

端到端语音合成中的韵律、情感建模研究

刘瑞 Rui Liu

Email: liurui_imu@163.com

4 Dec 2021

OUTLINE

- Introduction
- Expressive TTS Training with frame and style reconstruction Loss
- GraphSpeech
- StrengthNet

OUTLINE

- Introduction
- Expressive TTS Training with frame and style reconstruction Loss
- GraphSpeech
- StrengthNet

Introduction

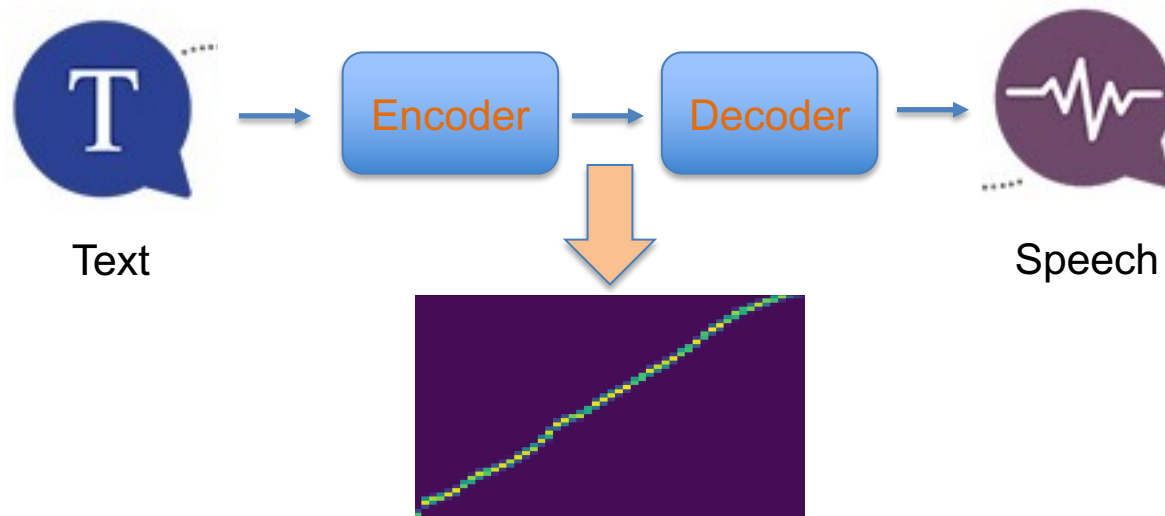
- Text-to-speech (TTS) seeks to synthesize human-like natural sounding voice for input text.



Taylor, Paul. *Text-to-speech synthesis*. Cambridge university press, 2009.

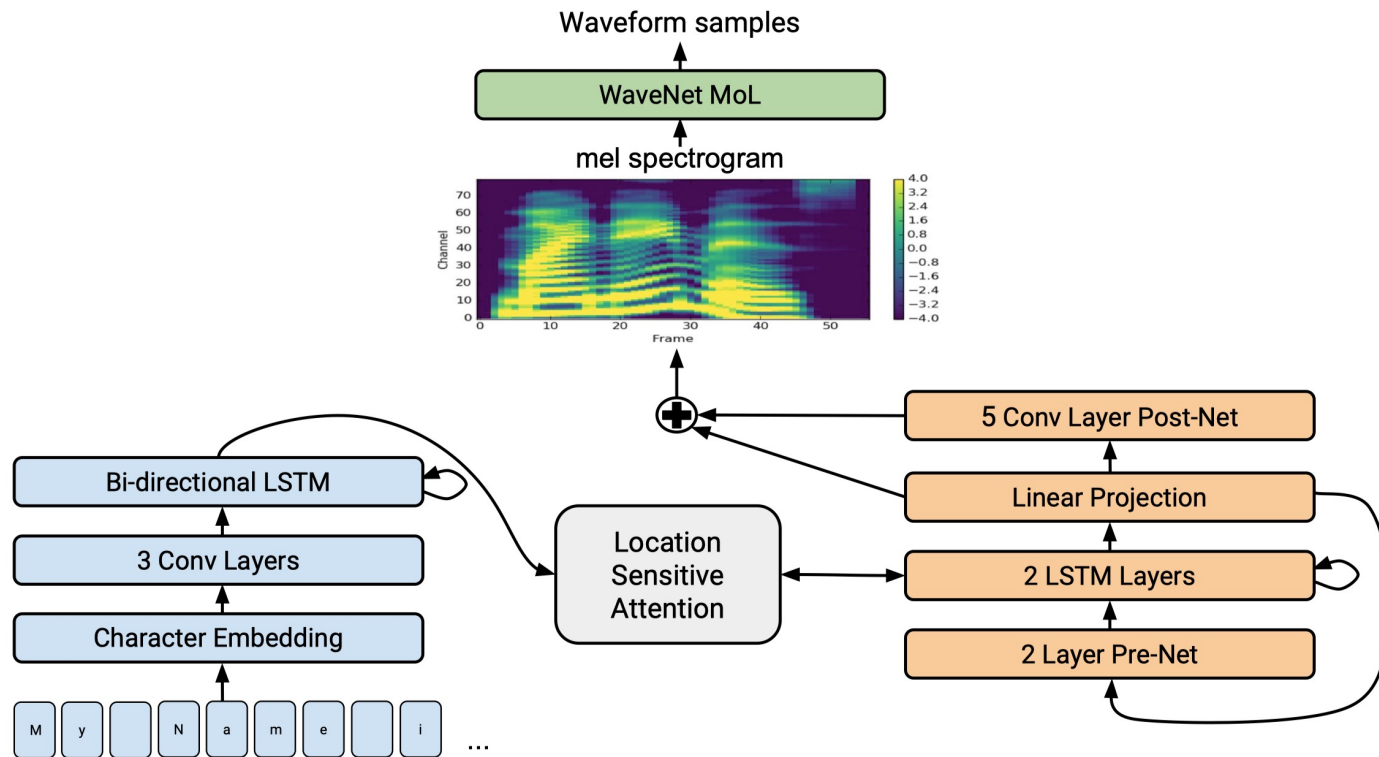
Introduction

- End-to-end TTS models simplify the synthesis pipeline with an **encoder decoder** network.



Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *3rd International Conference on Learning Representations, ICLR 2015*. 2015.

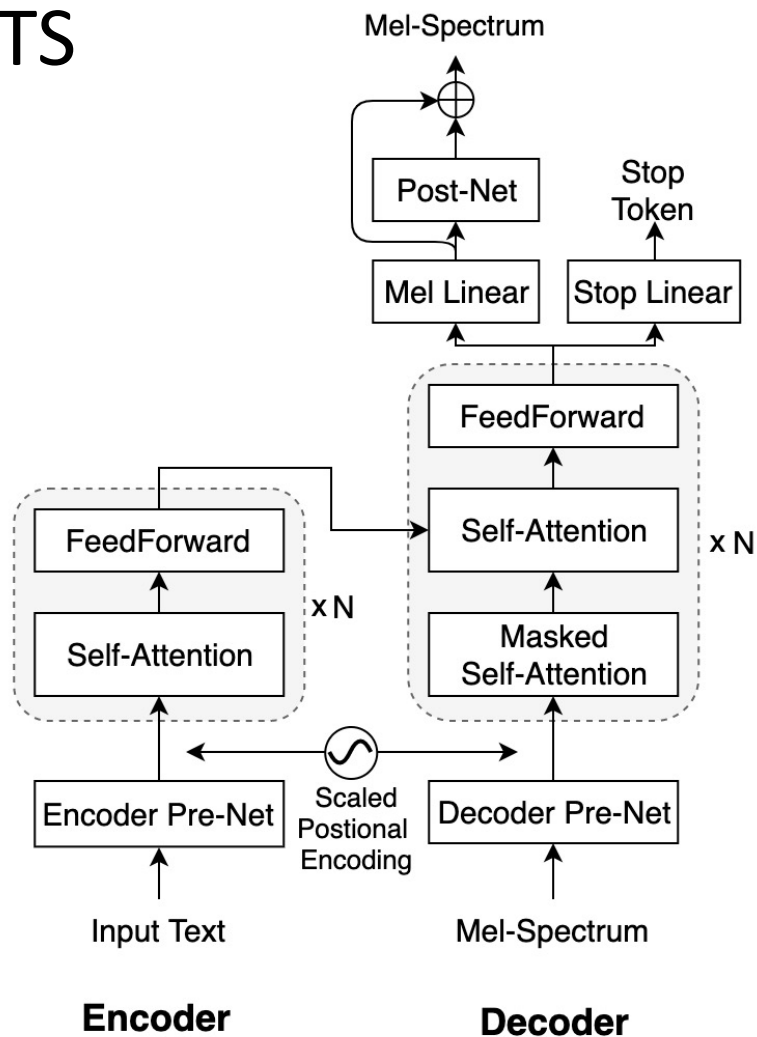
Tacotron TTS



Yuxuan Wang, et al. "Tacotron: Towards End-to-End Speech Synthesis." INTERSPEECH 2017

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." ICASSP 2018.

