Master's Thesis Presentation for the 143th Dissertation Evaluation

A Study on Mel-Spectrogram Synthesis for Multi-Speaker Text-To-Speech Task based on High-Low Frequency Separation

고주파-저주파 분리 기법을 활용한 다화자 음성 합성의 멜 스펙트로그램 생성 연구

Dahyun Song

hyeons.opy@gmail.com

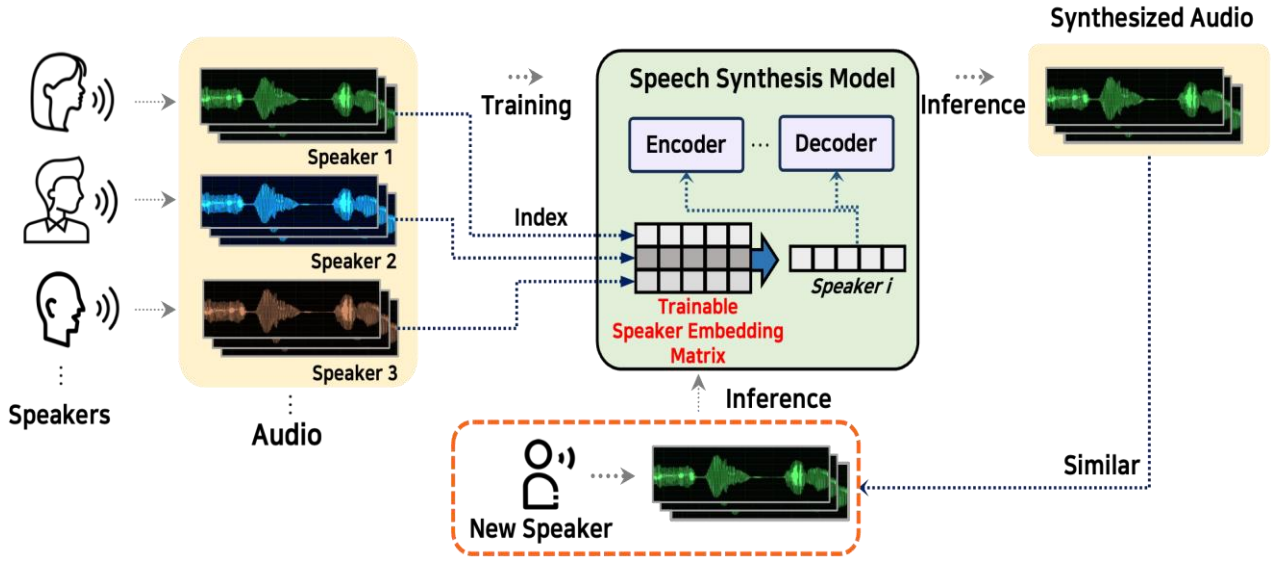
25th Apr, 2025

Contents

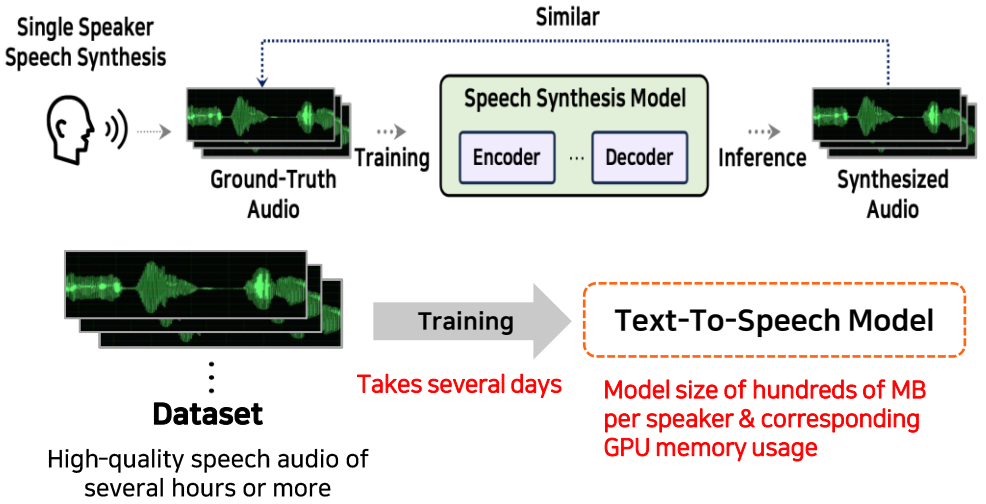
Introduction

Multi-Speaker Text-to-Speech

- Multi-Speaker Text-to-Speech (MS-TTS) with speaker representation learning are used in applications such as Interactive Voice Response and Voice User Interfaces, addressing the need for natural speech generation for new speakers or low-resource languages with minimal data [1].
- While conventional Text-to-Speech requires additional training for multiple speakers, MS-TTS equipped with speaker representation learning enables personalized speech synthesis from a short reference sample without fine-tuning, significantly enhancing real-world applicability [2].



[Figure 1] Illustration of speaker embedding extraction from multi-speaker data.



