ICASSP 2020 virtual conference: 4 to 8 May, 2020
Machine Learning for Speech Synthesis I: TU2.PA.10

45th IEEE International Conference on Acoustics, Speech and Signal Processing
ICASSP 2020
May 4 — 8, 2020, Barcelona, Spain

Transformer-based text-to-speech with weighted forced attention

Takuma Okamoto[1], *Tomoki Toda*[2,1], *Yoshinori Shiga*[1] and *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology (NICT), Japan
[2]Nagoya University, Japan

NICT National Institute of Information and Communications Technology

Outline

- Introduction
- Problems and purpose

- Proposed Transformer-based acoustic model with weighted forced attention
- FastSpeech without duration predictor

- Experiments

- Additional experiments (Not included in proceeding)

- Portable real-time neural TTS demo system on a laptop with a GPU

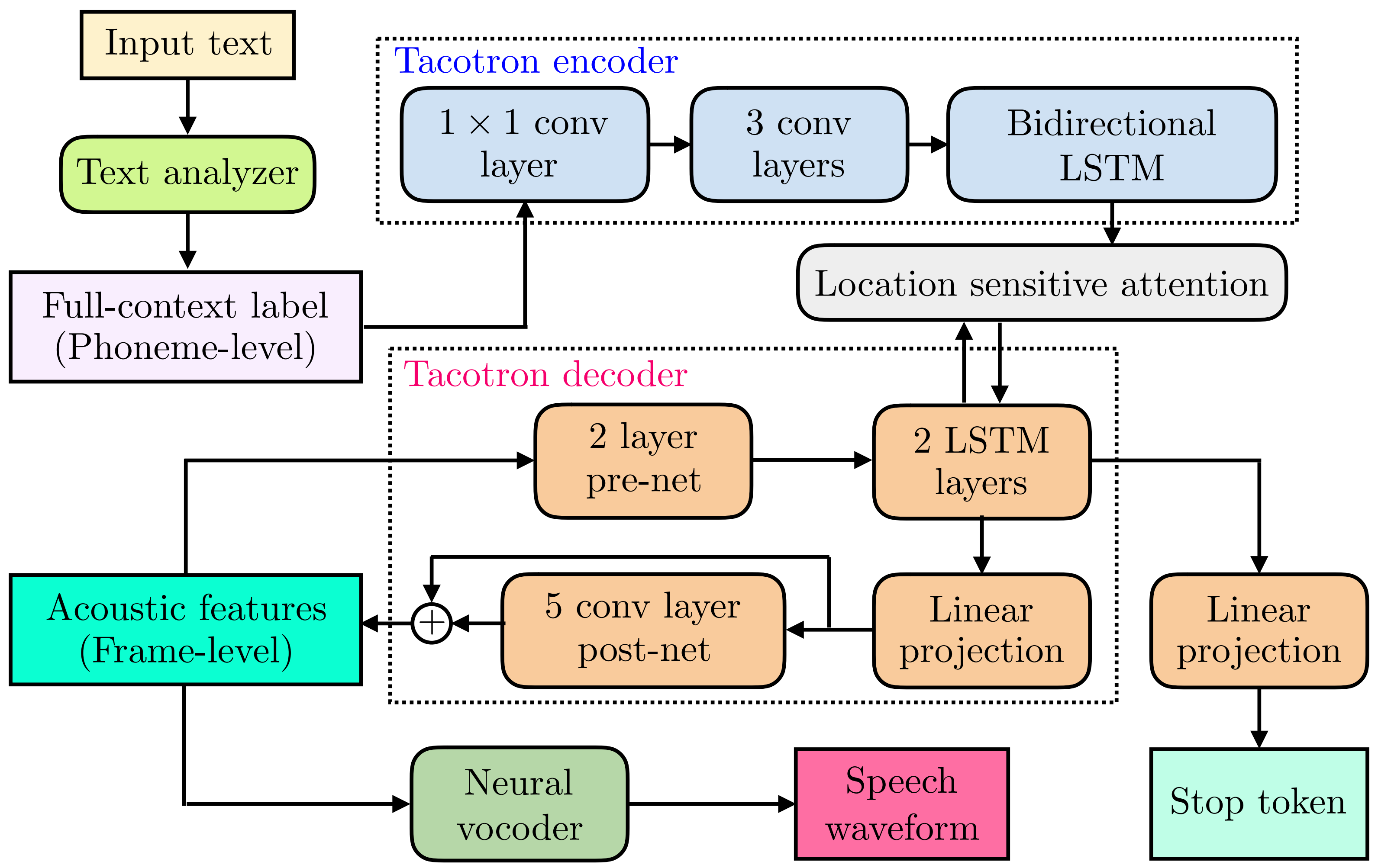
- Conclusions

Introduction

- High-fidelity text-to-speech (TTS) systems
  - WaveNet outperformed conventional TTS systems in 2016 -> End-to-end neural TTS
- Tacotron 2 (+ WaveNet vocoder) J. Shen *et al.*, ICASSP 2018
  - Text (English) -> [Tacotron 2] -> mel-spectrogram -> [WaveNet vocoder] -> speech waveform
  - Jointly optimizing text analysis, duration and acoustic models with a single neural network
    - No text analysis, no phoneme alignment, and no fundamental frequency analysis

Realizing high-fidelity speech synthesis comparable to human speech!!