Transformer-TTS

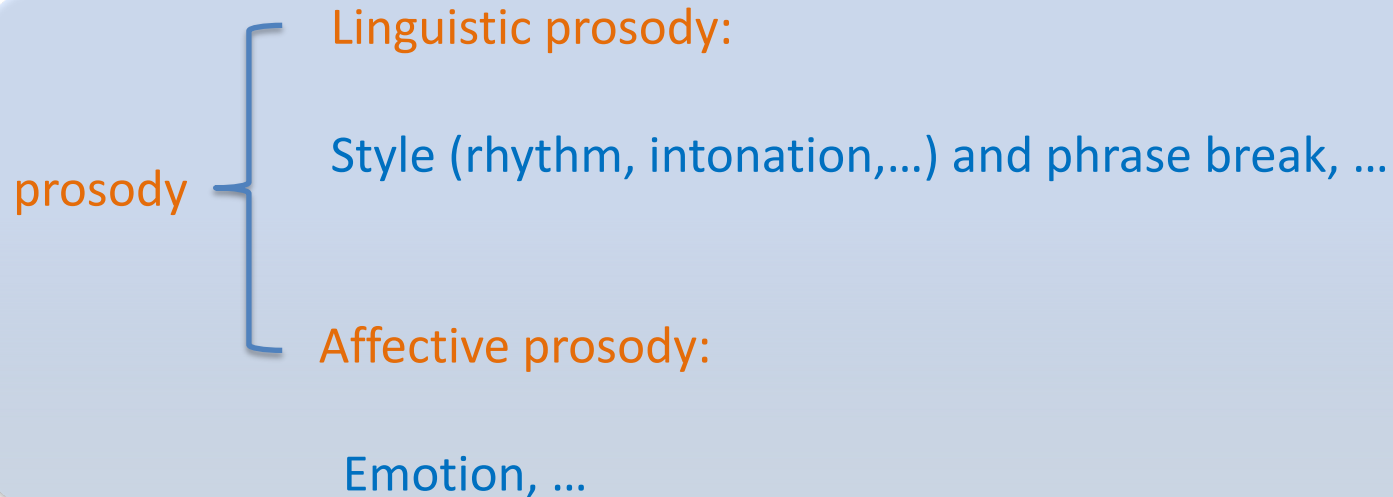


Naihan Li, et al., "Neural speech synthesis with transformer network," AAAI-2019

端到端语音合成中的韵律、情感建模研究

Introduction

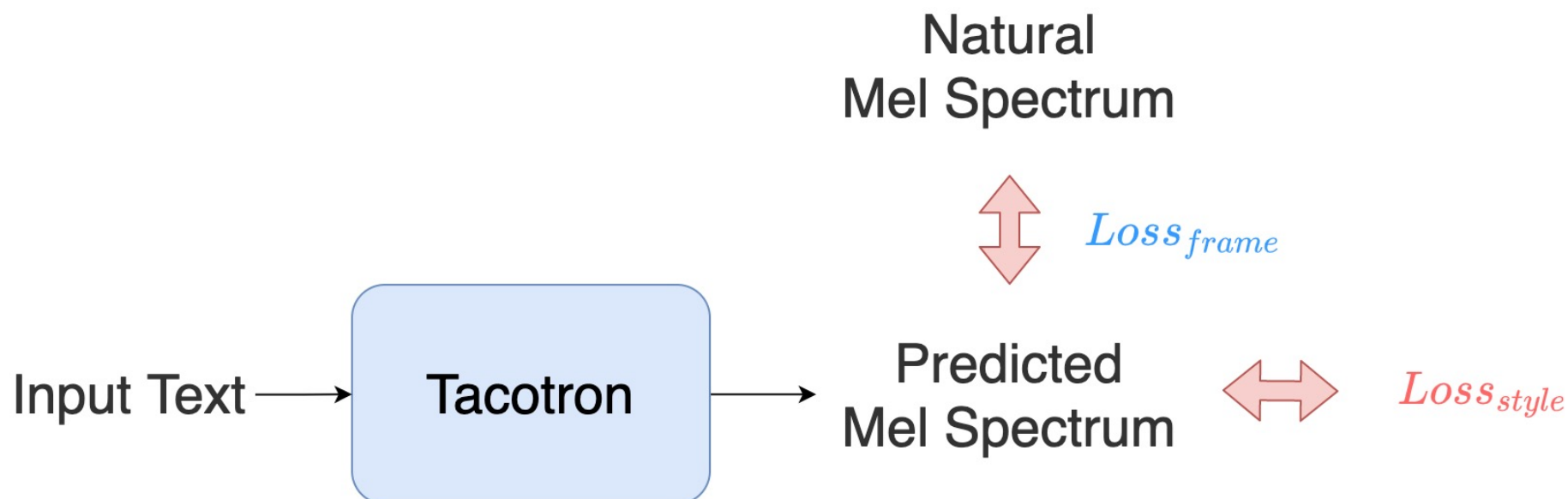
- However, it doesn't model the complex **prosody/style, emotion** information within a sentence.



Xu, Yi. "Speech prosody: A methodological review." *Journal of Speech Sciences* 1.1 (2011): 85-115.

OUTLINE

- Introduction
- Expressive TTS Training with frame and style reconstruction Loss
- GraphSpeech
- StrengthNet



Expressive TTS Training with frame and style reconstruction Loss

R. Liu, B. Sisman, G. Gao and H. Li, "Expressive TTS Training With Frame and Style Reconstruction Loss," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1806-1818, 2021, doi: 10.1109/TASLP.2021.3076369.

Tacotron-PL (Perceptual Loss)

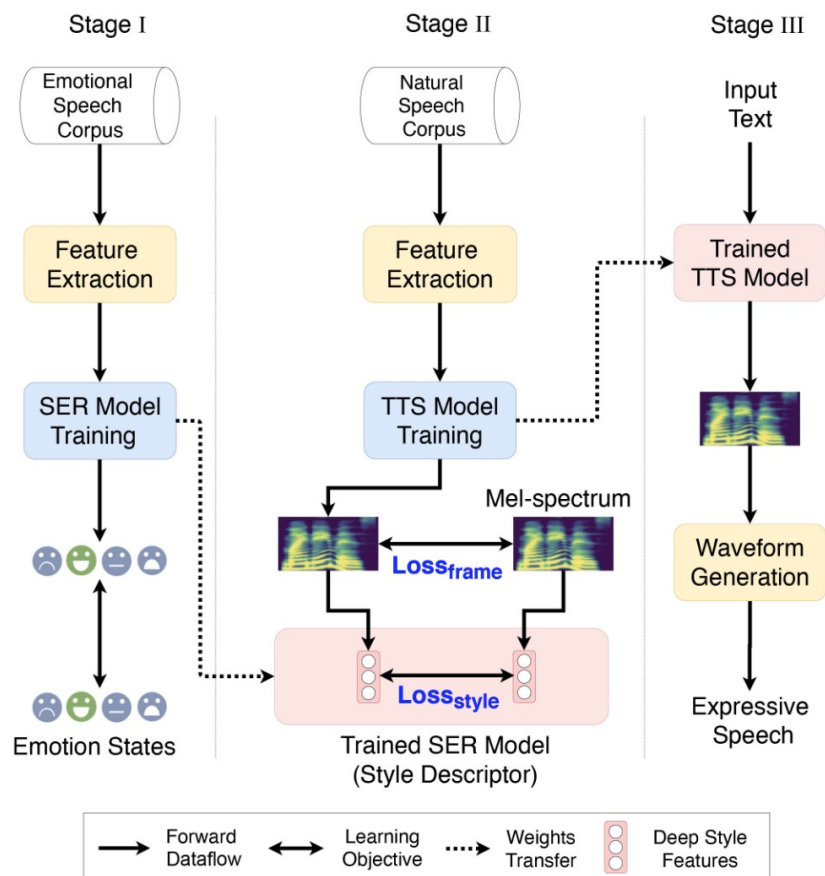


Fig. 2. Overall framework of a Tacotron-PL system in three stages: Stage I for training of style descriptor; Stage II for training of Tacotron-PL; Stage III for run-time inference.