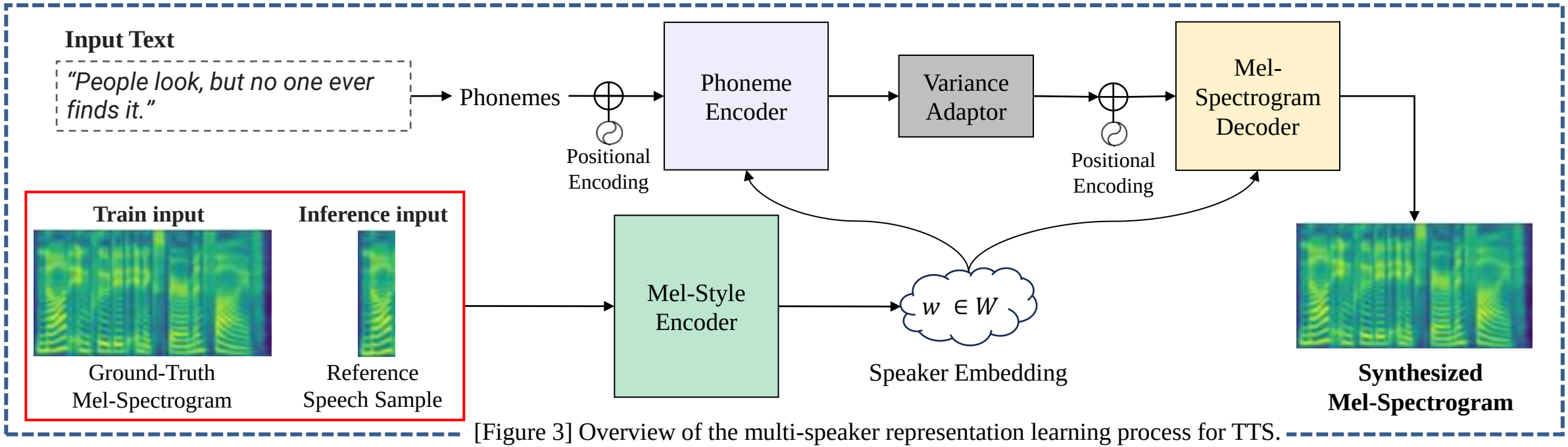
[Figure 2] Framework of Text-to-Speech

[1] Luo, R., Tan, X., Wang, R., Qin, T., Li, J., Zhao, S., Chen, E., Liu, T.-Y., 2021. LightSpeech: Lightweight and fast text to speech with neural architecture search. In: Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, June 6-11, 2021, pp. 5699–5703.
[2] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., GÅNolge, E., Ponti, M. A., 2022. YourTTS: Towards zero-shot multispeaker TTS and zero-shot voice conversion for everyone. In: Proceedings of the 39th International Conference on Machine Learning, Virtual Conference, July 18-24, 2022, pp. 2709–2720.

Introduction

Objective and Standard Procedure

- MS-TTS equipped with speaker representation learning aims to synthesize natural speech with fast inference speed for unseen speakers not included in training without requiring fine-tuning, using a few seconds of reference audio and text as input to generate a speech waveform [3].
- Conventional methods perform extracting speaker embeddings from reference audio and conditioning a TTS model, often using meta-learning, convolutional structures, or conditional layer normalization to improve speaker adaptation and synthesis quality [4, 5, 6].



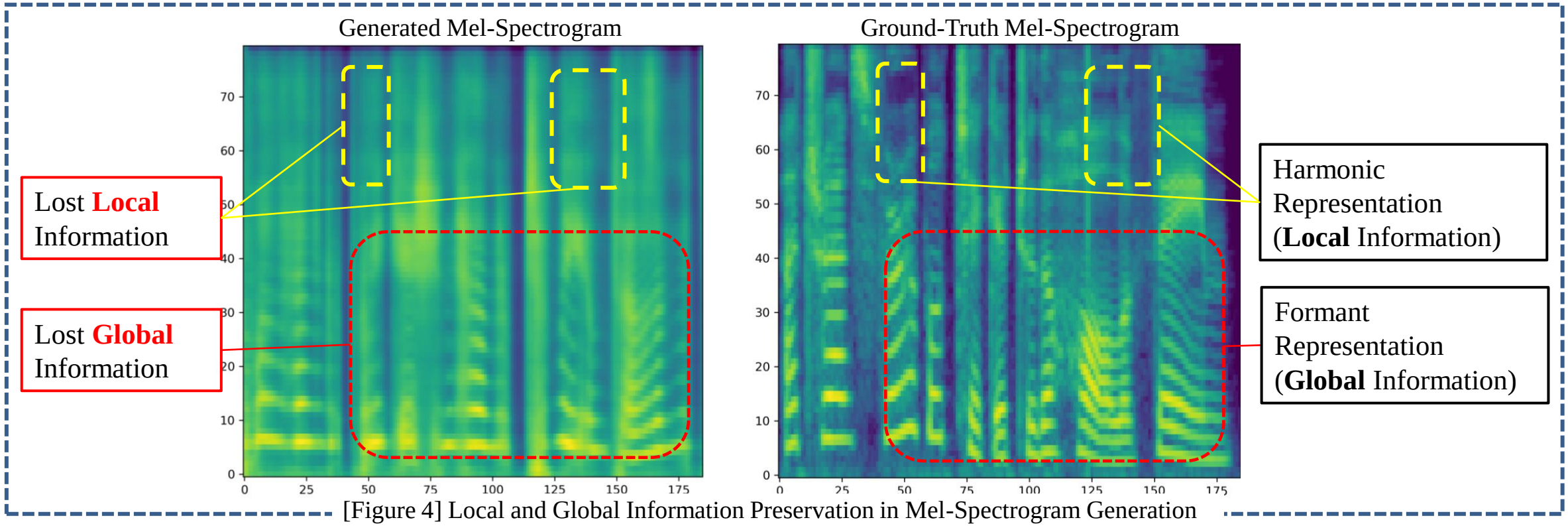
[Figure 3] Overview of the multi-speaker representation learning process for TTS.

[3] Atienza, R., 2023. Efficientspeech: An on-device text to speech model, In: Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 4-9, 2023, pp. 1–5.
[4] Yoon, H., Kim, C., Um, S., Yoon, H.-W., Kang, H.-G., 2023. SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. IEEE Signal Processing Letters, 30, 593–597.
[5] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: Proceedings of the 9th International Conference on Learning Representations, Virtual Conference, May 3-7, 2021, pp. 1–15.
[6] Min, D., Lee, D. B., Yang, E., Hwang, S. J., 2021. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation. In:Proceedings of the 38th International Conference on Machine Learning, Virtual Conference, July 18-24, 2021, pp. 7748–7759.

Introduction

Problem and Strategy

- Existing study [4] exhibits limitations in capturing frequency band characteristics of mel-spectrograms due to its global and local information mixing process, and for introducing an application that satisfies users, real-time processing speed and low latency are essential.
- We propose a method that separates the frequency spectrum into high- and low-frequency bands, enhancing expressiveness in the high-frequency band with deeper layers and improving Real-Time Factor (RTF) in the low-frequency band using a lightweight architecture.