  - Problem
    - NOT directly applied to pitch accent languages
- Tacotron 2 with full-context label input
  - Capable for pitch accent languages (e. g. Japanese)
    - Realizing real-time neural TTS with Tacotron 2 and WaveGlow
  - Crucial problem for actual implementations
    - Sometimes unstable in inference (skip or stop)



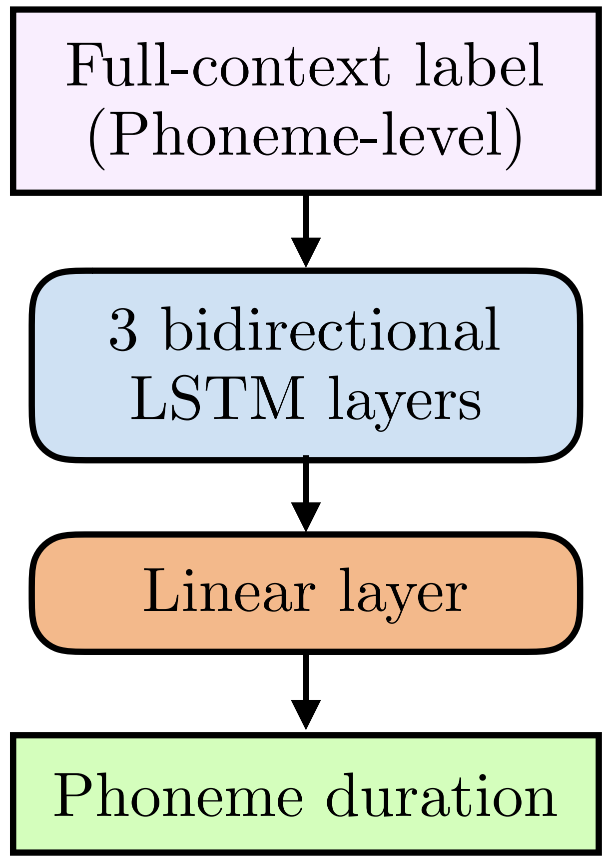
T. Okamoto *et al.*, Interspeech 2019

Tacotron-based stable neural TTS model

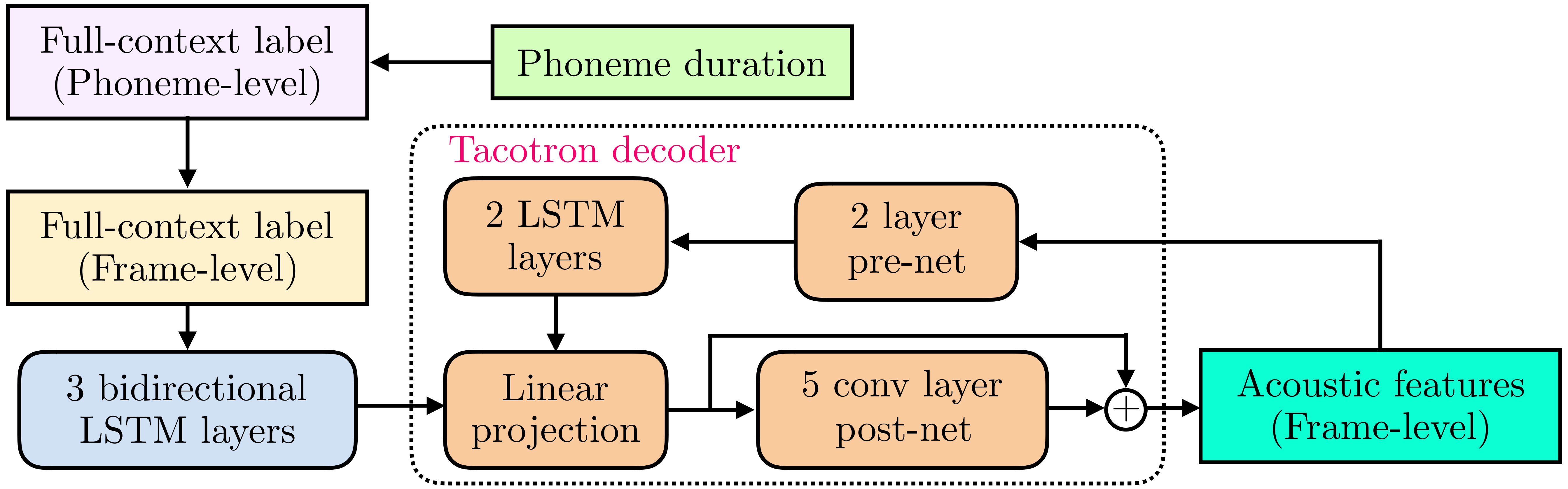
T. Okamoto *et al.*, ASRU 2019

High-fidelity and stable acoustic model (AM)

- Conventional bidirectional LSTM-based duration model <- more stable compared with sequence-to-sequence models
  - Trained with HMM-based forced alignment
- Tacotron-based acoustic model with full-context label input
  - HMM-based forced alignment in training
  - Predicted phoneme durations are used in inference



(a) Duration model



(b) acoustic model: BLSTM+Taco2dec

High-fidelity, real-time and stable TTS can be realized with WaveGlow vocoder!!