Tacotron-PL

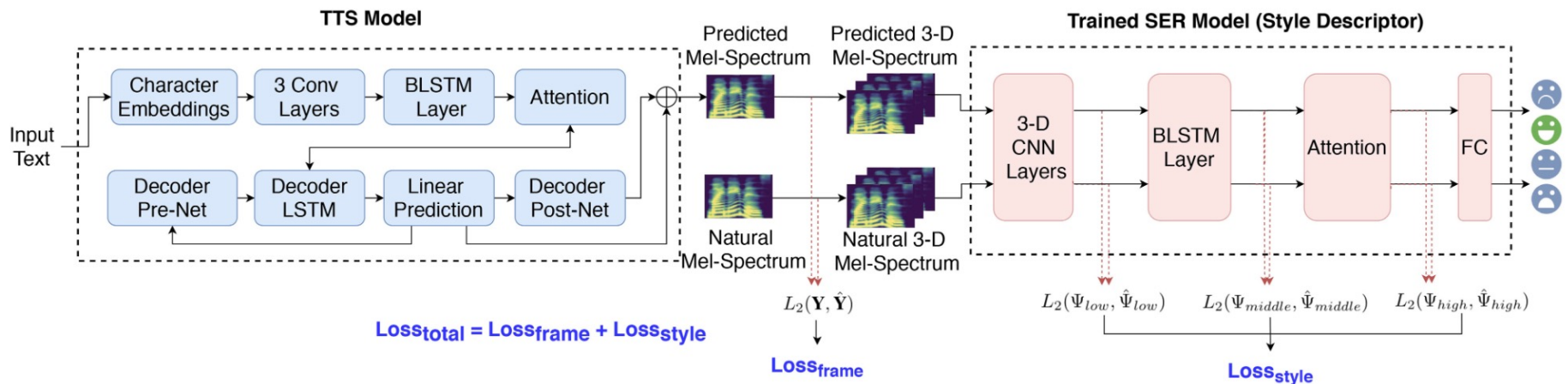


Fig. 3. Block diagram of the proposed training strategy, *Tacotron-PL*. A speech emotion recognition (SER) model is trained separately to serve as an auxiliary model to extract deep style features. A *style reconstruction loss*, $\text{Loss}_{\text{style}}$, is computed between the deep style features of the generated and reference speech at utterance-level.

Chen, Mingyi, et al. "3-D convolutional recurrent neural networks with attention model for speech emotion recognition." *IEEE Signal Processing Letters* 25.10 (2018): 1440-1444.

Experiments

Dataset	Description
IEMOCAP	10039 utterances happy, angry, sad, and neutral
LJ-Speech	13100 utterances ~24 hours

Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42.4 (2008): 335-359.

K. Ito, "The lj speech dataset," <https://keithito.com/LJ-SpeechDataset/>, 2017.

Experiments

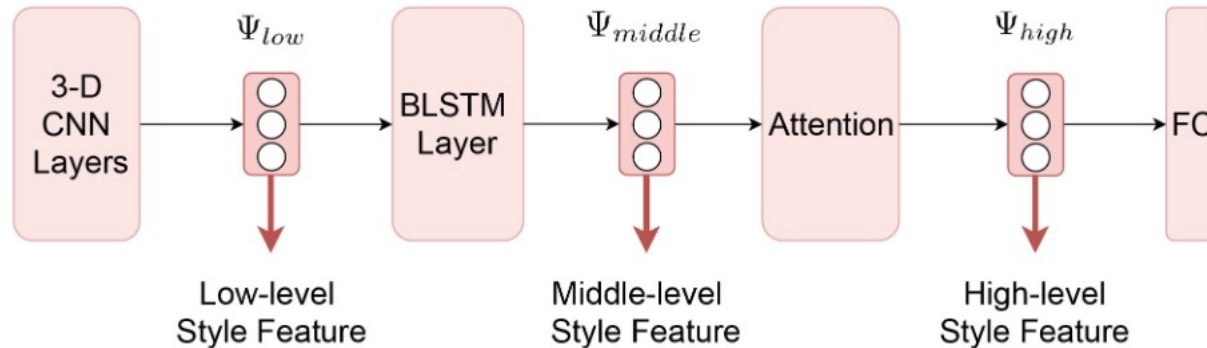
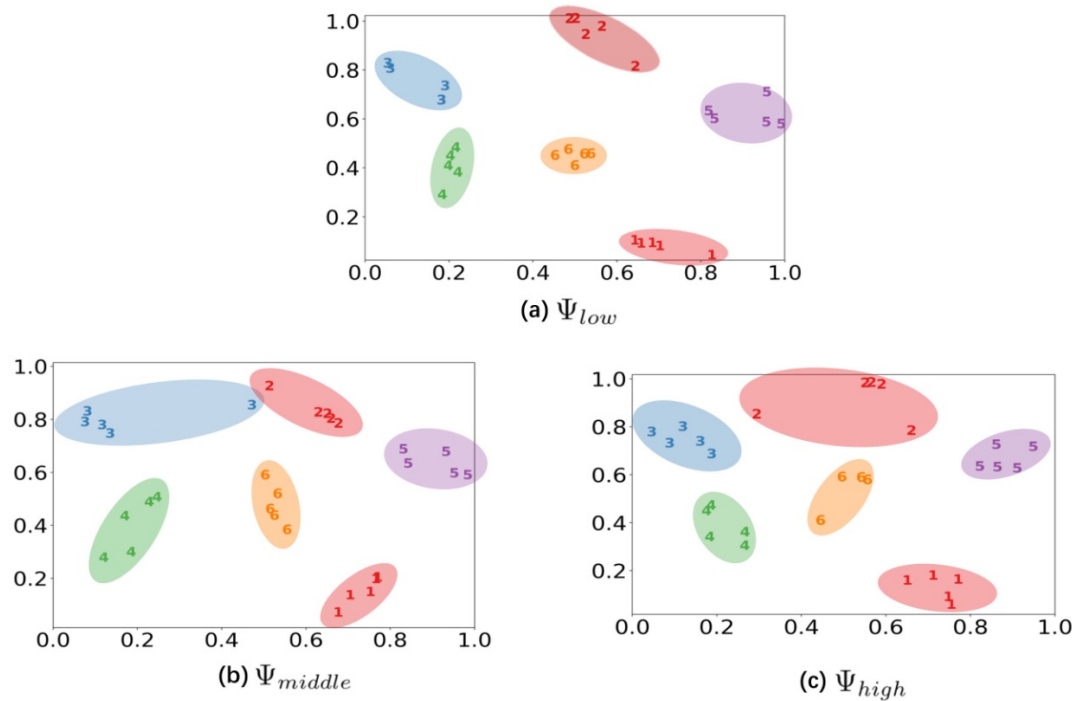


Fig. 5. Three level (low, middle and high) of deep style features extracted from SER-based style descriptors for computing style construction loss.

Baseline: Tacotron

Ours: Tacotron-PL(L); Tacotron-PL(M); Tacotron-PL(H); Tacotron-PL(LMH)

Experiments



- 1) Short question;
- 2) Long question;
- 3) Short answer;
- 4) Short statement;
- 5) Long statement
- 6) Digit string

Fig. 4. t-SNE plot of the distributions of deep style features Ψ_{low} , Ψ_{middle} and Ψ_{high} for six groups of utterances in LJ-Speech corpus. The list of utterances can be found at Table V in Appendix A.