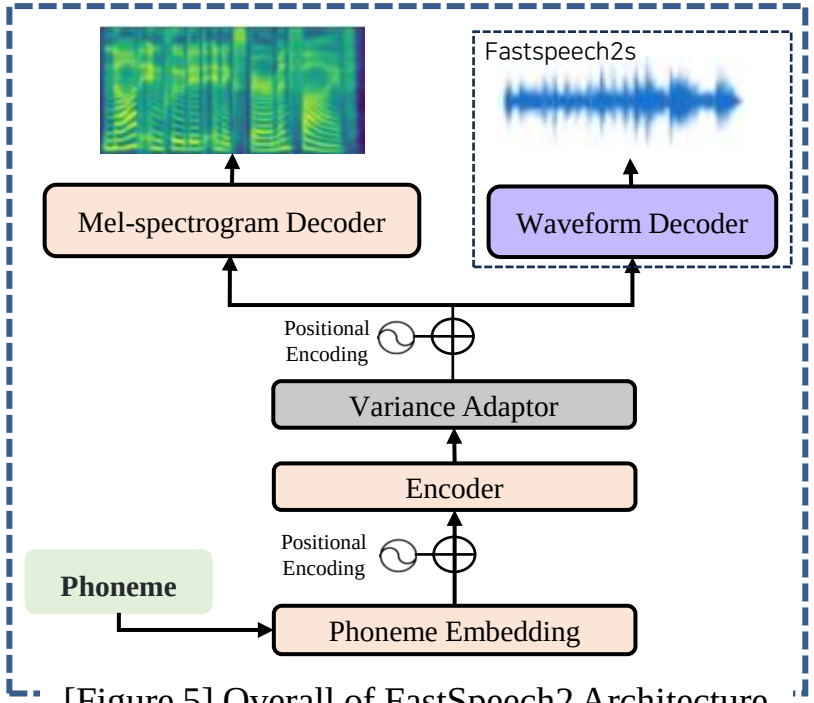
[Figure 4] Local and Global Information Preservation in Mel-Spectrogram Generation

[4] Yoon, H., Kim, C., Um, S., Yoon, H.-W., Kang, H.-G., 2023. SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. IEEE Signal Processing Letters, 30, 593–597.

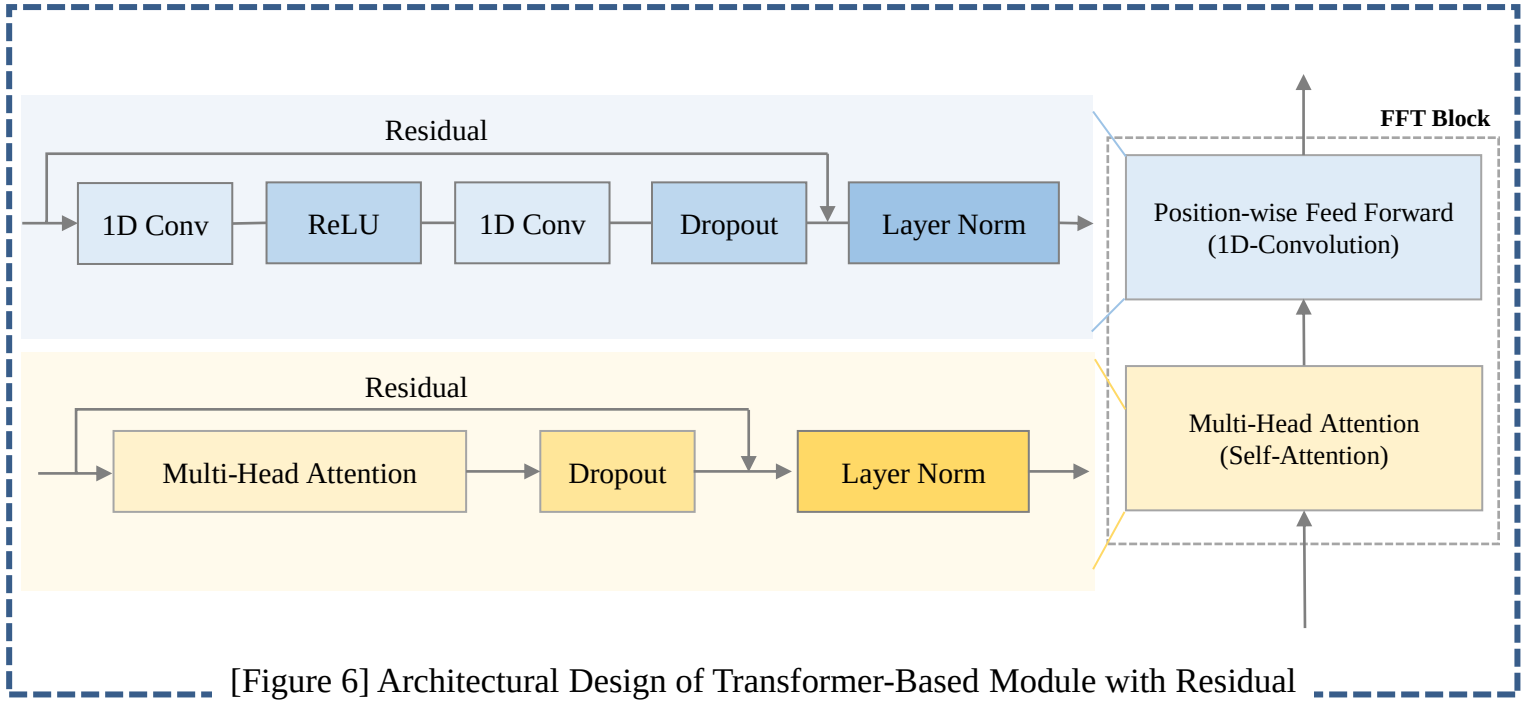
Related Work

FastSpeech2 [5]

- This study proposes a speech synthesis model that addresses the one-to-many mapping problem in non-autoregressive TTS by utilizing prosodic features such as pitch, energy, and duration as conditional inputs and directly training from ground-truth data without a teacher model.
- In this study, simultaneously performing pitch, energy, duration prediction, and mel-spectrogram generation in the Variance Adaptor improves speech quality, but the computational complexity limits real-time inference in resource-constrained environments.