Problems and purpose

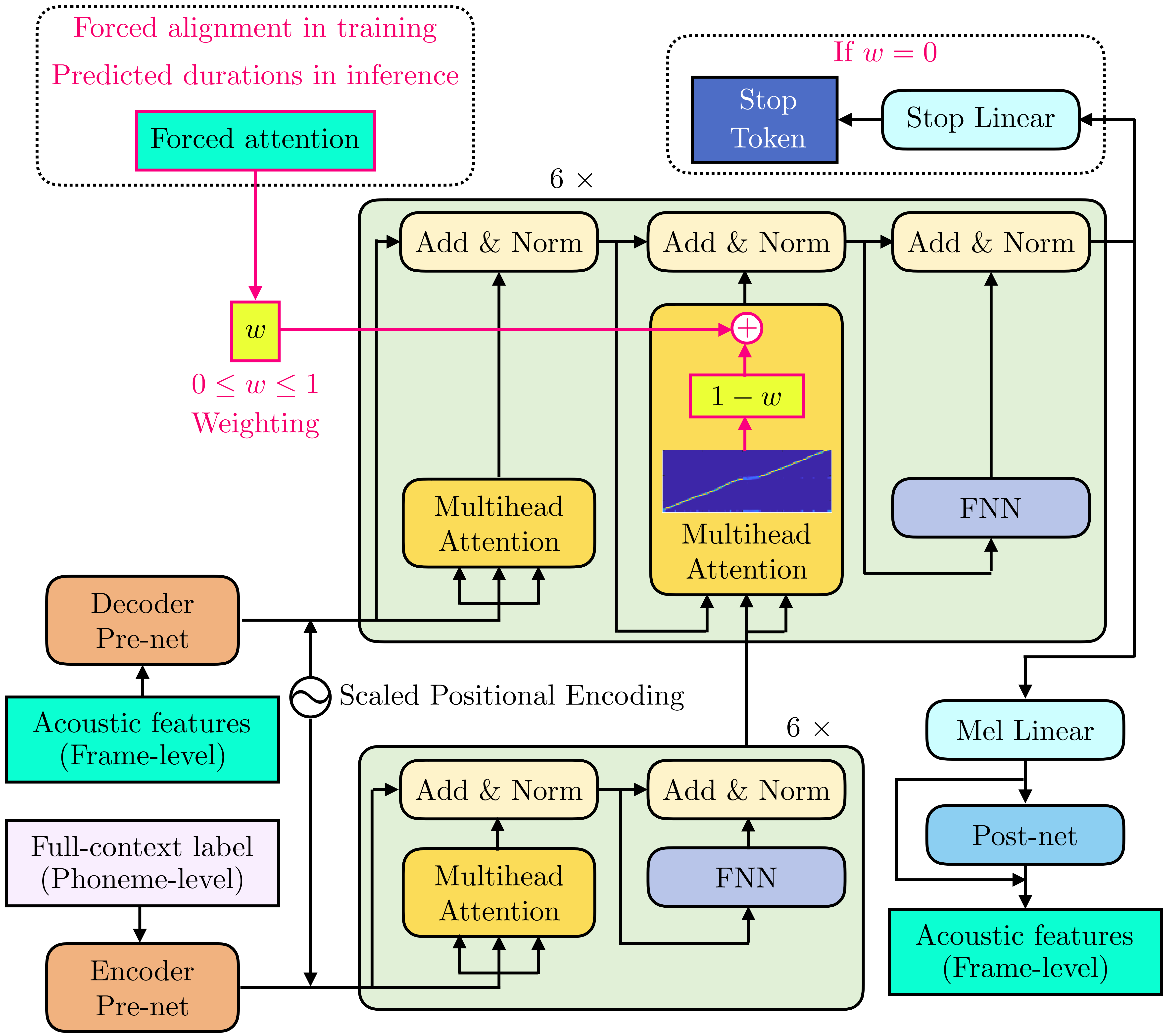
- Problem in RNN-based models (Tacotron 2 and BLSTM+Taco2dec)
  - Slower training period than CNN- and self-attention-based models (Transformer and FastSpeech)
- Problem in sequence-to-sequence models (Tacotron 2 and Transformer)
  - Sometimes unstable in inference (skip or stop)
    - Stable inference with phoneme durations (BLSTM+Taco2dec and FastSpeech)
- Problems in self-attention-based acoustic models (Transformer and FastSpeech)
  - Only phoneme input is investigated for English TTS
  - Teacher-student training (teacher Transformer) is required for FastSpeech