Experiments

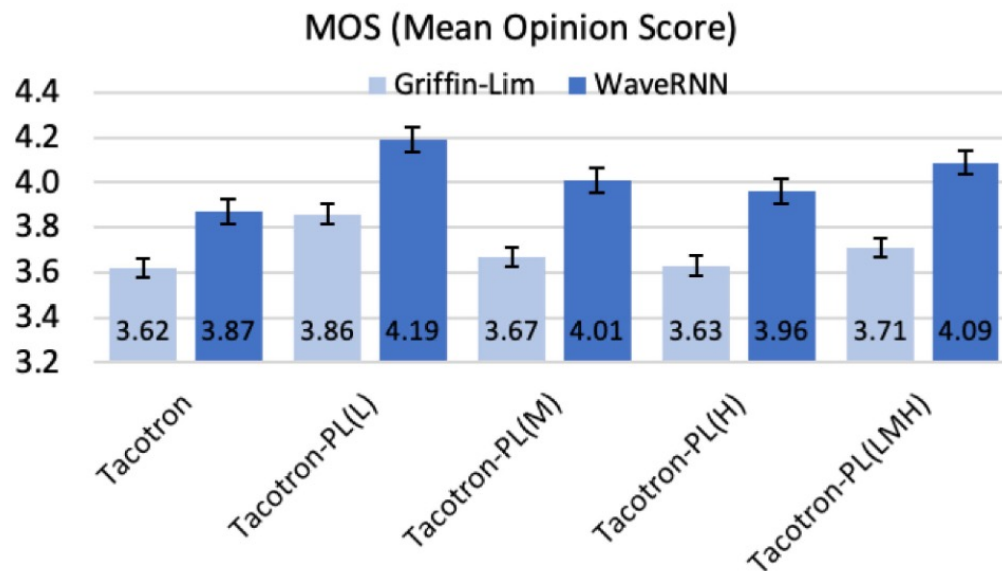
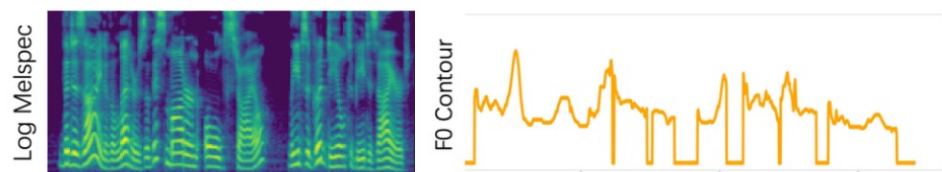
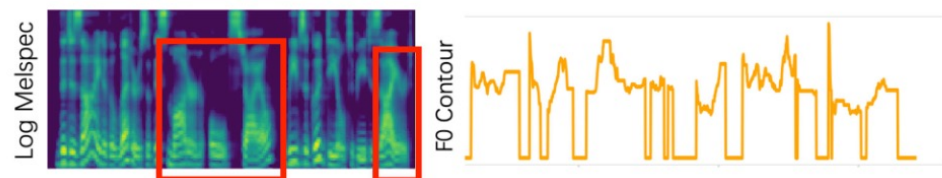


Fig. 7. The mean opinion scores (MOS) of five systems evaluated by 15 listeners, with 95% confidence intervals computed from the t-test.

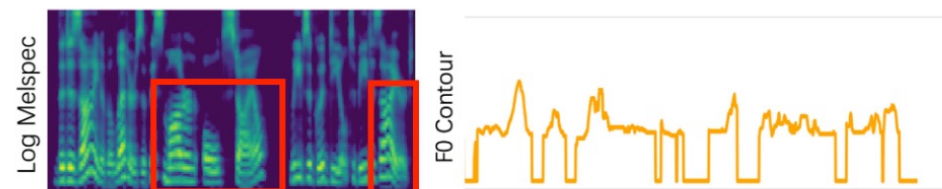
Experiments



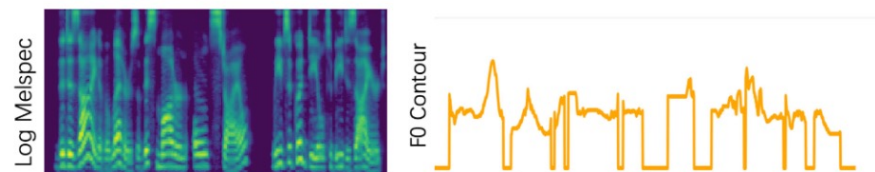
(a) Ground Truth



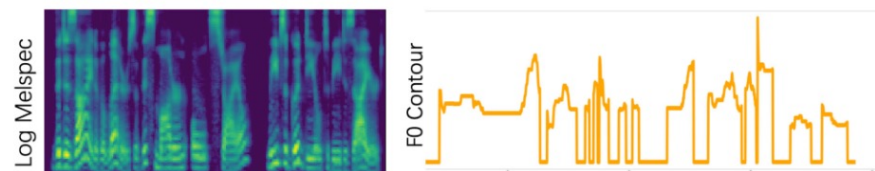
(b) Tacotron



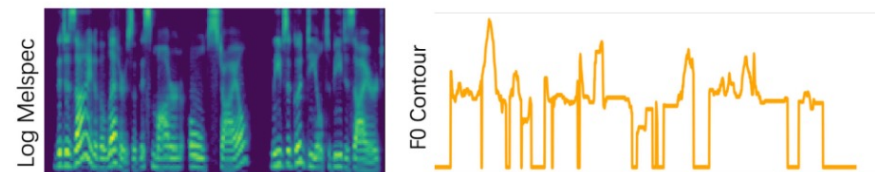
(c) Tacotron-PL(L)



(d) Tacotron-PL(M)














































































(e) Tacotron-PL(H)



(f) Tacotron-PL(LMH)

Speech Samples:

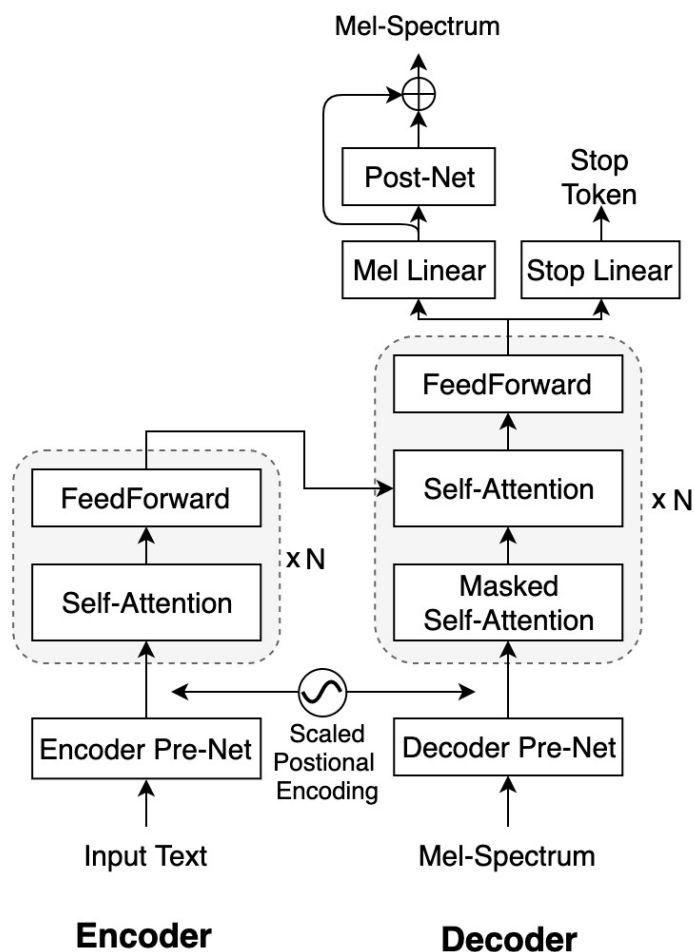
	Tacotron	Tacotron-PL(L)	Tacotron-PL(M)	Tacotron-PL(H)	Tacotron-PL(LMH)
[1]	 0:00  	 0:00  	 0:00  	 0:00  	 0:00  
[2]	 0:00  	 0:00  	 0:00  	 0:00  	 0:00  
[3]	 0:00  	 0:00  	 0:00  	 0:00  	 0:00  
[4]	 0:00  	 0:00  	 0:00  	 0:00  	 0:00  
[5]	 0:00  	 0:00  	 0:00  	 0:00  	 0:00  

Demo : <https://ttslr.github.io/Expressive-TTS-Training-with-Frame-and-Style-Reconstruction-Loss/>

端到端语音合成中的韵律、情感建模研究

OUTLINE

- Introduction
- Expressive TTS Training with frame and style reconstruction Loss
- **GraphSpeech**
- StrengthNet



Transformer-TTS

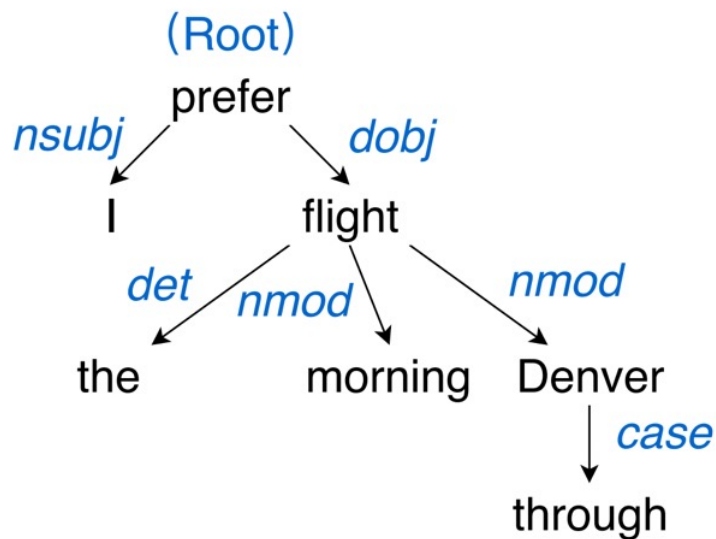
$$s_{ij} = f(x_i, x_j) = x_i W_q^T W_k x_j$$

$$x_i \xleftrightarrow{r_{ij}} x_j \quad ???$$

Vaswani, Ashish, et al. "Attention is all you need." *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017.