[Figure 5] Overall of FastSpeech2 Architecture



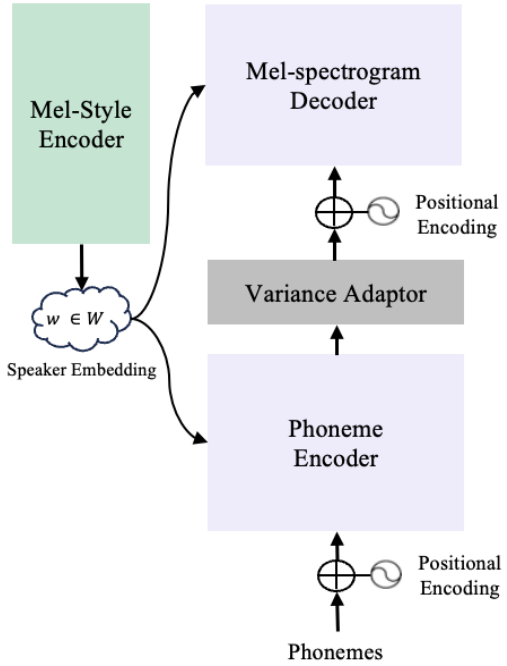
[Figure 6] Architectural Design of Transformer-Based Module with Residual Connections and Multi-Head Attention Mechanism

[5] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: Proceedings of the 9th International Conference on Learning Representations, Virtual Conference, May 3-7, 2021, pp. 1–15.

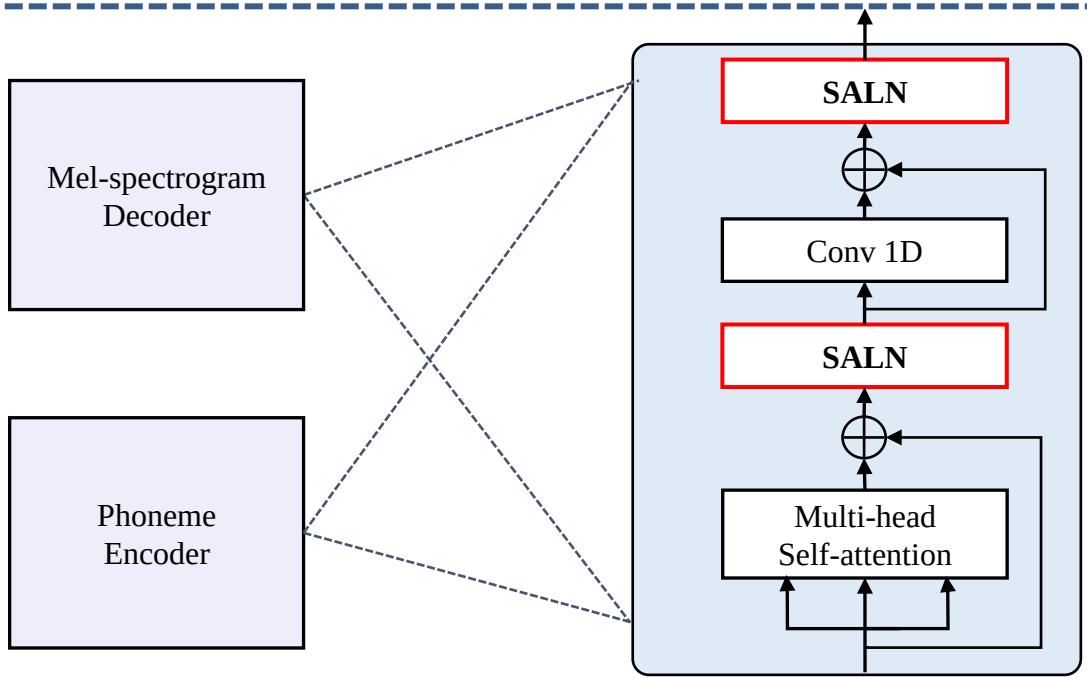
Related Work

Meta-StyleSpeech [6]

- This study proposes a multi-speaker adaptive TTS model using Style-Adaptive Layer Normalization (SALN), which aligns input text features based on a style vector extracted from short reference audio, enabling efficient synthesis without extensive fine-tuning.
- SALN is proposed to improve adaptability to unseen speakers by adjusting input features through style vectors extracted from reference audio, but the extra computations increase the computational complexity, making real-time inference difficult in resource-constrained environments.



[Figure 7] The architecture of StyleSpeech



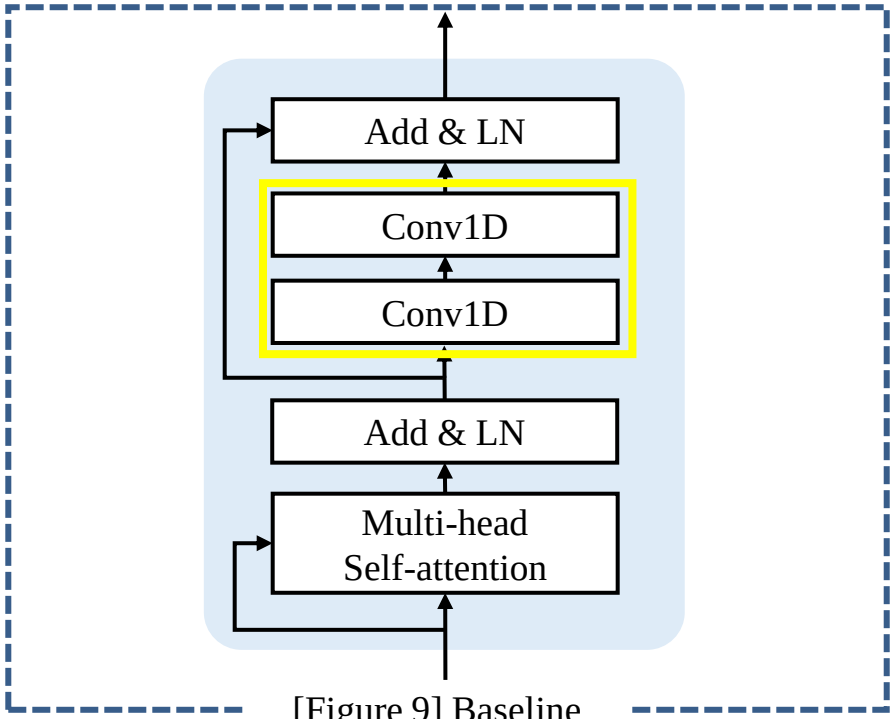
[Figure 8] Structure of Phoneme Encoder and Mel-spectrogram Decoder

[6] Min, D., Lee, D. B., Yang, E., Hwang, S. J., 2021. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation. In:Proceedings of the 38th International Conference on Machine Learning, Virtual Conference, July 18-24, 2021, pp. 7748–7759.