- Purpose of this study
  - Investigating Transformer- and FastSpeech-based AMs with full-context label input for pitch accent languages
  - Introducing HMM-based phoneme alignment to Transformer- and FastSpeech-based AMs
    - Stable inference for Transformer-based TTS
    - Removing teacher-student training and duration predictor in FastSpeech

Transformer-based TTS with weighted forced attention

- **Transformer-based TTS** N. Li *et al*., AAAI 2019
  - Feedforward network and self-attention instead of RNN
    - Faster training than RNN-based models (e. g. Tacotron 2)
    - Only phoneme input is investigated for English TTS

- **Proposed Transformer-based TTS**
  - Full-context label input for pitch accent languages
  - Introducing wighted forced attention
    - HMM-based forced alignment in training
    - Duration predicted by conventional model in inference
    - Both multihued attention and predicted duration are simultaneously used with a weighting factor
    - For case of $w = 1$, hidden features from encoder is too redundant (ASRU 2019) -> importance of weighting
    - Proposed AM can be trained without "stop token" loss



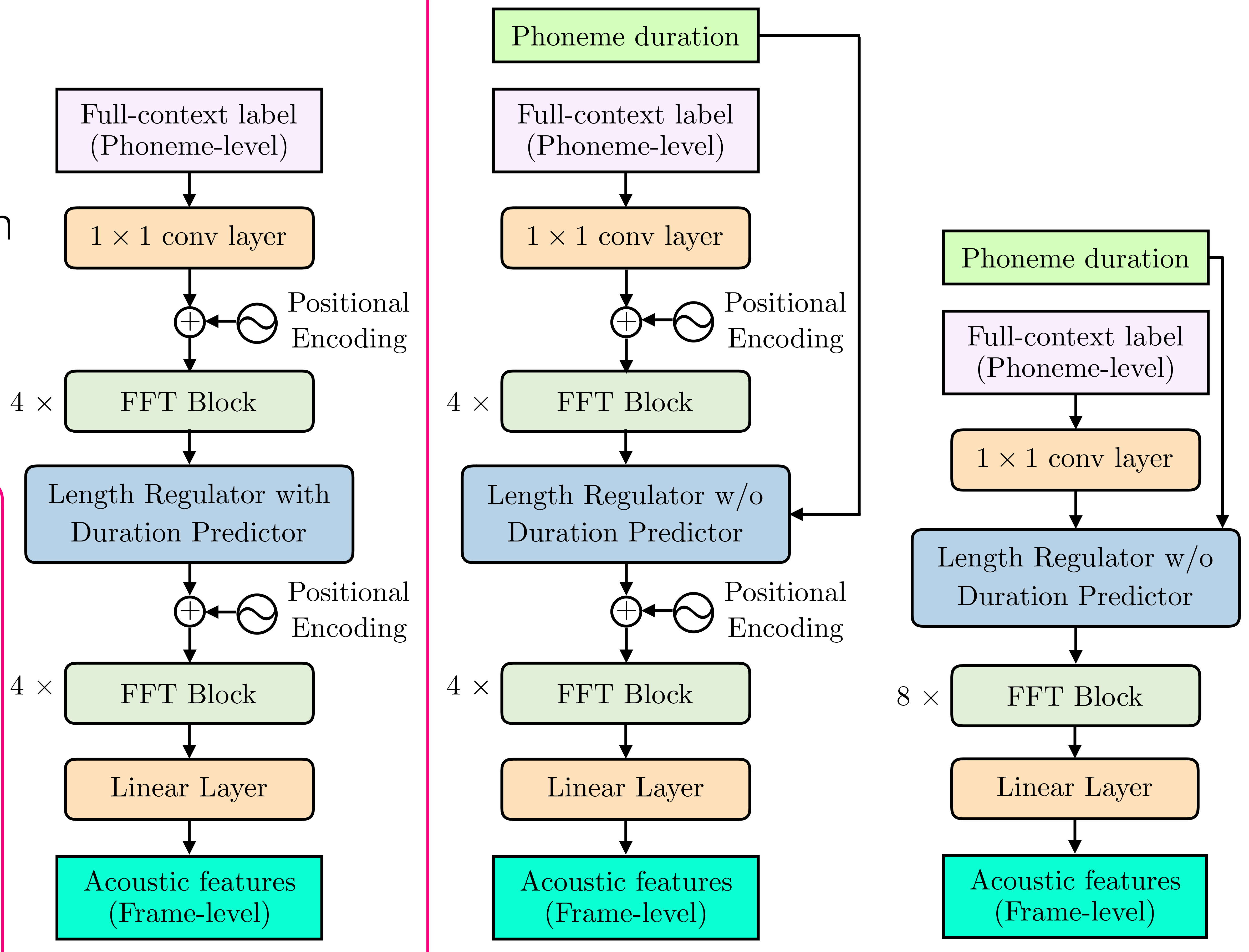
High-fidelity and stable TTS with faster training than RNN-based models is expected!!