Text: “I prefer the morning flight through Denver.”
(我更喜欢上午飞丹佛的班机)

Syntax Tree



Guo Haohan, et al. "Exploiting Syntactic Features in a Parsed Tree to Improve End-to-End TTS." INTERSPEECH 2019: 4460-4464

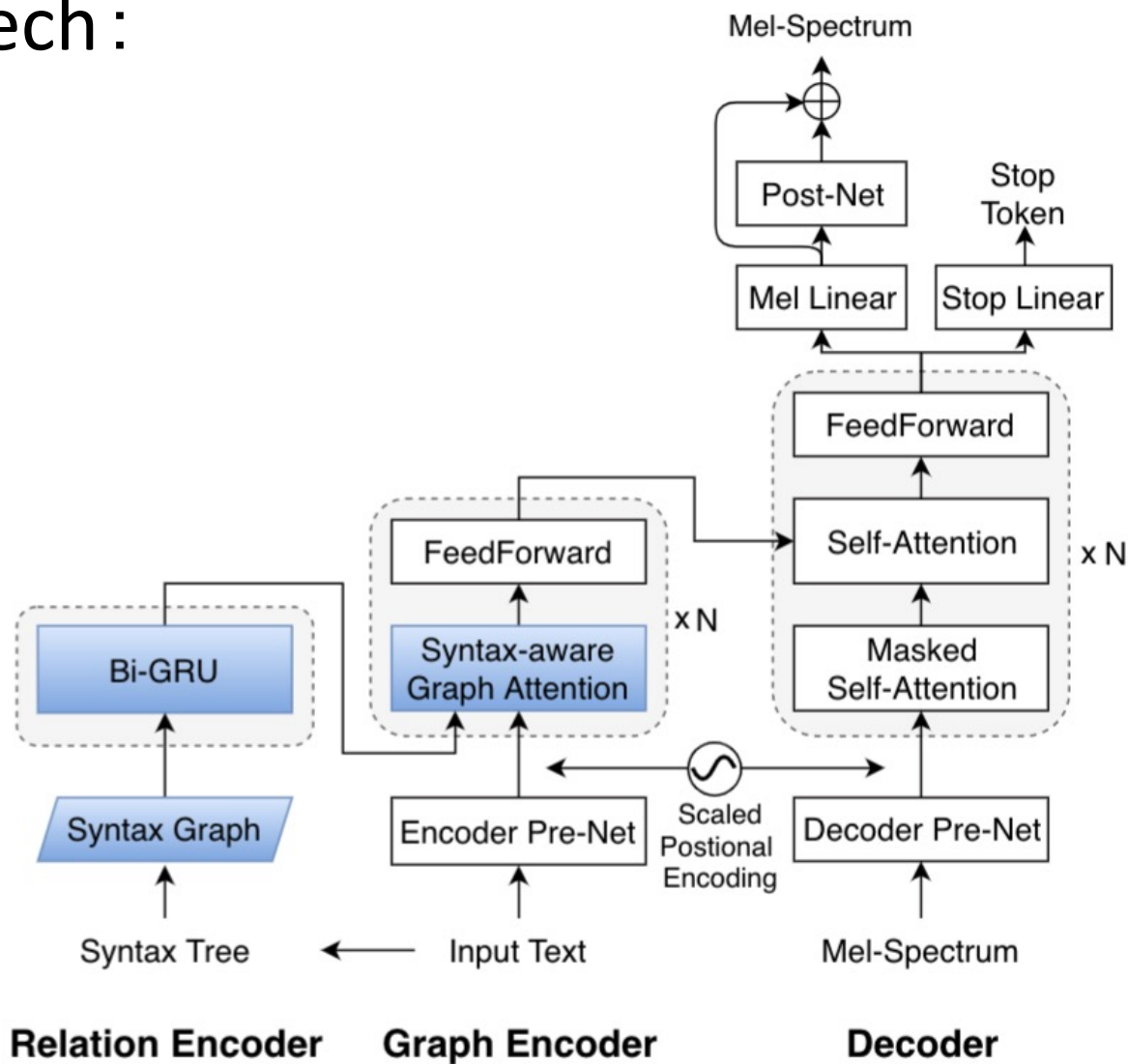
Song, Changhe, et al. "Syntactic representation learning for neural network based TTS with syntactic parse tree traversal." ICASSP 2021

GraphSpeech : Syntax-aware Graph Attention Network for Neural Speech Synthesis

Liu, Rui, Berrak Sisman, and Haizhou Li. "GraphSpeech: Syntax-aware graph attention network for neural speech synthesis." *ICASSP 2021*

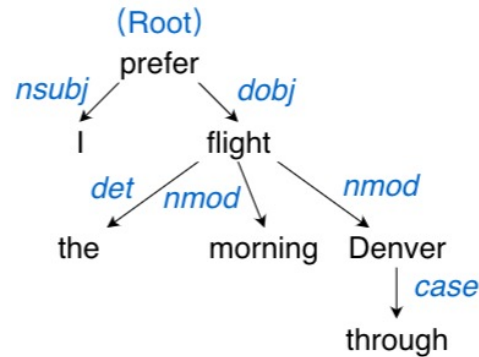
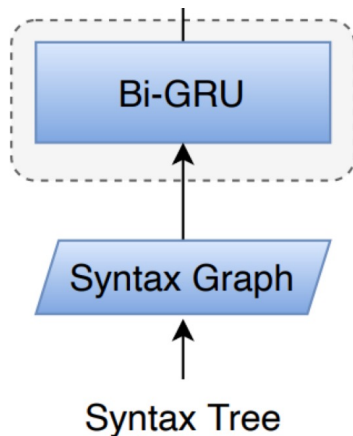
端到端语音合成中的韵律、情感建模研究

GraphSpeech :

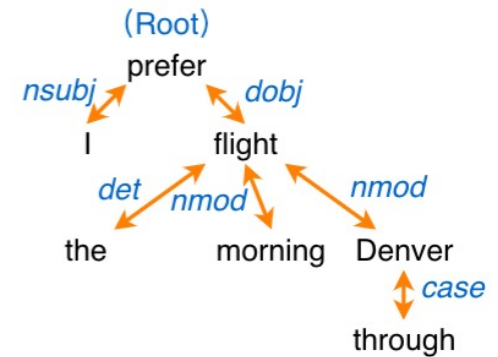


GraphSpeech

Relation Encoder:



(a) Syntax Tree



(b) Syntax Graph

$$[sp_1, \dots, sp_t, \dots, sp_{n+1}] = [e(i, k_1), e(k_1, k_2), \dots, e(k_n, j)]$$

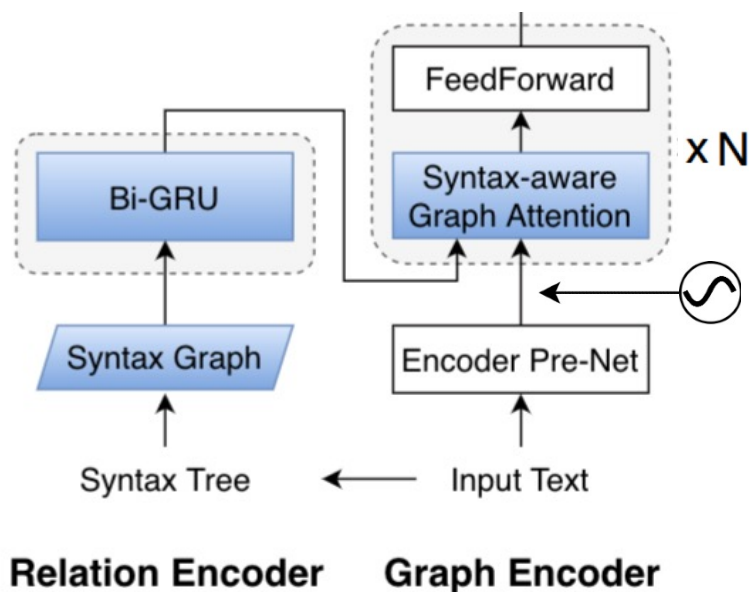
$$\begin{aligned} \vec{s}_t &= \text{GRU}_f(\vec{s}_{t-1}, sp_t) \\ \overleftarrow{s}_t &= \text{GRU}_b(\overleftarrow{s}_{t+1}, sp_t) \end{aligned}$$

$$r_{ij} = [\vec{s}_{n+1}; \overleftarrow{s}_0]$$

Scarselli, Franco, et al. "The graph neural network model." *IEEE transactions on neural networks* 20.1 (2008): 61-80.