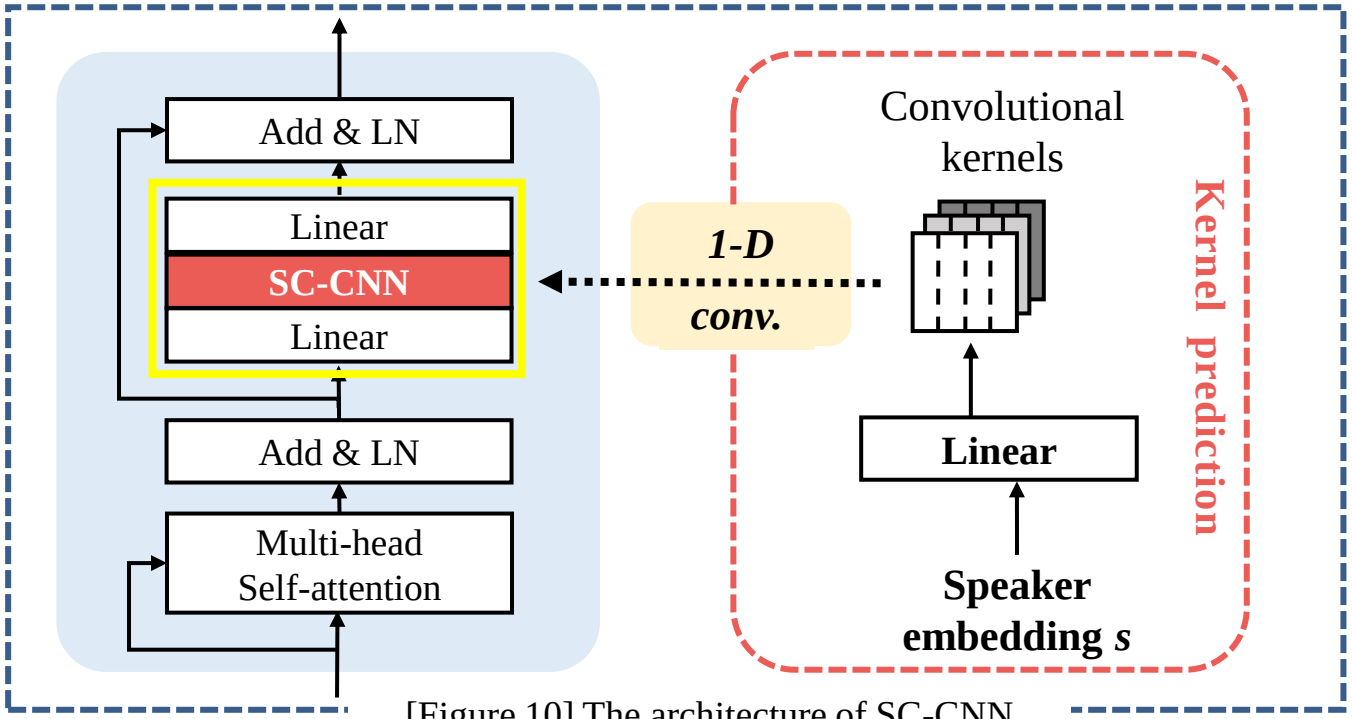
Related Work

SC-CNN [4]

- This research captures diverse speaker characteristics in phoneme sequences by predicting convolutional kernels reflecting unique speaker features through trained embeddings and applying them via 1-D convolution and speaker-conditioning layers to the acoustic model.
- The model facilitates real-time inference in resource-constrained environments, however its linear layer architecture limits the capture of local speaker characteristics, leading to reduced speaker adaptability and accuracy for unseen speakers in MS-TTS with speaker representation learning.



[Figure 9] Baseline



[Figure 10] The architecture of SC-CNN

[4] Yoon, H., Kim, C., Um, S., Yoon, H.-W., Kang, H.-G., 2023. SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. IEEE Signal Processing Letters, 30, 593–597.

Related Work

Common Drawback and Solution

- Existing models process both high- and low-frequency components through a unified pipeline, making it difficult to accurately reflect the distinct characteristics of global and local information, which ultimately limits speech quality and speaker expressiveness.
- We divide the frequency spectrum into high- and low-frequency bands, using deeper layers for enhanced expressiveness in the high band and a lightweight architecture for improved RTF in the low band.

[Table 1] Counterpart Summary

Model	Author	Publication	Summary	Drawback
FastSpeech2	Yi Ren et al.	International Conference on Learning Representations	• Using pitch, energy, and duration information to solve the one-to-many mapping problem in non-autoregressive TTS prevents information loss and reduces the information gap between text and mel-spectrogram.	• The Variance Adaptor and additional preprocessing steps increase complexity, limiting real-time inference in resource-constrained environments.
Meta-StyleSpeech	Dongchan Min et al.	International Conference on Machine Learning	• The research proposes Style-Adaptive Layer Normalization (SALN) to adjust input text features using style vectors from reference audio, enabling efficient multi-speaker TTS for unseen speakers without extensive fine-tuning.	• Inference Limitations in Resource-Constrained Environments Due to Style Vector Extraction and Feature Adjustment
SC-CNN	Hyungchan Yoon et al.	IEEE Signal Processing Letters	• Uses a CNN-based simple structure to enable fast and efficient voice synthesis.	• Insufficient capture of global speaker characteristics

Proposed Method

Notation and Definition

- Each data sample comprises a speaker identifier sid, an input text sequence $text$, and the corresponding Ground-Truth mel_{target} , further divided into low-frequency bands mel^{lo}_{target}(bins 1–40) and high-frequency bands mel^{hi}_{target} (bins 41 and above).
- The mel-style encoder, Enc_s takes a reference speech X as input and mel-style encoder extract a vector $w \in \mathbb{R}^N$ which contains the style such as speaker identity and prosody of given speech X.