FastSpeech without duration predictor

- **FastSpeech** Y. Ren et al., NeurIPS 2019
  - Feedforward Transformer without any recurrent connections
    - Not only fast training but also fast inference
  - Duration predictor trained from teacher Transformer's attention
    - Duration and acoustic models are jointly trained
  - Teacher-student training for improving synthesis accuracy

- **FastSpeech without duration predictor**
  - Duration and acoustic models are separately trained
    - HMM-based forced alignment in training
    - Durations predicted by conventional model in inference

- **FastSpeech with simple structure**
  - Without encoder-decoder structure and positional encodings



(a) Default (b) w/o-DP (c) Simple

They are expected to concentrate to optimize only acoustic features for higher accuracy!!

Experimental conditions

- Speech corpus: Sampling frequency: 24 kHz
  - Japanese female corpus: about 22 h (test set: 80 utterances)

- Acoustic models
  - Input: full-context label vector (130 dim)
  - Output acoustic feature: Mel-spectrograms (80 dim)
  - Sequence-to-sequence models
    - Tacotron 2 (Interspeech 2019), Transformer (FNN: default), Transformer (Conv1D used in FastSpeech)
  - Pipeline models with BLSTM-based duration model
    - BLSTM
    - BLSTM+Taco2dec(ASRU 2019)
    - Proposed Transformer with weighted forced attention (weightings are 0.2, 0.5, 0.7 and 1.0)
    - FastSpeech (default) with HMM-based forced alignment without teacher Transformer
    - Fastspeech without duration predictor
    - Fastspeech with simple structure

- Neural vocoder: WaveGlow with 512 channels

Results of training period (TP) and real-time factor (RTF)

Evaluation condition

- Using an NVIDIA Tesla V100 GPU in inference
- Simple PyTorch implementation

Notations

- TF: Transformer
- WFA: Weighted forced attention
- FS: FastSpeech
- DP: duration predictor

Method	TP (days)	AM RTF	Total RTF
(A):Tacotron 2	24	0.063	0.13
(B):TF (FNN)	6	0.55	0.62
(C):TF (Conv1D)	6	0.55	0.62
(D):BLSTM	3	0.015	0.12
(E):BLSTM+Taco2dec	12	0.061	0.13
(F)-(I):TF-WFA	6	0.55	0.62
(J):FS (Default)	6	0.004	0.070
(K):FS (w/o-DP)	6	0.004	0.072
(L):FS (Simple)	6	0.004	0.072
Duration model	2	-	0.002
WaveGlow vocoder	30	-	0.066

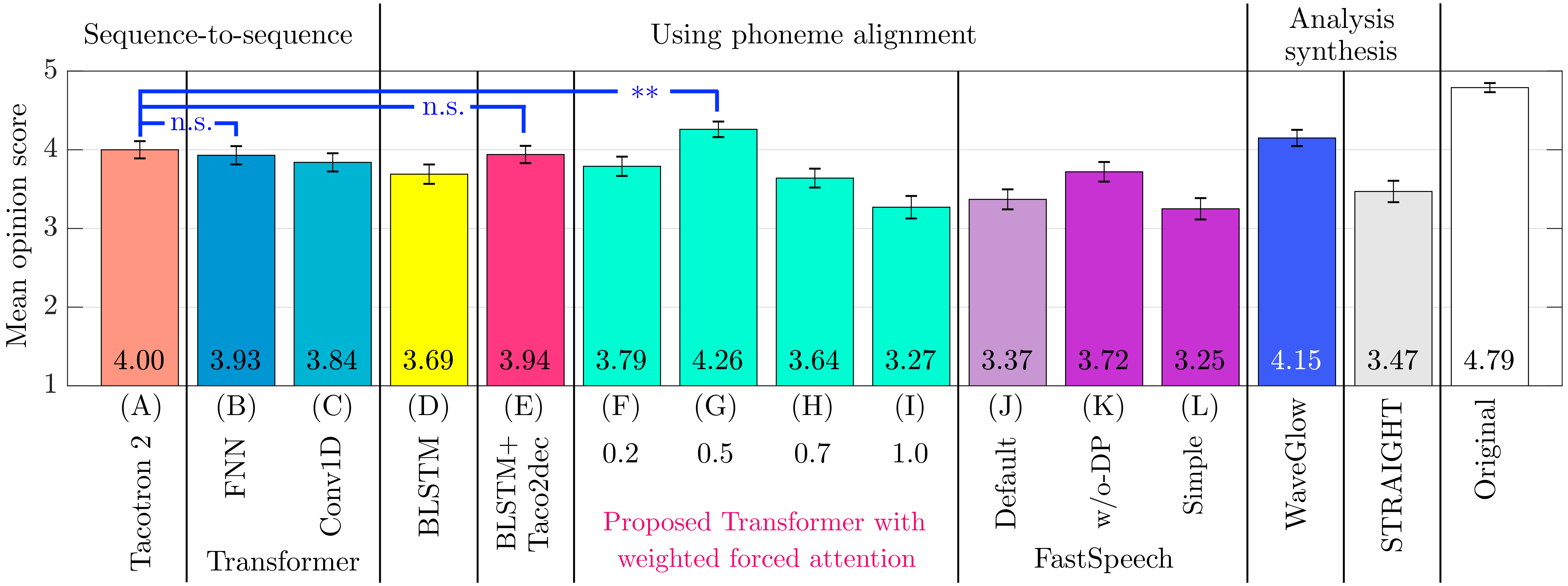
Results

- All models can realize real-time neural TTS with a GPU although Transformer-based model is not so fast
- Transformer and FastSpeech can realize faster training than Tacotron 2 and BLSTM+Taco2dec
- FastSpeech can realize fastest inference speed compared with other AMs