GraphSpeech

Graph Encoder:

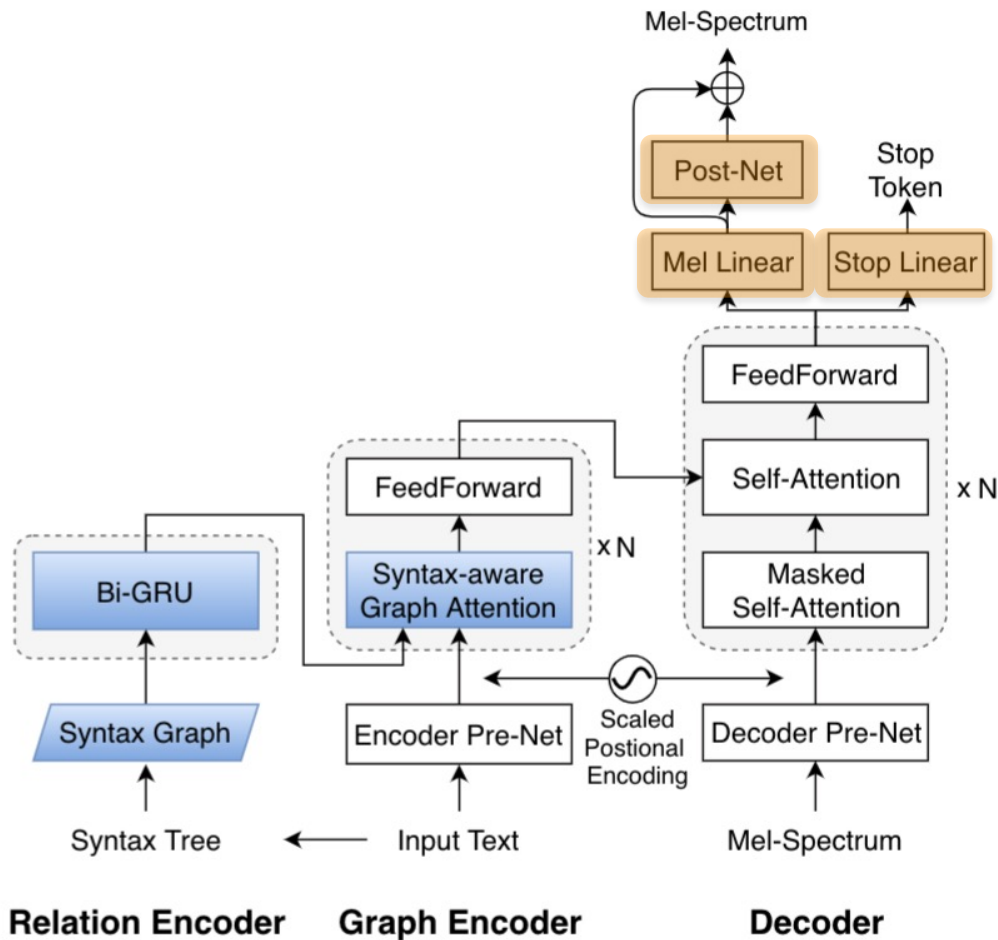


$$[r_{i \rightarrow j}; r_{j \rightarrow i}] = W_r r_{ij}$$

Syntax-aware Graph Attention:

$$\begin{aligned}
 s_{ij} &= g(x_i, x_j, r_{ij}) \\
 &= (x_i + r_{i \rightarrow j}) W_q^T W_k (x_j + r_{j \rightarrow i}) \\
 &= \underbrace{x_i W_q^T W_k x_j}_{(a)} + \underbrace{x_i W_q^T W_k r_{j \rightarrow i}}_{(b)} \\
 &\quad + \underbrace{r_{i \rightarrow j} W_q^T W_k x_j}_{(c)} + \underbrace{r_{i \rightarrow j} W_q^T W_k r_{j \rightarrow i}}_{(d)}
 \end{aligned}$$

GraphSpeech



Decoder :

Mel Linear

Post-Net and

Stop Linear

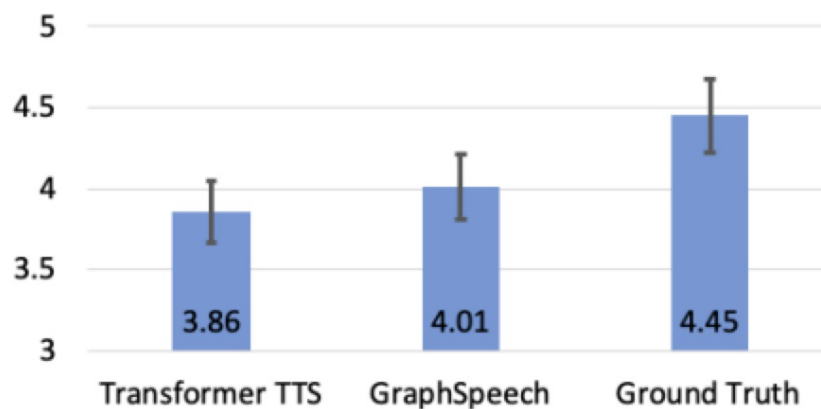
Experiments

- Database: [LJ-Speech](#)
- Syntactic dependency parsing: [Stanza](#)
- Vocoder: [Griffin-Lim algorithm](#)
- Edge embedding: 200-D
- GRU size: 200
- Relation Encoding: 200-D
- Character Embedding: 256-D
- Mel-spectrum: 80
- N (block numbers): 6
- Attention head: 4

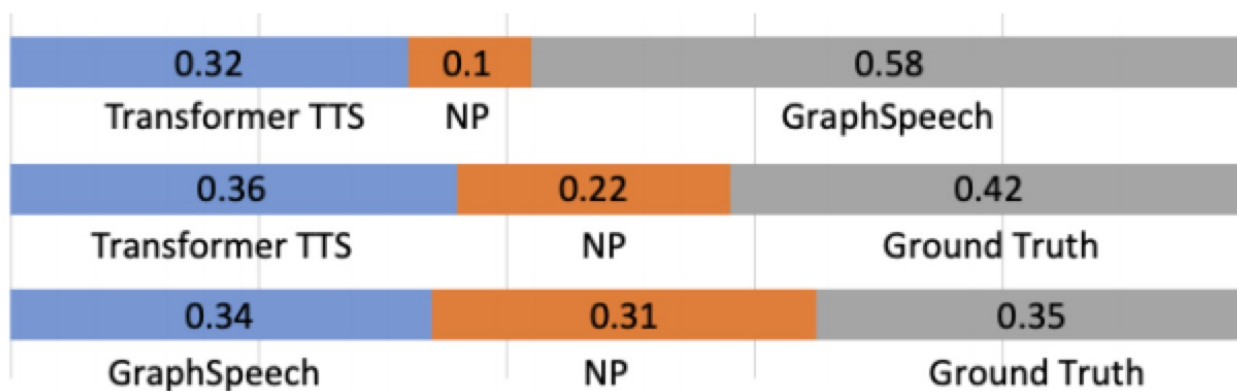
Peng Qi et al., “Stanza: A Python natural language processing toolkit for many human languages,” ACL-2020, pp.101-108.

Experiments

MOS Evaluation



AB Preference Test



Speech Samples:

Transformer TTS

GraphSpeech

[1] "We conduct objective evaluations to assess the performance of our proposed framework ."

▶ 0:00 / 0:06 — 🔊 ⋮

▶ 0:00 / 0:06 — 🔊 ⋮

[2] "The main contributions of this paper are listed as follows ."