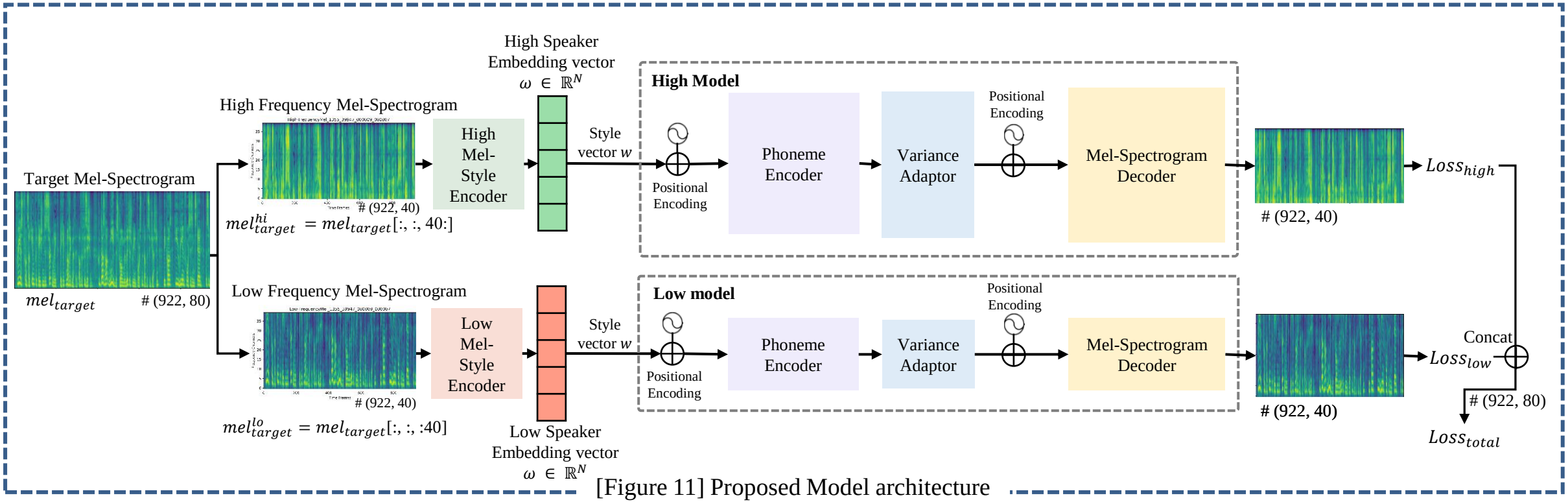
[Table 2] Notation Symbol Definition

Notation	Name	Definition
$sid, text$	Speaker ID, Text sequence	Speaker identifier and Input text sequence
X	Reference speech	Mel-Style Encoder takes a reference speech as input
mel_{target}	Target Mel-spectrogram	Ground-truth Mel-spectrogram
$mel^{lo}_{target} = mel_{target}[:, :, :40]$	Low-frequency bands	Low-frequency bands (1–40 bins) of the target Mel-spectrogram
$mel^{hi}_{target} = mel_{target}[:, :, 40:]$	High-frequency bands	High-frequency bands (41 bins and above) of the target Mel-spectrogram
$w \in \mathbb{R}^N$	Style vector	Contains the style of speaker identity and prosody of given speech
$v_{(\cdot)}, g_{(\cdot)}, b_{(\cdot)}$	Direction, gain, bias	Denote the direction, gain and bias for kernels, respectively, obtained by passing the speaker embedding s through the linear projection layer
$(\cdot)_D, (\cdot)_P$	Depth- and Point-wise convolution	Represent the corresponding types of convolutions, depth- and point-wise

Proposed Method

Procedural Overview

- The proposed method divides the entire frequency range into different time segments and train each model separately, optimizing processing based on specific frequency characteristics.
- In the high-frequency range, the SC-CNN structure is retained to enhance expressiveness, while in the low-frequency range, it is removed and replaced with a lightweight structure to improve RTF by accelerating real-time processing.

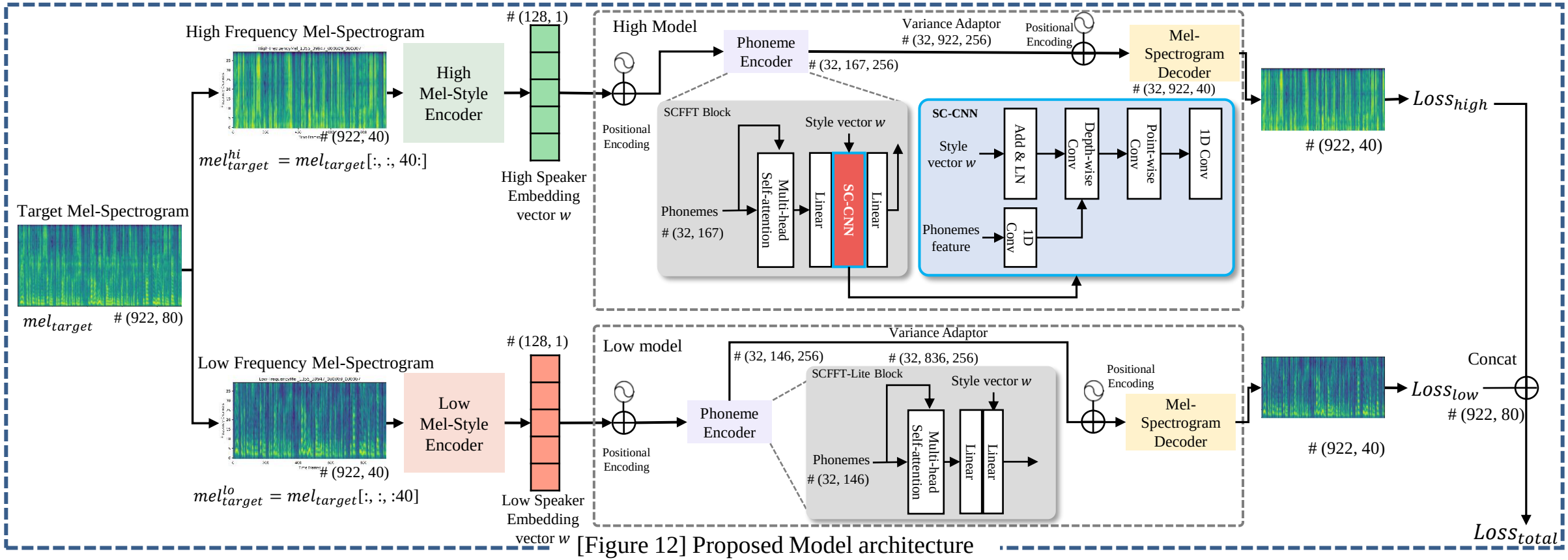


[Figure 11] Proposed Model architecture

Proposed Method

Details

- Proposed model processes high and low-frequency mel-spectrograms via High and Low Models, using Mel-Style encoder-extracted speaker embedding vectors to train spectrogram characteristics through phoneme encoder decoders and reducing total loss by merging outputs.
- High Model with Self-Attention and Multi-head Attention layers, Low Model SCFFT-Lite Block capture global information, and High Model Convolutional layers extract local features, training mel-spectrogram characteristics.