MOS results

Subjective evaluation

- Listening subjects: 20 Japanese native speakers
- 15 conditions x 20 utterances (successfully synthesized by all models) = 300 sentences / a subject

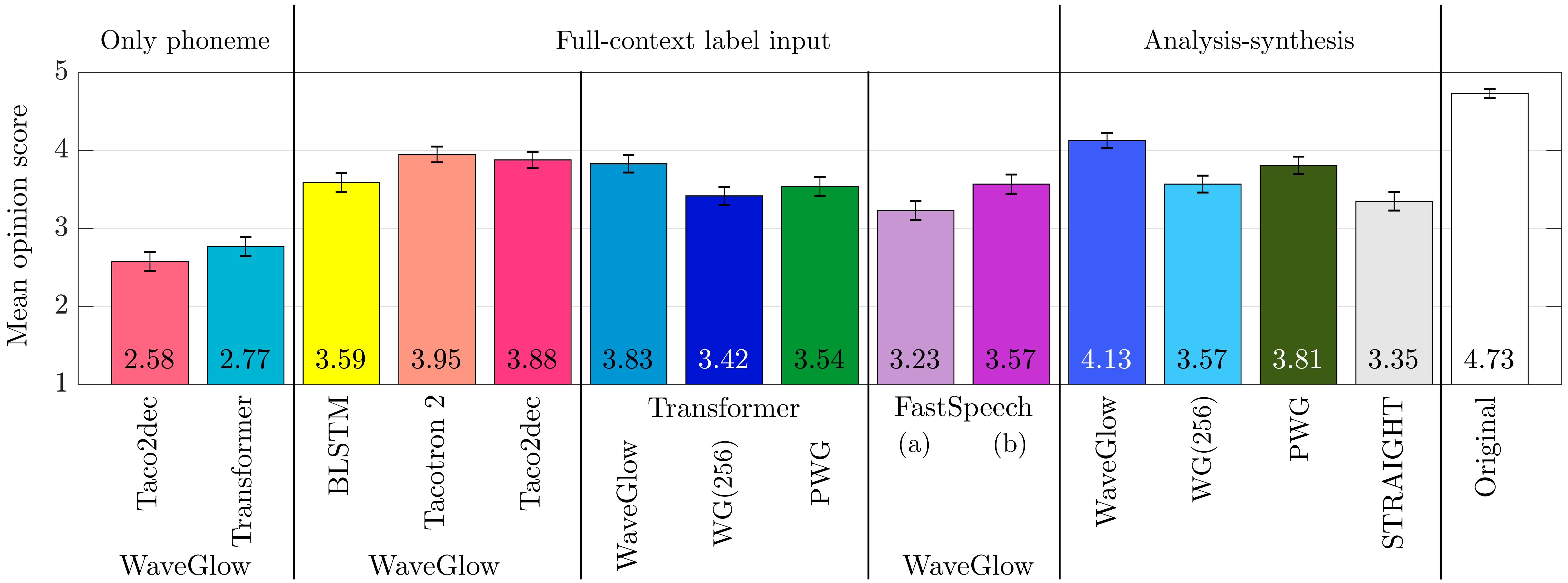


Results

- Proposed Transformer-based AM with a weighting factor of 0.5 can significantly outperform other models
- FastSpeech without duration predictor can realize higher synthesis quality than that with duration predictor
- Proposed Transformer-based AMs with weighted forced attention included some unsuccessfully synthesized samples
  - Encoder and decoder attentions were not diagonal

Additional results (Not included in proceeding)

- Additional experiments after submission of ICASSP 2020
  - Only phoneme input condition
  - Parallel WaveGAN (PWG): R. Yamamoto et al., ICASSP 2020
    - Training period: 2 days, Real-time factor: 0.031
  - Small WaveGlow model with 256 channels
    - Training period: 12 days, Real-time factor: 0.030