▶ 0:00 / 0:04 — 🔊 ⋮

▶ 0:00 / 0:04 — 🔊 ⋮

[3] "Graphical structure plays an important role in natural language processing ."

▶ 0:00 / 0:04 — 🔊 ⋮

▶ 0:00 / 0:04 — 🔊 ⋮

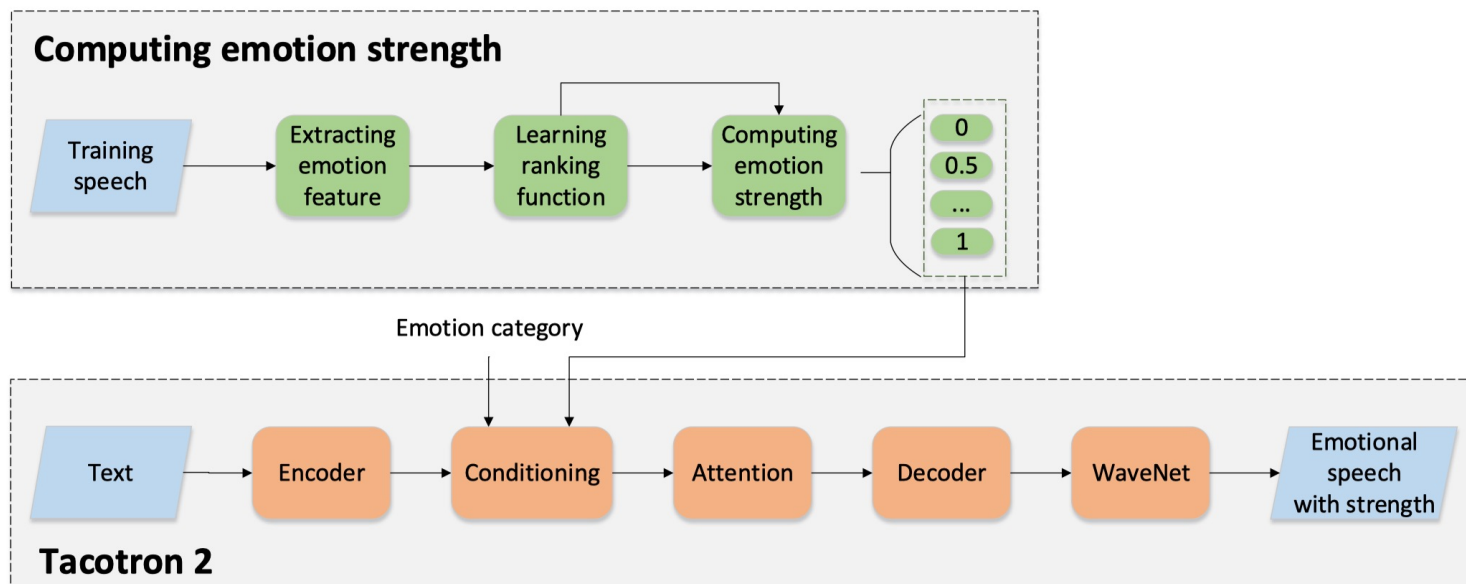
Demo: <https://ttslr.github.io/GraphSpeech>

端到端语音合成中的韵律、情感建模研究

OUTLINE

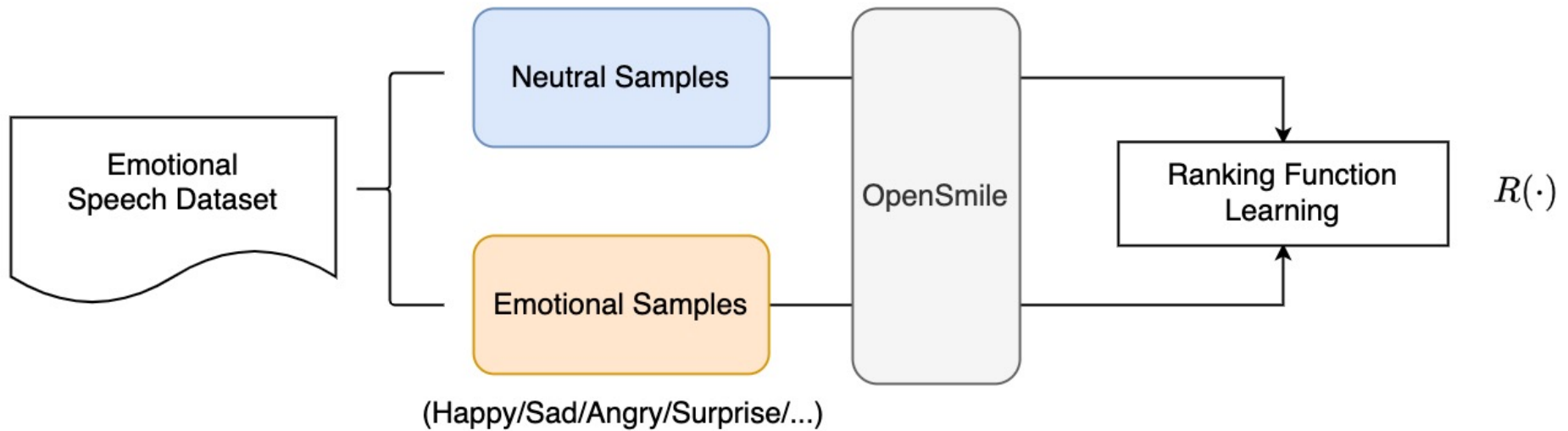
- Introduction
- Expressive TTS Training with frame and style reconstruction Loss
- GraphSpeech
- **StrengthNet**

Emotional Speech Synthesis with Strength Control



Zhu, Xiaolian, et al. "Controlling emotion strength with relative attribute for end-to-end speech synthesis." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

Lei, Yi, Shan Yang, and Lei Xie. "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis." *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

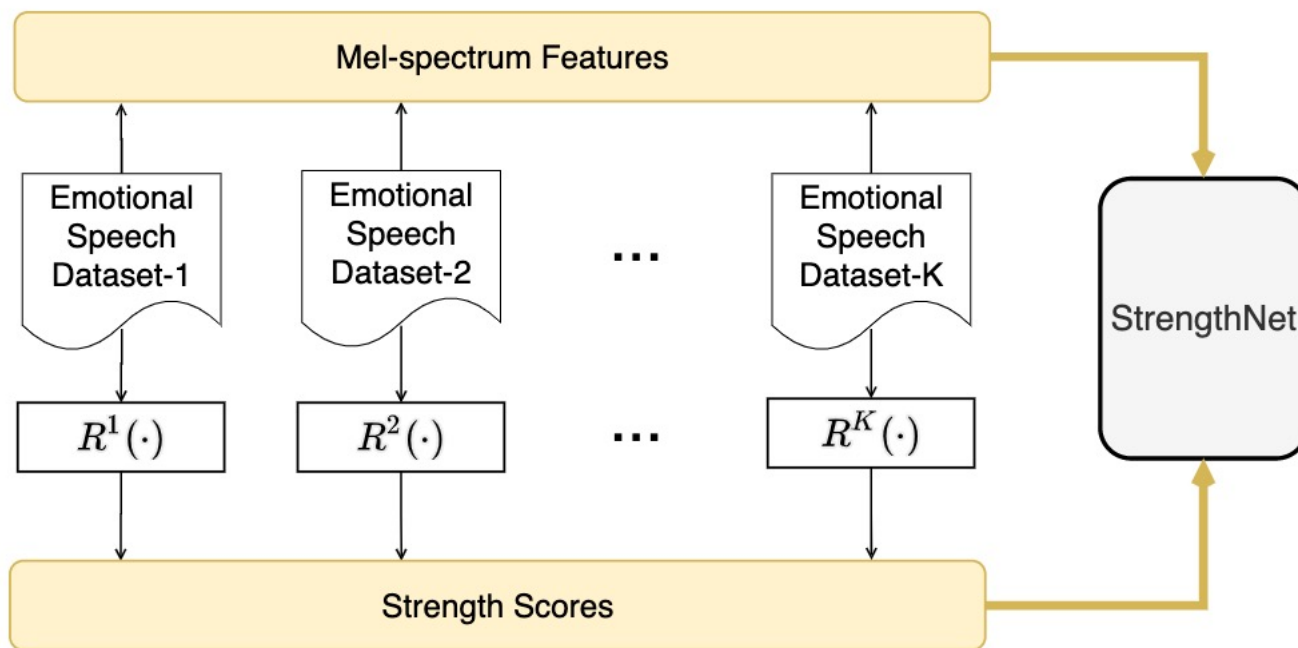


poor generalization

StrengthNet: Deep Learning-based Emotion Strength Assessment for Emotional Speech Synthesis

