Frame-Wise Breath Detection with Self-Training: An Exploration of Enhancing Breath Naturalness in Text-to-Speech

Dong Yang[1,2], Tomoki Koriyama[1], Yuki Saito[2]

[1]CyberAgent, Inc., Japan; [2]The University of Tokyo, Japan

⇧ Code & speech samples

CyberAgent AI Lab

Overview

Hmm	I	think	that	...	[Inhale]	we	should	[click]	go
(Filled Pause)				(Silence)	(Breath)			(Tongue Click)	

Breath synthesis remains underexplored in Text-to-Speech (TTS) research.

1. Breath Detection (Our Focus)

- a. Develop a rule-based approach for initial labeling
  - I. Avoid extensive manual annotation of training data
- b. Propose and train our breath detection model
  - I. Frame-wise detection with reduced computational costs
  - II. Self-training method on a large TTS corpus

2. Speech Synthesis (For Validation)

- a. Insert breath marks to text transcripts based on detection results
- b. Train a TTS model
  - I. Achieve more natural breath-contained synthetic speech

Acoustic Features

Duration[1]

Zero-Crossing Rate (ZCR)[2]

Definition: Rate of the audio signal changes its sign

For discrete sampled signal:

N: window length, $X = \{x[n]\}_{n=0}^{N-1}$: audio signal

$$ZCR(X) = \frac{1}{N-1}\sum_{n=1}^{N-1} 0.5|\,sgn(x[n]) - sgn(x[n-1])|$$

Variance of Mel-Spectrogram (VMS)

Definition: Var(Mel) in frequency domain

Normalized Average of VMS (NA-VMS)

Definition: mean of min-max normalized VMS

F: frame, $V = \{v[f]\}_{f=0}^{F-1}$: audio signal

$$NA-VMS(V) = \frac{1}{F}\sum_{f=0}^{F-1}\frac{v[f] - \min(V)}{\max(V) - \min(V)}$$

Dataset & Annotation

LibriTTS-R[3] Corpus + MFA[4] for text-speech alignment & pause recognition

Manual annotation for valid & test sets:

	Utterances	Pauses	Annotated breath
Validation set	520	2,049	400
Test set	455	2,051	480

Rule-based annotation for training set:

Class	Duration	Max(VMS)	Max(ZCR)	NA-VMS	Precision	Recall
Breath	> 300 ms	> 150	> 1 x 10^{-4}	> 0.6	0.982	0.450
Non-breath	-	< 150	< 5 x 10^{-5}	-	1.000	0.111

Self-Training Process

$\textit{label}(T, P, B, U)$:

All frames in training set: T
- MFA-recognized pauses: P
  - Annotated breath: B → 1
  - Annotated non-breath: U → 0
  - Unannotated: $P \setminus (B \cup U)$ → -100
- Non-pauses: $T \setminus P$ → 0

Algorithm: Self-training for breath detection models

Initialize:

$$k \leftarrow 0$$
$$Y^0 \leftarrow label(T, P, B, U)$$
$$D_\theta^0 \leftarrow BCE(D_\theta(T), Y^0)$$

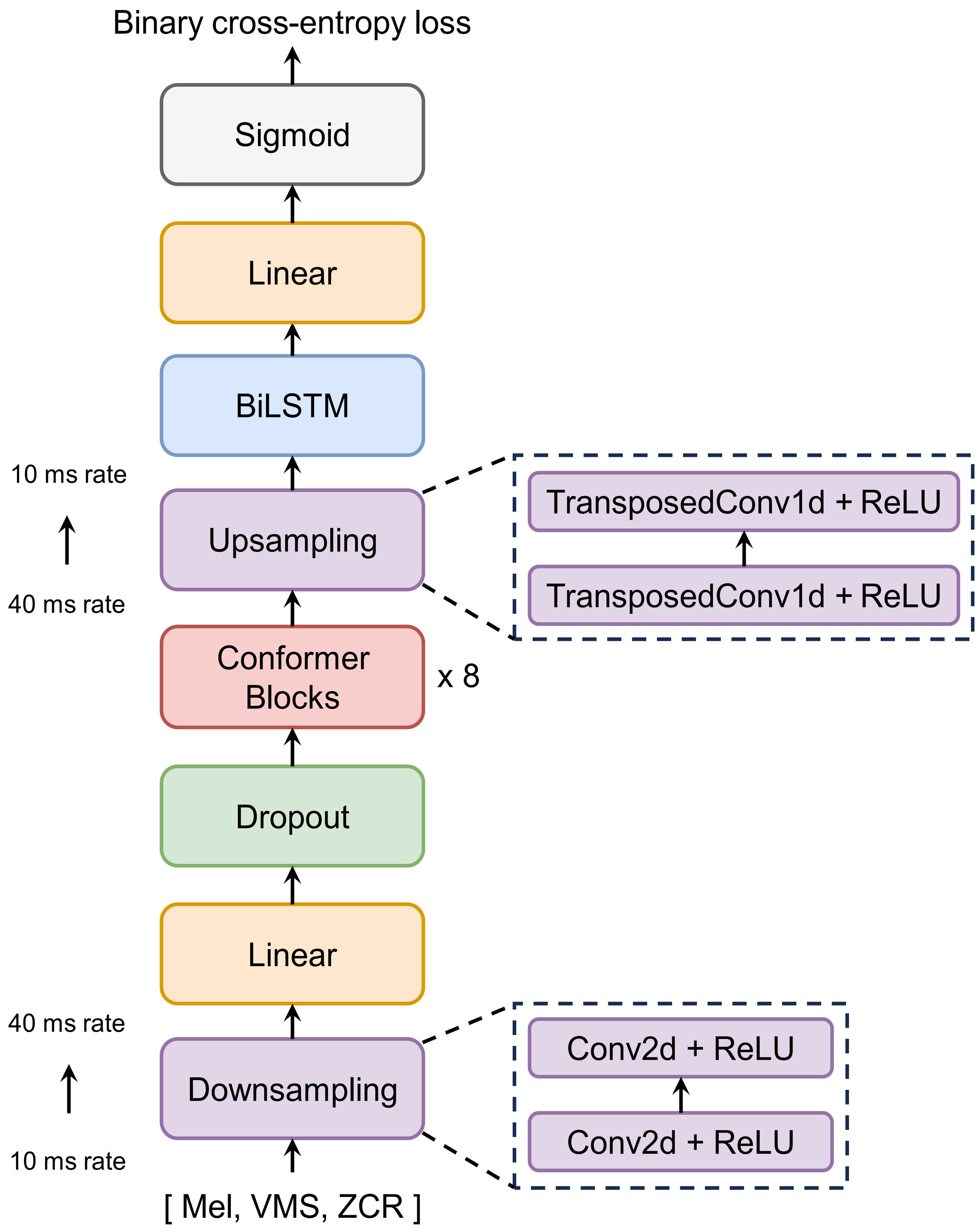
Repeat:

$$k \leftarrow k+1$$
$$\alpha^k \leftarrow argmin_{\alpha^k}|Precision(D_\theta^{k-1}(P_{valid}) > \alpha^k, B_{valid}) - (0.97 - 0.02 * k)|$$
$$\beta^k \leftarrow argmin_{\beta^k}|Precision(D_\theta^{k-1}(P_{valid}) < \beta^k, U_{valid}) - (0.97 - 0.02 * k)|$$
$$\hat{B} \leftarrow D_\theta^{k-1}(T) > \alpha^k$$
$$\hat{U} \leftarrow D_\theta^{k-1}(T) < \beta^k$$
$$Y^k \leftarrow label(T, P, B \cup \hat{B}, U \cup \hat{U}) \quad \triangleright \text{ Pseudo-labeling}$$
$$D_\theta^k \leftarrow BCE(D_\theta^{k-1}(T), Y^k)$$
$$\gamma^k \leftarrow argmax_{\gamma^k} IoU(D_\theta^k(T_{valid}) > \gamma^k, B_{valid})$$

Until: $IoU(D_\theta^k(T_{valid}) > \gamma^k, B_{valid}) < IoU(D_\theta^{k-1}(T_{valid}) > \gamma^{k-1}, B_{valid})$

Output: D_θ^{k-1}

Proposed Model



Experimental Results

Breath detection experiments:

Metric: intersection over union (IoU)

Iteration	Baseline[5]	Proposed
0	0.616	0.777
1	0.634	0.809
2	0.681	0.829
3	**0.710**	**0.836**
4	0.709	0.827

Training configurations:
- Optimizer: AdamW
- Scheduler: Linear
- Peak learning rate: 2 x 10^{-5}
- Epoch: 10
- Batch size: 64
- Dataset: train-clean-100 train-other-500

1. Our proposed model consistently outperformed the baseline model.
2. Both models achieved their peak IoU after the 3rd training iteration, where the models were considered the best-performing ones and used in the TTS experiments.

Ablation studies:

Model	Iteration	IoU
Proposed	0	0.777
w/o ZCR		0.631
w/o VMS		0.677
w/o non-breath		0.702
Proposed	1	0.809
w/o pseudo-label		0.740

1. ZCR and VMS in the input and the use of non-breath set proved to be critical.
2. Continued training without pseudo-labeling did not improve performance.

TTS experiments:

TTS model: VITS[6]; Dataset: train-clean-360

Model	MOS1 ± CI	MOS2 ± CI
Ground truth	4.03 ± 0.12	3.92 ± 0.13
VITS	3.35 ± 0.15	3.34 ± 0.17
VITS w/ baseline	3.27 ± 0.15	3.50 ± 0.14
VITS w/ proposed	**3.37 ± 0.14**	**3.55 ± 0.15**

MOS1:
- General evaluation
- Samples: Not all utterances included breath
- Conclusion: Inaccurate breath detection negatively affected the TTS training

MOS2:
- Breath-focus evaluation
- Samples: All utterances included breath
- Instruction: "Please focus on the breath sounds"
- Conclusion: Detected breath marks enhanced the naturalness of synthetic breath sounds

[1] N. Braunschweiler and L. Chen, SSW 2015.
[2] D. Ruinskiy and Y. Lavner, TASLP 2007.
[3] Y. Koizumi et al., INTERSPEECH 2023.
[4] M. McAuliffe et al., INTERSPEECH 2017.
[5] E. Sz´ekely et al., ICASSP 2019.
[6] J. Kim, J. Kong, and J. Son, ICML 2021.