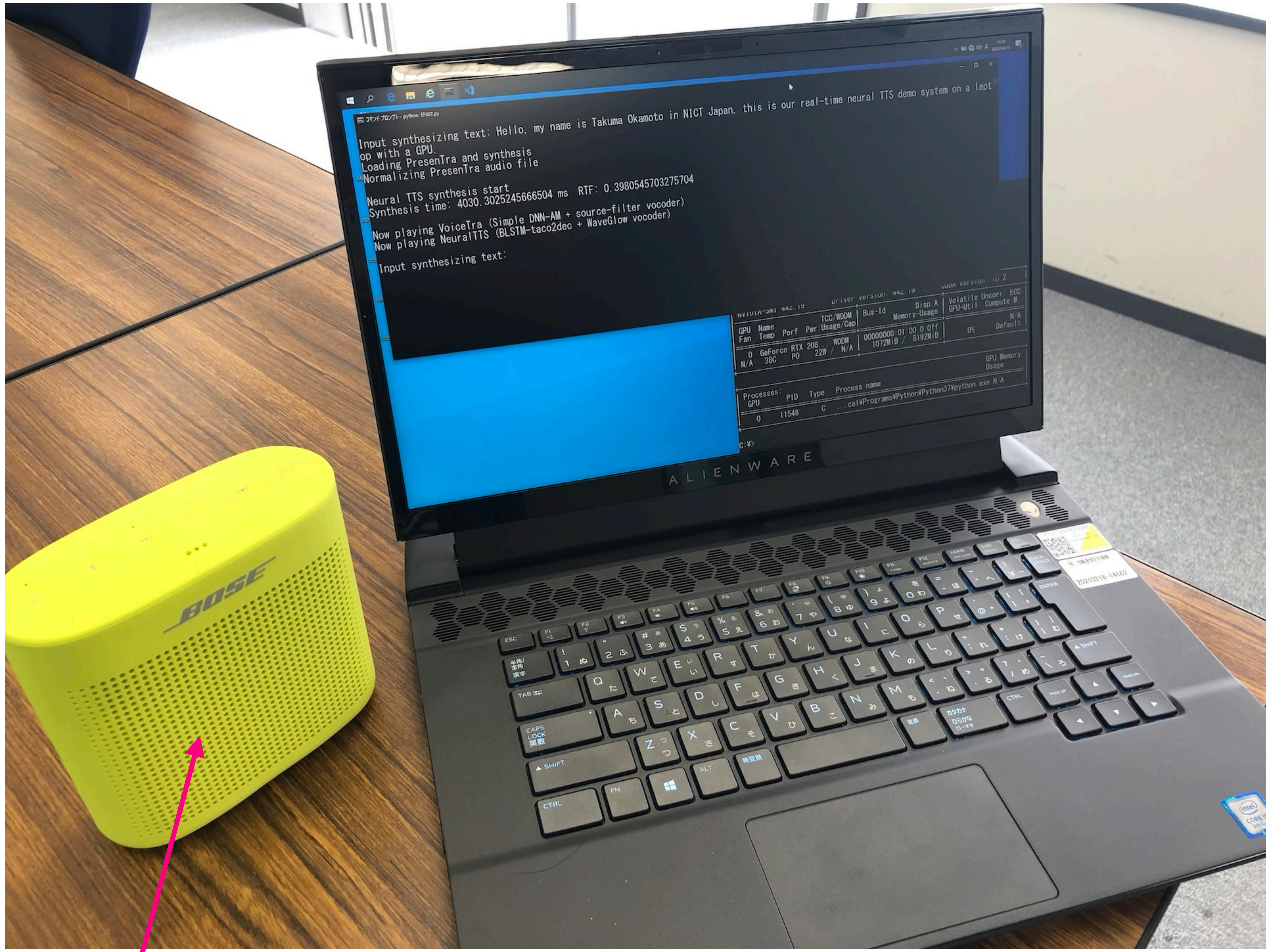


- Subjective evaluation
  - Listening subjects: 15 Japanese native speakers
  - 15 conditions x 20 utterances = 300 sentences
- Results
  - Parallel WaveGAN and small WaveGlow can realize faster training and inference than original WaveGlow
  - WaveGlow with 512 channels can realize higher synthesis quality than other models
  - Importance of full-context label input for Japanese TTS

Portable real-time neural TTS demo system

- Laptop: DELL ALIENWARE M15
  - GPU: NVIDIA GeForce RTX 2080
  - CPU: Intel Core i7-9750H 6 cores
  - Memory: 16 GB DDR4 2,666 MHz
  - 512 GB PCIe M.2 SSD
  - Windows 10 Professional

- Real-time neural TTS demo system
  - Simple PyTorch implementation
  - Acoustic model: BLSTM-Taco2dec
  - Neural vocoder: WaveGlow
  - Neural TTS models
    - 4 Japanese speakers (female and male)
    - 2 English speakers (female and male)
  - Total real-time factor: about 0.4