Liu, Rui, Berrak Sisman, and Haizhou Li. "StrengthNet: Deep Learning-based Emotion Strength Assessment for Emotional Speech Synthesis." *arXiv preprint arXiv:2110.03156* (2021). **Submitted to ICASSP 2022**

端到端语音合成中的韵律、情感建模研究



端到端语音合成中的韵律、情感建模研究

StrengthNet

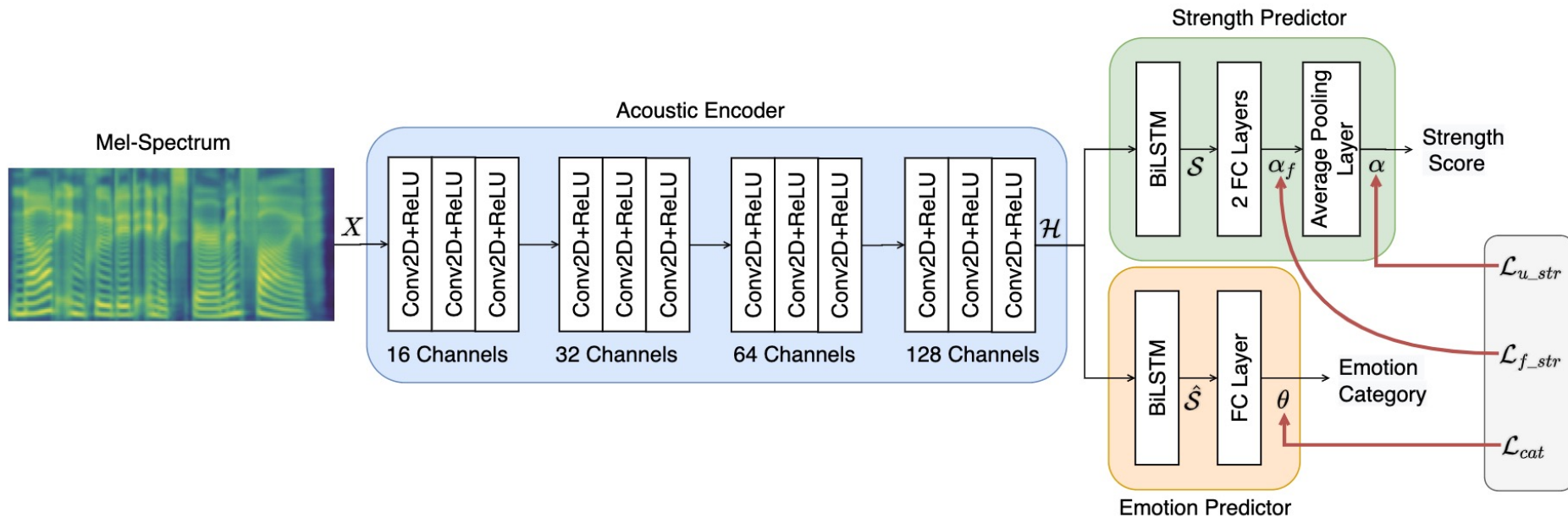


Fig. 1. The proposed StrengthNet that consists of Acoustic Encoder, Strength Predictor and Emotion Predictor.

Code: <https://github.com/ttslr/StrengthNet>

Experiments

Dataset	Description
ESD	10039 utterances (happy, angry, sad, and neutral)
RAVDESS	1440 utterances acted by 24 professional actors (calm, happy, sad, angry, fear, surprise, disgust, and neutral)
SAVEE	480 utterances (surprise, happy, sad, angry, fear, disgust, and neutral)

Zhou, Kun, Berrak Sisman, Rui Liu, and Haizhou Li.. "Emotional Voice Conversion: Theory, Databases and ESD." *arXiv preprint arXiv:2105.14762* (2021). accepted by **Speech Communication**.

Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." *PloS one* 13.5 (2018): e0196391.

Jackson, Philip, and SJUoSG Haq. "Surrey audio-visual expressed emotion (savee) database." *University of Surrey: Guildford, UK* (2014).

端到端语音合成中的韵律、情感建模研究

Experiments

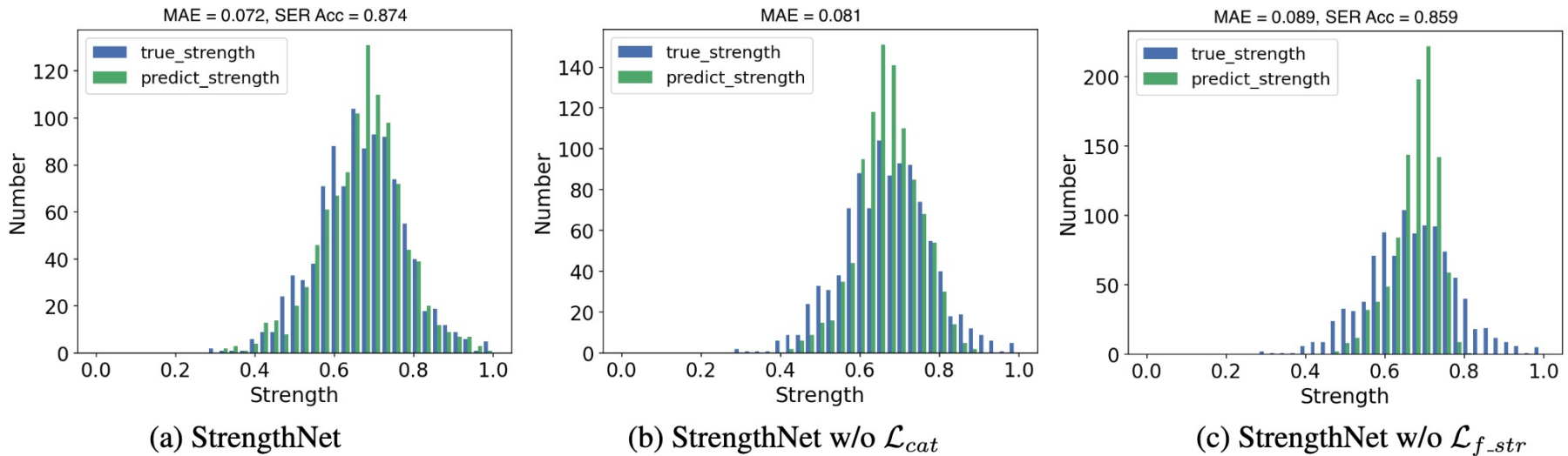


Fig. 3. Histogram of the utterance level strength predictions for (a) StrengthNet; (b) StrengthNet w/o \mathcal{L}_{cat} ; (c) StrengthNet w/o \mathcal{L}_{f_str} . The X-axis and Y-axis of subfigures represent the strength scores and the utterance number, respectively.

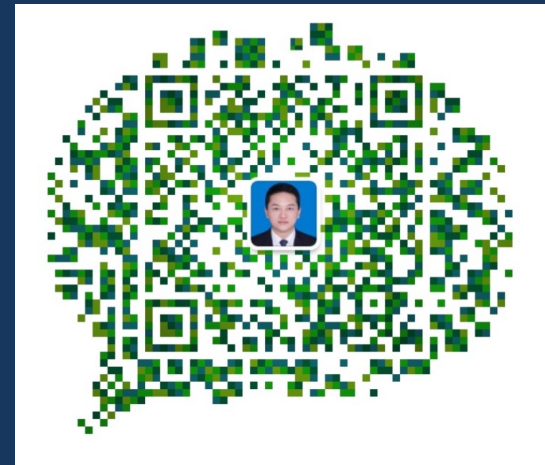
Experiments

Table 1. The MAE results for RAVDESS and SAVEE datasets in the comparison.

Method	MAE	
	RAVDESS	SAVEE
$R_{\text{RAVDESS}}(\cdot)$	NA	0.304
$R_{\text{SAVEE}}(\cdot)$	0.283	NA
$R_{\text{ESD}}(\cdot)$	0.266	0.272
StrengthNet _{ESD}	0.238	0.243
StrengthNet _{ESD+RAVDESS}	NA	0.173
StrengthNet _{ESD+SAVEE}	0.102	NA



Homepage: <https://ttslr.github.io/>



WeChat

Thank You!