[Figure 12] Proposed Model architecture

Experimental Settings

Dataset

- The experimental setup uses LibriTTS [7] subsets, including train-clean-100 and dev-clean, as source domain training data and test-clean as the target domain.
- The LibriTTS dataset comprises 251 speakers with 32,535 utterances in the train-clean-100 subset for training and 40 speakers each in dev-clean and test-clean subsets, consisting of 5,710 and 4,820 utterances, respectively, for validation and testing purposes.

[Table 3] Datasets Information

Dataset	Data Split	Subset	Total Speakers	Total utterances
LibriTTS [7]	Train	train-clean-100	251	32535
	Val	dev-clean	40	5710
	Test	test-clean	40	4820

Experimental Settings

Hyperparameter Settings

[Table 4] Hyperparameter Settings

Model	Batch size	Max iteration	Loss function	Optimizer	Learning rate
Yi Ren et al. [5]	64	200K	Mean absolute error, Mean Squared Error Loss	Adam	2e-4
Dongchan Min et al. [6]			Mean Squared Error Loss, Mean Absolute Error Loss, Cross Entropy Loss		
Hyungchan Yoon et al. [4]	16		Mean absolute error, Mean Squared Error Loss		

Evaluation Measures

- *Perceptual Evaluation of Speech Quality (PESQ)*

  = Mean Opinion Score – Listening Quality Objective (MOS-LQO)

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \qquad PESQ_{MOS} = \alpha + \frac{b}{1 + e^{(c \cdot D + d)}}$$

- *Mel Cepstral Distortion (MCD)*

$$MCD_{dB} = \frac{\alpha}{N} \sum_{t=0}^{N-1} \sqrt{\sum_{k=1}^{P} (MC_{syn}(t,k) - MC_{ref}(t,k))^2}$$

Experimental Results

Performance Comparison

- The proposed model preserves a low real-time factor (0.175), comparable to SC-CNN (0.168), while achieving higher perceptual quality (1.301 vs. 1.113) and lower mel-cepstral distortion (26.055 vs. 26.600), demonstrating a favorable trade-off between efficiency and speech quality.
- The proposed model achieves an RTF of 0.175, satisfying the real-time constraint (RTF < 1.0), unlike FastSpeech2 and Meta-StyleSpeech, which fail to meet the threshold, thereby validating its applicability to real-time TTS systems.

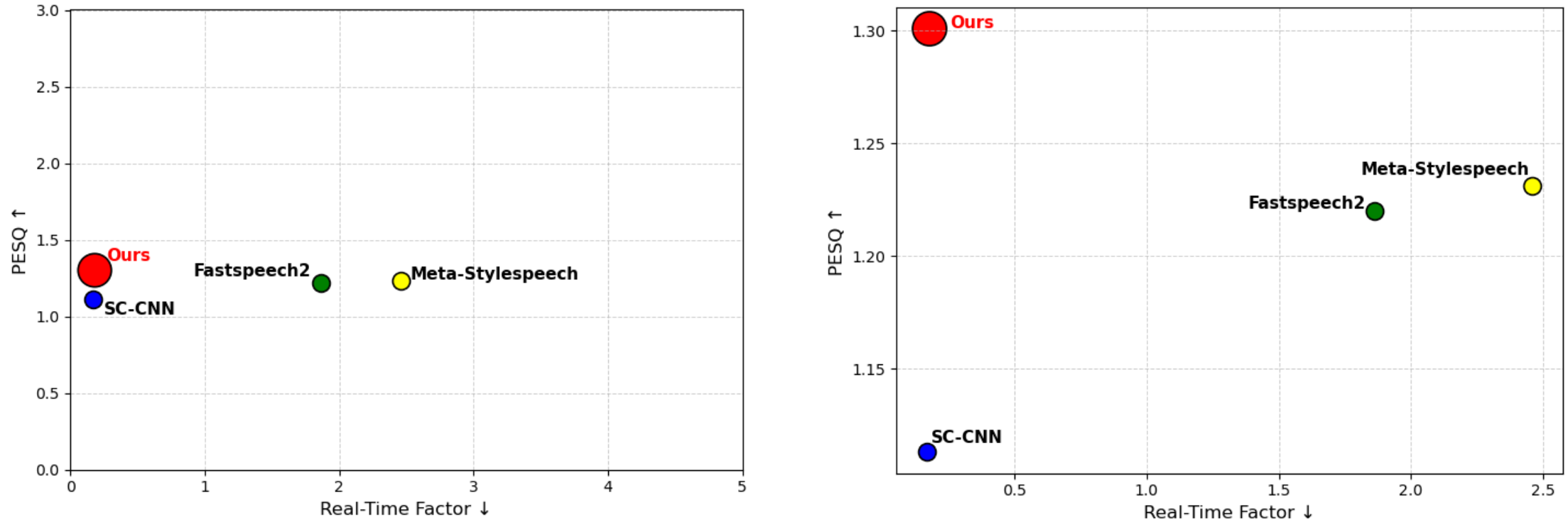
[Table 5] Performance evaluation through MCD, RTF, and PESQ. MCD measures the mel-cepstral distortion between synthesized and reference audio, with lower values indicating better quality. RTF represents the inference speed, where values below 1 indicate real-time capability. PESQ quantifies speech quality, with higher values corresponding to better perceptual quality.