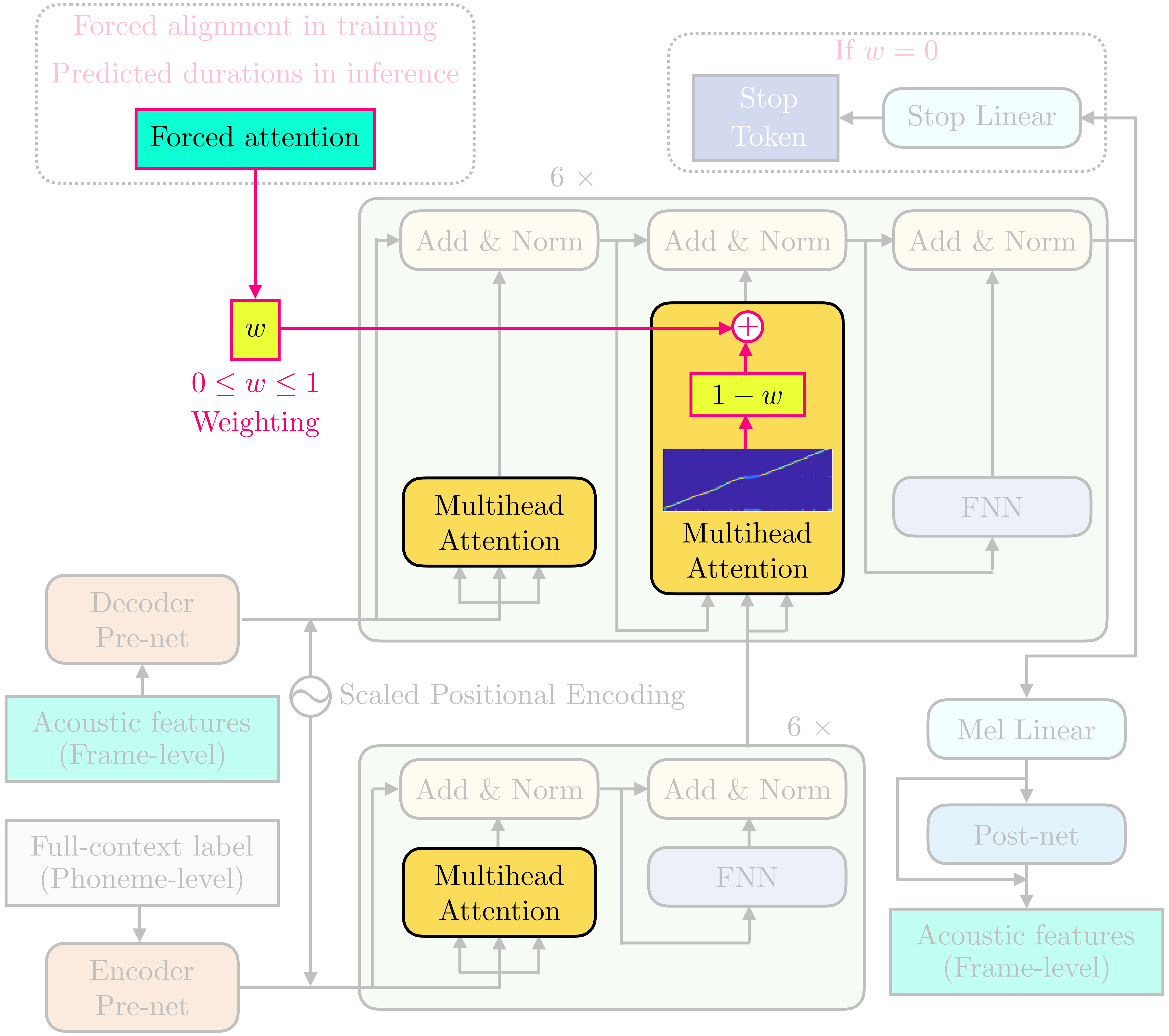
This mobile loudspeaker is a prize in a social event of ASRU 2019!!


Input synthesizing text:
GPU Name TCC/WDDM
Fan Temp Perf Pwr:Usage/Cap
Bus-Id Disp.A
Memory-Usage
Volatile Uncorr. ECC
GPU-Util Compute M.
0 GeForce RTX 208... WDDM
N/A 38C P0 22W / N/A
00000000:01:00.0 Off
1072MiB / 8192MiB
N/A
0% Default
Processes:
GPU PID Type Process name
GPU Memory
Usage
0 11548 C ...cal¥Programs¥Python¥Python37¥python.exe N/A
C:¥>
ALIENWARE

Conclusions

- Transformer-based TTS with weighted forced attention
  - Transformer- and FastSpeech-based AMs with full-context label input can also be successfully trained
  - Proposed Transformer-based AM with a weighting factor of 0.5 can significantly improve synthesis accuracy
  - FastSpeech without duration predictor can realize higher synthesis quality than that with duration predictor
  - Proposed Transformer-based AMs with weighted forced attention cannot improve synthesis stability
- Future work
  - Improving stability of transformer-based TTS for actual implementations by introducing trainable weighting factors
  - Introducing weighted forced attention to Tacotron 2
  - Introducing teacher-student training in FastSpeech-based AMs for higher synthesis accuracy
- Demo samples
  - Synthesized speech samples used in experiments are available
    https://ast-astrec.nict.go.jp/demo_samples/icassp_2020_okamoto/index.html

Thank you for your



If you have any questions, please contact us!!

okamoto@nict.go.jp