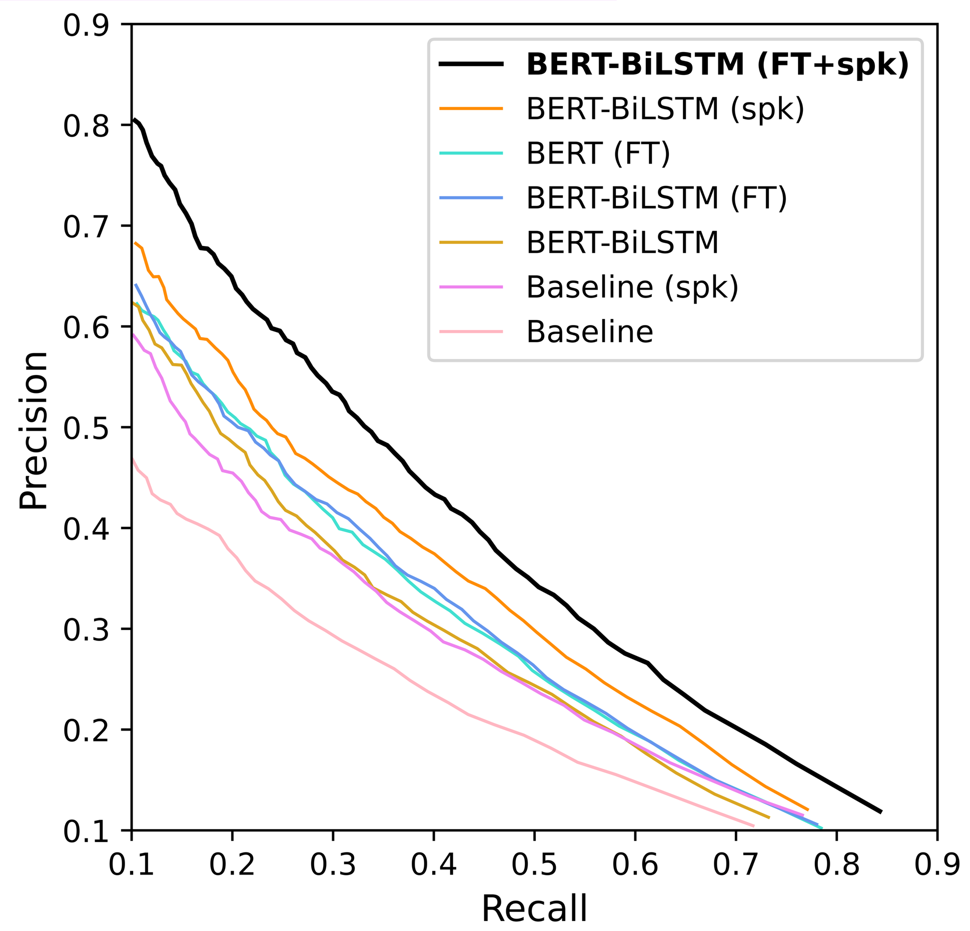
- CPI model

- Position and category prediction of both pauses
- Multi-task learning framework
- Two decoders: different distributions of the two pauses
- Category 0: placeholder
- The CPI model first predicts **Probability** that represents the occurrence of pauses and then outputs **Category** with the highest probability among categories 1–3.



- Inference of the CPI model

Objective evaluations (Predictive accuracy)



- Results of **RPI** model on RP position prediction

	Precision	Recall	$F_{0.5}$
BERT-BiLSTM (FT+spk)	**0.569**	**0.272**	**0.467**
BERT-BiLSTM (spk)	0.490	0.253	0.413
BERT (FT)	0.487	0.233	0.400
BERT-BiLSTM (FT)	0.467	0.246	0.396
BERT-BiLSTM	0.475	0.213	0.381
Baseline (spk)	0.446	0.209	0.364
Baseline [1]	0.393	0.187	0.322

- Different speakers have different styles for inserting RPs
- BERT fine-tuning + speaker embeddings
  - A large boost in predictive accuracy
- Using speaker embeddings in phrasing
  - Validity & generalizability

- Results of **CPI** model on position prediction

	Precision	Recall	F_β
RPs	0.575	0.261	$F_{0.5}$ = 0.463
PIPs	0.848	0.996	F_2 = 0.962

- Results of **CPI** model on 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**

- The CPI model retains the ability to predict the position of RPs
- The predictive accuracy of category 2 is lower than that of the others
  - There is still some works to be done in threshold choice of categorization

Subjective evaluations (Rhythm)

TTS model: FastSpeech 2
Vocoder: HiFi-GAN
The number of speakers: 16
Text-speaker pairs: 277
Test: AB preference test
The number of listeners: 30 per test
Platform: Amazon Mechanical Turk

Inserting normal pause marks at punctuation:
- FastSpeech2

Inserting predictive normal pause marks:
- Baseline, RPI, CPI (Position)

Inserting predictive categorized pause marks:
- CPI

Method A	Score	Method B
RPI	0.560 vs. 0.440	FastSpeech 2
RPI	0.537 vs. 0.463	Baseline
CPI	0.557 vs. 0.443	Baseline
RPI	0.448 vs. 0.512	CPI (Position)
RPI	0.460 vs. 0.540	CPI
RPI	0.561 vs. 0.490	RPI*
CPI	0.550 vs. 0.450	CPI*

- Subjective performance of the models
  - (Position): only predicting pause position
  - *: inputting with unmatched speaker embedding
  - Underlined scores: p-values below 0.05
- RPI ≈ Baseline, RPI ≈ RPI*
  RPI > FastSpeech2
  - Listeners are insensitive to the position of RPs
  - Listeners only become aware when listening to a long sentence without a pause
- CPI performed the best, RPI ≈ CPI(Position)
  - Inputting categorized pause phonemes to phoneme-based TTS models makes the rhythm of synthetic speech better
- CPI > CPI*
  - Listeners are sensitive to the difference arising from inserting different categories of pauses

References

[1] V. Klimkov et al., *Proc. INTERSPEECH*, 2017.
[2] K. Futamata et al., *Proc. INTERSPEECH*, 2021.
[3] A. Abbas et al., *Proc. INTERSPEECH*, 2022.
[4] E. Campione et al., *Proc. Speech Prosody*, 2002.