Model	RTF (< 1)	MCD(↓)	PESQ(↑)
Proposed	**0.175**	**26.055**	**1.301**
SC-CNN [4]	0.168	26.600	1.113
FastSpeech2 [5]	**1.863 (Fail)**	14.491	1.220
Meta-StyleSpeech [6]	**2.459 (Fail)**	14.570	1.231

Experimental Results

Performance Comparison

- The proposed model achieves the best perceptual speech quality (PESQ) among all compared methods while maintaining a real-time factor (RTF) well below 1, indicating both high-quality synthesis and suitability for real-time applications.
- Unlike other models, which show trade-offs between inference speed and speech quality, the proposed method effectively balances both metrics, dominating the Pareto front in the zoomed-in region.

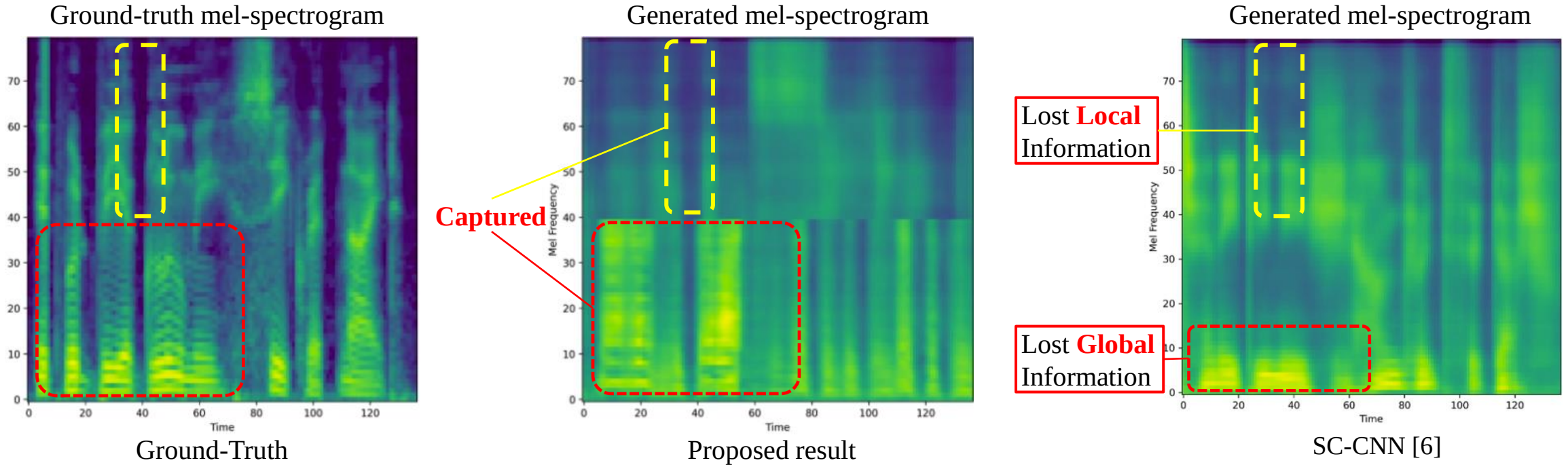


[Figure 13] Comparison of Multi-Speaker Model Performance : PESQ vs. RTF (Left: Full Range, Right: Zoomed-in View).

Experimental Results

In-depth Analysis

- The original mel-spectrogram fails to accurately capture the formant characteristics and fundamental phonation patterns, demonstrating significant limitations in representing harmonic representations and local details during silent intervals.
- Our method effectively reconstructs global information via clear vowel and fundamental frequency (F0) patterns in the low-frequency band while capturing local spectral characteristics and silent segments in the high-frequency band, enhancing overall speech synthesis quality.



[Figure 14] Comparative Analysis of Mel-Spectrogram Generation

Conclusion and Achievement

- This study aimed to address the information loss problem in existing speech synthesis models and overcome the limitations of local and global information capture to achieve both high precision and real-time processing speed.
- The proposed methodology separates the mel-spectrogram frequency spectrum into high and low-frequency bands, improving expressiveness with deeper and improving real-time processing speed using a lightweight architecture.
- We evaluate our method on the LibriTTS dataset, and our approach outperforms existing methods by achieving lower real-time factor and higher perceptual speech quality in speech synthesis experiments.

Achievement

- Conferences
  - Recent Research Trends of CNN-based Music Emotion Recognition, ***IEIE***, 2023
  - A Simple Localization and Classification Framework Based on Template Matching and Self-Attention for TAO Treatment Diagnosis, ***ICCE***, 2024
  - Briefing of Text-To-Speech Trends Toward Zero-Shot Multi-Speaker Synthesis, ***ICCE***, 2025
- Others
  - Medical Technology for Patients with Thyroid-associated Ophthalmopathy, ***AI Technology-based Employment and Start-up Workshop***, 2023, 3rd Prize
  - ***2023 Singapore K-Tech Festival - AI/ML Innovation Research Center Booth***, Singapore, Singapore, 2023

Thank you

Reference

[1] Luo, R., Tan, X., Wang, R., Qin, T., Li, J., Zhao, S., Chen, E., Liu, T.-Y., 2021. LightSpeech: Lightweight and fast text to speech with neural architecture search. In: Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, June 6-11, 2021, pp. 5699–5703.

[2] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., GÅNolge, E., Ponti, M. A., 2022. YourTTS: Towards zero-shot multispeaker TTS and zero-shot voice conversion for everyone. In: Proceedings of the 39th International Conference on Machine Learning, Virtual Conference, July 18-24, 2022, pp. 2709–2720.

[3] Atienza, R., 2023. Efficientspeech: An on-device text to speech model, In: Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 4-9, 2023, pp. 1–5.

[4] Yoon, H., Kim, C., Um, S., Yoon, H.-W., Kang, H.-G., 2023. SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. IEEE Signal Processing Letters, 30, 593–597.

[5] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: Proceedings of the 9th International Conference on Learning Representations, Virtual Conference, May 3-7, 2021, pp. 1–15.

[6] Min, D., Lee, D. B., Yang, E., Hwang, S. J., 2021. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation. In:Proceedings of the 38th International Conference on Machine Learning, Virtual Conference, July 18-24, 2021, pp. 7748–7759.

[7] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association, Interspeech, Graz, Austria, September 15-19, 2019, pp. 1526–1530.